Fachartikel 2008

Prof. Dr. Michael Braun

Georg-Simon-Ohm-Hochschule für angewandte Wissenschaften – Fachhochschule Nürnberg
Keßlerplatz 12
90489 Nürnberg

ISSN 1867-4585 Schriftenreihe der Georg-Simon-Ohm-Hochschule Nürnberg Nr. 17, Nürnberg 2009

1. Inhaltsverzeichnis

Hedge-Fonds: Attrition Rate und optimaler Leverageeinsatz

Volker Gronau

Lohäckerstraße 83 a

90427 Nürnberg

Abstract

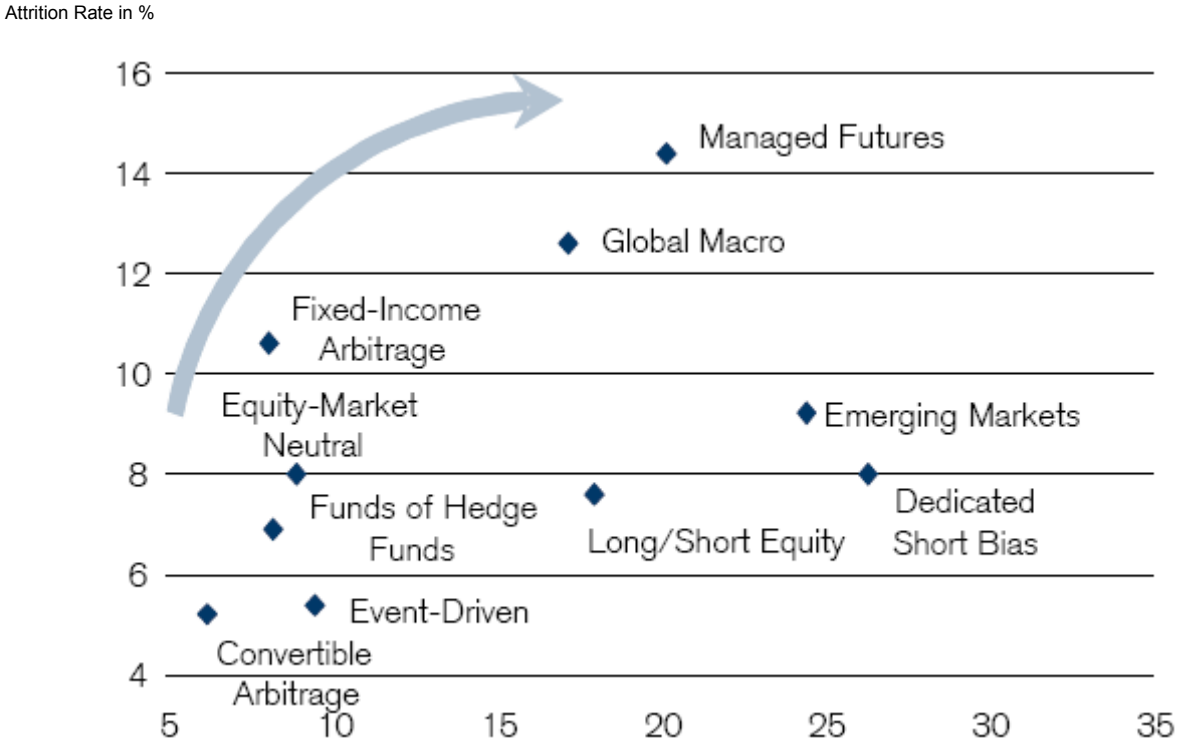
Hedge-Fonds-Investoren haben größtes Interesse daran, die besten Hedge-Fonds an den Märkten auszuwählen. Sie unterziehen die Fondsstrategie daher einer gründlichen Analyse und führen zwecks Identifizierung der besten Akteure umfassende Due-Diligence-Prüfungen durch. Trotzdem können Hedge-Fonds auch in Schwierigkeiten geraten, sodass jedes Jahr eine große Zahl von ihnen liquidiert wird, meist wegen einer enttäuschenden Performance. Laut Hedge Fund Research (HFR) wurden zwischen dem 1. und 3. Quartal 2007 814 neue Hedge-Fonds lanciert, während 409 aufgelöst wurden. Das Risiko eines Scheiterns ist selbst für eine Branche ein Problem, die gemäß (den höheren) Schätzungen bis zu knapp 10.000 Fonds umfasst. Da das Hauptziel eines Hedge-Fonds-Investors im Kapitalerhalt besteht, kommt der Vermeidung von Hedge-Fonds-Mißerfolgen in einem Fund of Hedge Funds größte Priorität zu.

1. Hedge-Fonds: Attrition Rate und optimaler Leverageeinsatz

1.1. Die Misserfolgsquote von Hedge-Fonds

Hedge-Fonds-Investoren haben größtes Interesse daran, die besten Hedge-Fonds an den Märkten auszuwählen. Sie unterziehen die Fondsstrategie daher einer gründlichen Analyse und führen zwecks Identifizierung der besten Akteure umfassende Due-Diligence-Prüfungen durch. Trotzdem können Hedge-Fonds auch in Schwierigkeiten geraten, sodass jedes Jahr eine große Zahl von ihnen liquidiert wird, meist wegen einer enttäuschenden Performance. Laut Hedge Fund Research (HFR) wurden zwischen dem 1. und 3. Quartal 2007 814 neue Hedge-Fonds lanciert, während 409 aufgelöst wurden. Das Risiko eines Scheiterns ist selbst für eine Branche ein Problem, die gemäß (den höheren) Schätzungen bis zu knapp 10.000 Fonds umfasst. Da das Hauptziel eines Hedge-Fonds-Investors im Kapitalerhalt besteht, kommt der Vermeidung von Hedge-Fonds-Mißerfolgen in einem Fund of Hedge Funds größte Priorität zu.

Hedge-Fonds-Stile mit volatilen Renditen/hohem Tail Risk weisen überdurchschnittliche Abgangsraten und damit höhere Misserfolgsrisiken auf. Managed Futures und Global Macro haben das höchste spezifische Risiko.



Quelle: Liang und Park (2005)

1.2. Unterscheidung zwischen Abgangs- und Misserfolgsquote

Die Suche nach präzisen Zahlen zu gescheiterten Hedge-Fonds gestaltet sich schwierig, denn die Ergebnisberichterstattung der Hedge-Fonds an Datenbanken erfolgt auf freiwilliger Basis. Ein Fonds kann die Übermittlung seiner Zahlen einstellen, weil er erfolgreich war und seine optimale Größe erreicht hat (und der Manager daher nicht länger nach neuen Investoren sucht) oder weil die Performance enttäuschend war und der Fonds daher kaum Chancen hat, neue Anleger für sich zu gewinnen. Als Abgangsrate (Attrition Rate) wird der jährliche Prozentsatz an Fonds bezeichnet, die ihre Berichterstattung an eine Datenbank einstellen. Frühe Studien der Hedge-Fonds-Branche gingen davon aus, dass Hedge-Fonds, die ihre Performance nicht mehr offenleg-

ten, gescheitert seien. Auf dieser Basis ergaben sich eher alarmierend hohe Misserfolgsquoten (Failure Rates) von schätzungsweise 7,1-8,7%.
Man sollte sich daher vor Augen führen, dass ein Abgang nicht auch gleich eine Liquidation oder gar ein Scheitern impliziert. Ein Fonds, der seine Performancezahlen nicht weiter in eine Datenbank einspeist, hat also nicht unbedingt seinen Betrieb eingestellt. Schneeweiss et. al. (2007) weisen darauf hin, dass gemäß jüngsten Research-Erkenntnissen mehr als 40% der Fonds, die eine Datenbank verlassen, weiterhin an andere Datenbanken Bericht erstatten. Darüber hinaus können sich Manager auch aus anderen Gründen als einem Scheitern zur Liquidation eines Fonds entscheiden. Und schließlich erhalten die Anleger in der Regel auch dann einen Grossteil ihres anfänglich investierten Kapitals zurück, wenn der Fonds aus operativen, mit der Performance zusammenhängenden oder rechtlichen Gründen (Betrug, irreführende Buchführung oder Ergebnisberichterstattung) tatsächlich scheitert.

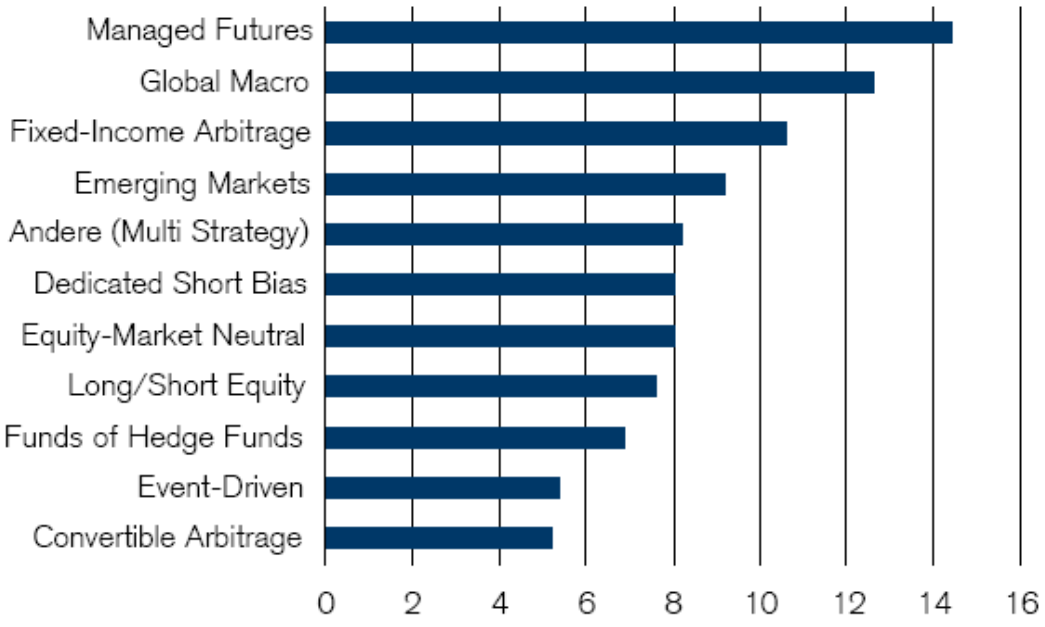
Hedge-Fonds können ihre Berichterstattung an eine Datenbank also aus den verschiedensten Gründen einstellen:

- Schwache Performance, die im Prinzip einem Scheitern gleichkommt;
- Gute Performance, sodass kein zusätzliches Kapital erforderlich ist;
- Fusionen und Namensänderungen, die in keinem direkten Zusammenhang mit der Performance stehen.

Laut Grecu, Malkiel und Saha (2006) ist eine schwache Wertentwicklung der häufigste Grund, der Hedge-Fonds von einer weiteren Berichterstattung absehen lässt. Dennoch sind die folgenden drei Szenarien in der Praxis auseinander zu halten:

- **Abgang (Attrition):** Der Fonds stellt die Berichterstattung ein und wird aus der Datenbank gestrichen.
- **Liquidation:** Der nicht mehr Bericht erstattende Fonds stellt seinen Betrieb ein und wird aufgelöst.
- **Misserfolg (Failure):** Der Hedge-Fonds wird aus operativen, mit der Performance zusammenhängenden oder rechtlichen Gründen liquidiert.

Historische Attrition Rates von Hedge Fonds, aufgeschlüsselt nach Hedge-Fonds-Strategien (1994-2003)



Quelle: Chan et al. (2005), Europäische Zentralbank

Durchschnittliche jährlich Attrition Rates in %, 1994 – 2003

Eine Schätzung der Failure Rate ist aufschlussreich, weil sie neben typischen ertragsrelevanten Parametern wie Volatilität oder Verlustrisiko eine zweite Risikodimension erfasst. Hedge-Fonds sind allgemein dafür bekannt, dass ihre Erträge und Betas mittlere Standardabweichungen aufweisen, was sie für Anleger zu Diversifikationszwecken attraktiv macht. Indessen vermag eine derartige grundsätzliche Risiko-Ertrags-Analyse über das Risiko, einen Hedge-Fonds mit besonders schlechter Performance zu wählen, keinen Aufschluss zu geben.

1.3. Unterschiedliche Hedge-Fonds-Stile mit unterschiedlichen Attrition Rates

Die Abgangsrate von Hedge-Fonds kann zwischen den einzelnen Strategien stark variieren (vergleiche „Historische Attrition Rates von Hedge Fonds“). Es überrascht nicht, dass volatilere Stile höhere Attrition Rates aufweisen (vergleiche „Hedge-Fonds-Stile“). Mit 14,4% und 12,6% verzeichneten Managed Futures und Global Macro beispielsweise die höchsten Abgangsquoten aller Hedge-Fonds-Stile, was angesichts der hohen Volatilität der beiden Ansätze kaum erstaunt.

Die Stile Fixed-Income Arbitrage (10,6%) und Emerging Markets (9,2%) folgten knapp dahinter.

Der Attrition-Wert der Emerging-Market-Fonds ist vermutlich Ausdruck der diversen Krisen, welche die Schwellenländer in den 1990er Jahren durchliefen, und er könnte daher in Zukunft durchaus niedriger ausfallen.

Convertible Arbitrage (5,2%) und Event-Driven (5,4%) glänzten mit den geringsten Attrition Rates.

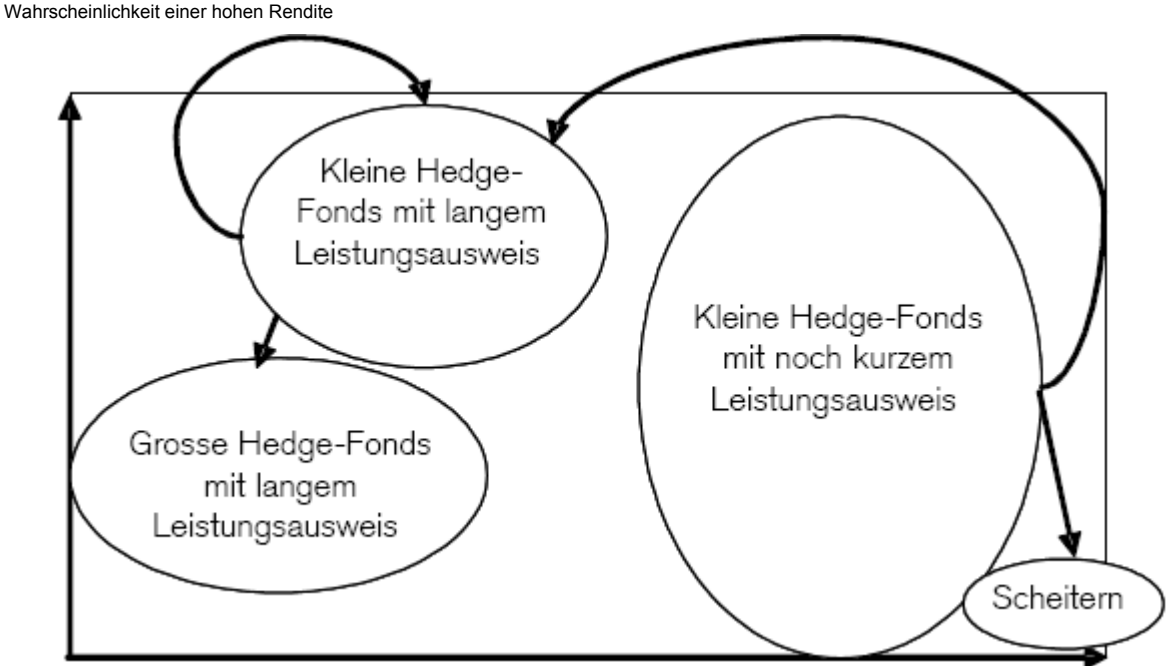
1.4. Einfluss der Hedge-Fonds-Grösse und der Ertragsvolatilität auf die Ausfallquote

Malkiel und Saha (2005) haben belegt, dass die Größe eines Hedge-Fonds und die Standardabweichung der Erträge in den letzten zwölf Monaten einen statistisch bedeutenden Einfluss (Größe: negativ; Ertragsvolatilität: positiv) auf die Wahrscheinlichkeit dafür haben, dass ein Hedge-Fonds in finanzielle Schieflage gerät. Darüber hinaus ist gemäß Grecu, Malkiel und Saha (2006) die Wahrscheinlichkeit eines Scheiterns für große Hedge-Fonds mit guter Performance geringer.

Diese Erkenntnisse scheinen auch intuitiv richtig, sind doch größere Hedge-Fonds ihrer Natur nach konservativer. Sie generieren einen höheren Ertragsanteil mit Managementgebühren und sind weniger stark von Performancegebühren abhängig. Dies steigert den Anreiz zur Risikoreduktion und folglich zur Wahrung der vorhandenen Asset- und Ertragsbasis.

Dieser Befund deckt sich mit Getmanskys (2005) Erkenntnissen, wonach die Beziehung zwischen Performance und Vermögensumfang anfänglich positiv ist. Wenn erfolgreiche Hedge-Fonds indes über ihre optimale Grösse hinaus wachsen, nimmt die Performance tendenziell eher ab als zu.

1.5. Hedge-Fonds: Risiko/Ertrag relativ zur Größe und zum Leistungsausweis



Quelle: Credit Suisse

1.6. Attrition und Failure Rates in der Praxis

Laut Getmansky (2005) betrug die Abgangsquote für Hedge-Fonds zwischen 1994 und 2000 durchschnittlich 7.1%. Eine weitere Studie der Hennesse Group gelangte zum Schluss, dass sich die Attrition Rate für Hedge-Fonds mit Vermögen von über USD 10 Mio. von 1999 bis 2006 auf durchschnittlich 5,2% p.a. belief.

Am höchsten war die Abgangsquote im Jahr 2000 (6,4%), sie bildete sich aber danach kontinuierlich zurück und lag 2006 bei 5,1%. Damit resultierte für die letzten acht Jahre ein Durchschnittswert von 5,2%.

Eine Studie von Chan et al. (2005) ergab für den Zeitraum 1994-2003 eine durchschnittliche Attrition Rate von 8,8%. Liang und Park (2007) wiesen für die Jahre 1995-2004 einen nahezu identischen Wert von 8,7% nach.

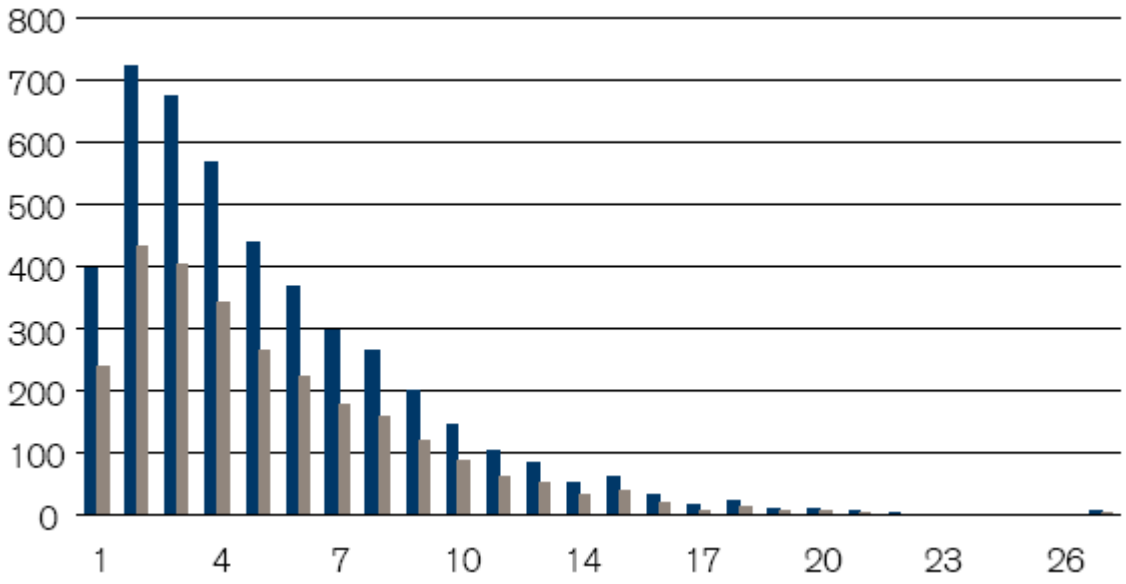
Allerdings gelangt diese Studie auch zum Schluss, dass die tatsächliche Failure Rate bei lediglich 3.1% lag. Dieses Resultat weicht beträchtlich von früheren Werten ab, weil in diesem Fall striktere Regeln zu Ermittlung der „effektiven Misserfolgsquote“ angewendet wurden.

Diese 3,1%ige Wahrscheinlichkeit eines Scheiterns bezieht sich auf die individuellen Fonds, ist aber auf vermögensgewichteter Basis wahrscheinlich noch niedriger. Rund 84% der von Hedge-Fonds gemanagten Aktiven konzentrieren sich in Fonds, die verwaltete Vermögen von USD 500 Mio. oder mehr aufweisen (Quelle: HFR).

Die Chancen dieser Exponenten, den nächsten Zyklus zu überleben, stehen ziemlich gut.

Die Zahl der Fonds, die ihre Berichterstattung an die TASS-Datenbank einstellten, ging nach 5 bis 7 Jahren merklich zurück.

Anzahl der Fonds, die ihre Performance-Berichterstattung einstellten



Quelle: Grecu, Malkiel und Saha (2006)

- Anzahl Fonds, die ihre Performance-Berichterstattung einstellen
- Minus 40%, die weiterhin an eine andere Datenbank Bericht erstatten

1.7. Der Lebenszyklus von Hedge-Fonds

Unzureichende Größe, operative Probleme und eine relativ zu anderen Fonds enttäuschende Performance sind die Hauptgründe für hohe Attrition Rates. Darüber hinaus reduziert eine schwache Wertentwicklung auch die Chancen, in den nächsten Jahren Performancegebühren verdienen zu können („High Watermark"-Prinzip), was ebenfalls für die Schließung eines Hedge-Fonds sprechen könnte.

Des Weiteren steigert auch der Konkurrenzdruck in diversen Hedge-Fonds-Strategien die Wahrscheinlichkeit eines Scheiterns. Grecu, Malkiel und Saha (2006) haben gezeigt, dass die Wahrscheinlichkeit für ein Scheitern („Hazard Rate") im fünften Jahr des Bestehens eines Fonds am höchsten ist und danach kontinuierlich zurückgeht.
Aus obiger Abbildung „Anzahl der Fonds, die ihre Performance-Berichterstattung einstellten" ist ersichtlich, dass die Zahl der Fonds, die ihre Berichterstattung an die TASS-Datenbank einstellten, nach dem fünften bis siebten Jahr ihres Bestehens markant fällt.

Mit anderen Worten: Haben Hedge-Fonds die ersten paar kritischen Jahre erst einmal überstanden, verfügen sie über gute Chancen, während längerer Zeit zu überleben.

1.8. Implikationen für die Anleger

Zusammenfassend lässt sich sagen, dass Anleger, die nach überdurchschnittlichen Hedge-Fonds-Erträgen streben, sich auf kleinere Fonds konzentrieren sollten, die möglichst schon seit ein paar Jahren bestehen sowie einen soliden Leistungsausweis und ein angemessenes Risikomanagement aufweisen.

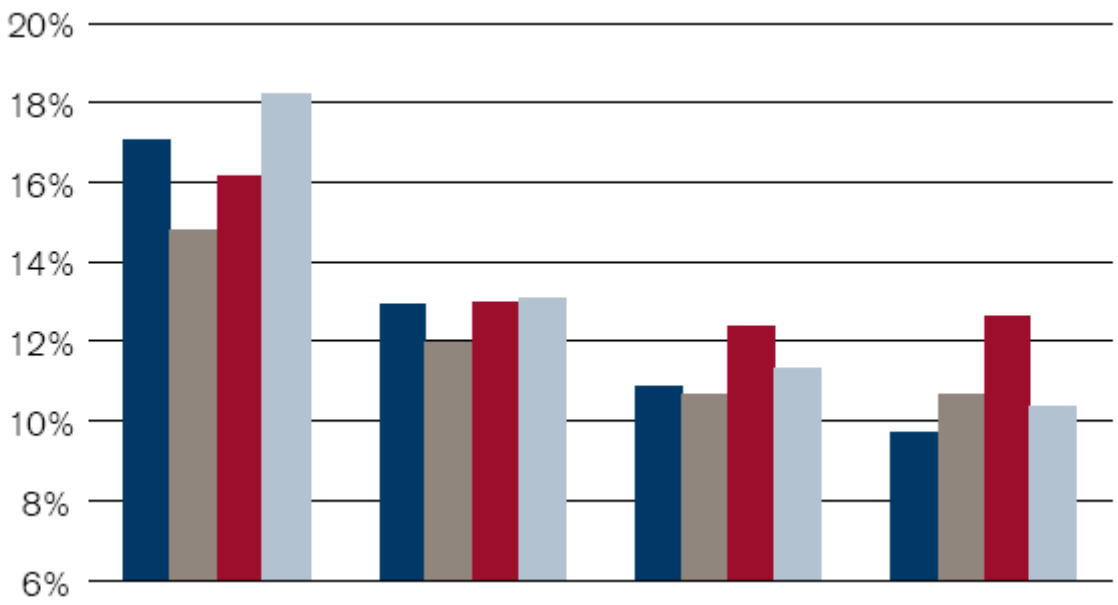
Der einzige Nachteil dieses Ratschlags ist es offensichtlich, dass derartige Fonds unter Umständen nur bedingt bereit sind, substanzielle Neugeldzuflüsse aufzunehmen, da diese die Renditen beeinträchtigen könnten. Weniger risikobereite Anleger, deren Hauptinteresse bei der Portfoliodiversifikation liegt, sollten sich dagegen in großen, etablierten Hedge-Fonds engagieren, da diese in der Vergangenheit mit der geringsten Wahrscheinlichkeit in Zahlungsverzug gerieten.

Da neue, üblicherweise kleinere Hedge-Fonds in den ersten Jahren ihres Bestehens höhere Renditen erzielen (siehe nachfolgende Abbildung), sollten risikobewusste Anleger den Kauf einer Art „Venture Capital Hedge Fund Portfolio“ in Betracht ziehen oder die Zusammenstellung eines eigenen Portfolios aus kleinen, noch jungen Hedge-Fonds prüfen, die von erfahrenen Spezialisten geführt werden und über dementsprechend gute Erfolgschancen verfügen.

Sie sollten also ein gewisses Risiko für Misserfolge zu schultern bereit sein, zumal dieses durch die Chance aufgewogen wird, künftig höhere Gewinne einzufahren, da kleinere Hedge-Fonds mit erfahrenen Managern überdurchschnittliche Erträge anstreben.

Neuere Hedge-Fonds zeigen eine tendenziell bessere Performance als ältere, da die Manager Ersterer bestrebt sind, sich einen guten Leistungsausweis aufzubauen und neue Investoren zu gewinnen, um wachsen zu können.

Rendite in %



Quelle: Hedge Fund Research

■ 1-j. annualisiert ■ 5-j. annualisiert
■ 3-j. annualisiert ■ 10-j. annualisiert

1.9. Hedge-Fonds und Fremdkapitaleinsatz: Risiken und (Ertrags-)Vorteile

Die partielle Finanzierung von Finanzanlagen mit Fremdkapital (Leverage) wird oft als Zeichen spekulativen Verhaltens angesehen.

Allerdings setzt eine Vielzahl von Akteuren der Finanzbranche regelmäßig Fremdmittel ein. Der Bankensektor mit seinen fremdkapitalisierten Bilanzen ist dafür nur das offensichtlichste Beispiel. Leverage ist ein fester Be-

standteil von Hedge-Fonds-Investitionen und kann bei vorsichtigem Einsatz die Erträge verbessern, ohne das Risiko übermäßig zu erhöhen.

Einige der schlimmsten Beispiele für das Scheitern von Hedge-Fonds (Long-Term Capital Management [LTCM], Amaranth) waren indes auf einen aggressiven Fremdkapitaleinsatz und übertriebene Risikobereitschaft zurückzuführen.
Im Folgenden werden Indizien analysiert, die Rückschlüsse auf das angemessene Leverage-Niveau sowohl für breit diversifizierte Hedge-Fonds-Anlagen als auch für Einzelstrategien ermöglichen.

Zwei Beispiele aus dem täglichen Geschäftsleben sollen diese Erläuterungen leichter verständlich machen.

- Ein Einfamilienhaus im Wert von USD 1 Mio. wird beim Kauf mit 40% Eigenkapital und 60% Fremdkapital (in Form einer Hypothek) finanziert. Damit weist die Investition in dieses neue Eigenheim einen Leverage-Faktor von 2,5 auf (Leverage = Gesamtwert dividiert durch Eigenkapital). Da der Kauf eines Hauses in der Regel als sicheres Investment gilt, scheint ein Leverage-Faktor von 2,5 in diesem Fall völlig in Ordnung.

- Die Bilanzen praktisch sämtlicher börsenkotierter und nicht börsenkotierter Unternehmen sind teilweise fremdfinanziert, d.h. sie setzten sich aus Eigenkapital und Verbindlichkeiten zusammen. Gemäß den Eigenkapitalvorschriften Basel II können sich z.B. Banken bis zum 10-12-fachen ihres Eigenkapitals mit Fremdmitteln finanzieren.

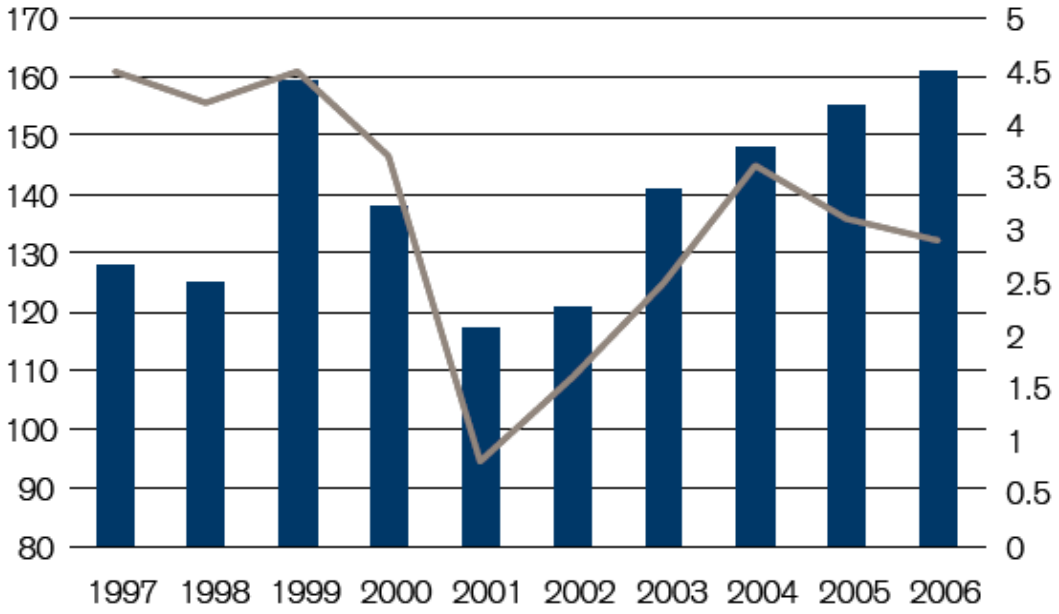
Die Bilanz einer finanziell soliden Bank mit einem Eigenkapitalanteil von 10% weist beispielsweise einen Leverage Faktor von 10 auf.

Leverage an sich ist kein spekulatives Konzept. Investitionen mit Leverage sind an den Finanzmärkten sehr verbreitet. Allerdings orientiert sich die Höhe des Leverage in der Regel an der Voraussagbarkeit der künftigen Cashflows. Je höher die Prognosesicherheit hinsichtlich der Cashflows, desto höher ist auch der Leverage, der sich auf Investments eingehen lässt. Versorgungsunternehmen sind für ihre stabilen Cashflows bekannt und finanzieren ihre Bilanzen dementsprechend stark mit Fremdmitteln.

1.10. Die Leverage-Faktoren von Hedge-Fonds

Gemäß Schätzungen von McKinsey & Company (2007) beträgt der durchschnittliche Leverage-Faktor in der Hedge-Fonds-Branche rund 3,0. Folglich umfassen die Bruttoinvestitionen im Wert von rund USD 6 Bio., die von den Hedge-Fonds verwaltet werden, rund USD 2 Bio. an Anlegerkapital. Das Brutto-Marktexposure von Hedge-Fonds relativ zu den verwalteten Gesamtvermögen orientiert sich eng an der zugrunde liegenden Risikoabneigung (vergleiche folgende Abbildung).

Die Risikobereitschaft von Hedge-Fonds und deren Leverage entwickelten sich tendenziell in Einklang mit dem allgemeinen Konjunkturzyklus.



Quelle: McKinsey & Company (2007), Bloomberg

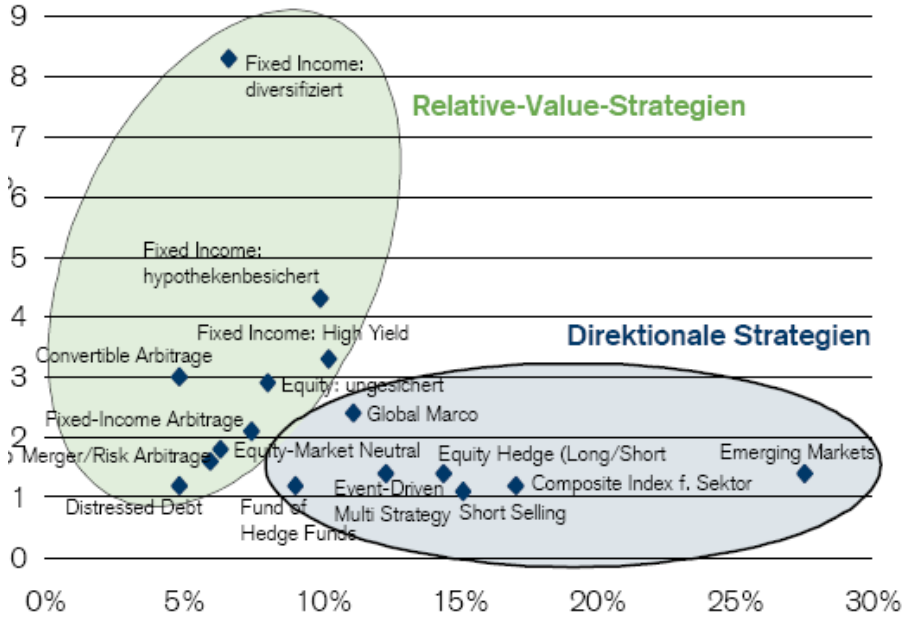
- Brutto-Marktexposure Hedge-Fonds in % verwaltetes Vermögen (Skala links)
- Veränderung US-BIP in % YoY (Skala rechts)

Es ist bemerkenswert, dass rund 50% der Hedge-Fonds über einen nur geringen oder gar keinen Leverage verfügen (gemäß Finanzstabilitätsbericht der Europäischen Zentralbank). Lediglich 13% der Hedge-Fonds wiesen einen Leverage-Faktor von mehr als 2,0 auf, und diese Exponenten konzentrierten sich in den Stilen Fixed-Income Arbitrage und Convertible Arbitrage.

Gut geführte Hedge-Fonds mit einem weit zurückreichenden und soliden Leistungsausweis stützen sich in den wenigsten Fällen auf Leverage als bestimmenden Ertragsfaktor.

Im Allgemeinen setzen direktionale Hedge-Fonds-Strategien weniger stark auf Leverage als Relative-Value-Strategien.

Vermögensgewichteter durchschnittlicher Leverage der einzelnen Strategien



Quelle: Bertelli (2007)

1.11. Leverage von Strategie zu Strategie unterschiedlich

Schneeweis, Martin, Kazemi und Karavas (2004) haben gezeigt, dass verschiedene Hedge-Fonds-Strategien in der Regel unterschiedlich hohe Fremdkapitaleinsätze aufweisen. Sie gelangten indes auch zur Erkenntnis, dass sich die risikobereinigten Wertentwicklungen innerhalb einer einzelnen Hedge-Fonds-Strategie zwischen Hedge-Fonds mit über- oder unterdurchschnittlichem Leverage statistisch nicht unterscheiden.

Dies bestätigt die an der Wall Street weit verbreitete Einschätzung, wonach gute Manager Fremdkapital nur sparsam einsetzen und ihre Erträge stattdessen mit überlegenen Fähigkeiten erzielen. Im Gegensatz dazu setzen schlechte Manager in guten Zeiten auf einen übermäßigen Leverage als hauptsächlichen Ertragstreiber, während sie bei Marktschocks schmerzhafte Verluste erleiden.

Als Faustregel lässt sich feststellen, dass sich wenig volatile Strategien wie Fixed-Income Arbitrage tendenziell stärker auf Fremdkapital abstützen, um ihre Ertragsziele zu erreichen. Sie sind aber in turbulenten Marktphasen auch höheren Ereignisrisiken (Tail Risks) unterworfen.

1.12. Leverage-Einsatz durch Hedge-Fonds - ein Beispiel

Eine der Haupteigenschaften, die Hedge-Fonds definiert und von anderen Anlageklassen unterscheidet, ist der Einsatz von Fremdkapital zur Ertragsverbesserung.

Im Prinzip kann ein Hedge-Fonds auf zwei verschiedene, sich ergänzende Arten Leverage eingehen:

Entweder indem er Fremdkapital aufnimmt (Broker-Kredit), oder indem er außerbilanzielle Positionen, wie z.B. Derivate oder strukturierte Notes, zur Renditesteigerung verwendet.

Zur Vereinfachung wird davon ausgegangen, dass der risikofreie Satz für die Kreditaufnahme oder Kreditgewährung 5% beträgt. Ein Hedge-Fonds mit einem verwalteten Vermögen von USD 40 Mio. könnte sodann rund USD 60 Mio. an Kredit aufnehmen, um den Kauf von Wertpapieren im Wert von USD 100 Mio. zu finanzieren.

Wenn der Wert dieser Titel auf USD 120 Mio. steigt, beträgt die auf einem ursprünglichen Leverage-Faktor von 2,5 beruhende Eigenkapitalrendite 43% [(120-100-3):40)].

Dasselbe Ergebnis hätte auch mit Derivaten wie Futures oder Optionen erreicht werden können. Das Ausmaß des von Hedge-Fonds eingesetzten Leverage lässt sich im Allgemeinen nicht genau beziffern, da sie ihre Positionen nicht offenlegen. Indes vermitteln empirische Studien gewisse Einsichten in die Verschuldungsniveaus spezifischer Hedge-Fonds-Stile.

1.13. Wir wirkt sich Leverage in verschiedenen Marktzyklen auf das Risiko-Ertrags-Profil eines Portfolios aus?

Unter normalen Marktbedingungen sorgt Leverage für zusätzliche Liquidität an den Finanzmärkten. Darüber hinaus reduziert der Fremdkapitaleinsatz die Kreditkosten und versetzt Hedge-Fonds in die Lage, kleine Preisabweichungen zu nutzen und illiquide Wertpapiere zu erwerben.

Damit akzentuiert Leverage die Preisschwankungen und verstärkt bei Marktverwerfungen die Abwärtsrisiken.

Betrachten wir abermals das Beispiel eines Wertpapierportfolios im Wert von USD 100 Mio., das mit USD 40 Mio. an Eigenkapital und USD 60 Mio. in Form eines Broker-Kredits finanziert wurde. Wenn die Marktkurse der betreffenden Wertschriften um durchschnittlich 10% fallen, schrumpft das Eigenkapital von USD 40 Mio. auf USD 30 Mio., während der Leverage-Faktor (wie gesagt die Verhältniszahl Gesamtwert/Eigenkapital) von 2,5 auf 3,0 steigt.

Falls der Fonds aus Gründen des Risikomanagements einen konstanten Leverage anstrebt, könnte der Manager gezwungen sein, Vermögensanteile in Höhe von USD 15 Mio. abzustoßen, um den Leverage-Faktor wieder auf 2,5 zu drücken. Nach diesem Schritt hielte der Hedge-Fonds noch Wertschriften im Wert von USD 75 Mio., die mit USD 30 Mio. an Eigenkapital und USD 45 Mio. an Krediten finanziert werden. Wenn mehrere Hedge-Fonds gleichzeitig durch einen solchen Marktabschwung zu Verkäufen gezwungen werden, kann dies bei den betroffenen Wertschriften eine markant höhere Volatilität zur Folge haben.

1.14. Bestimmung des optimalen Leverage für Hedge-Fonds-Anlagen – Ergebnisse der Credit Suisse

Die obige Abbildung (Vergleich „Relative Value“ gegen „Direktionale“ Strategien) belegt eine inverse Beziehung zwischen dem Leverage und der Volatilität der einzelnen Hedge-Fonds-Strategien.

Die Manager tendierten dazu, die (oft bescheidenen) Renditen volatilitätsarmer Strategien mittels Leverage zu steigern. Manager von Einzelfonds müssen die potenziellen Vorteile einer partiellen Fremdkapitalfinanzierung ihrer Positionen gegen die damit einhergehenden Ausfallrisiken abwägen.

Gewisse Strategien, wie z.B. Fixed-Income Arbitrage oder Merger Arbitrage, reagieren sensibel auf Ereignisrisiken. Ein übermäßiger Leverage kann Verluste akzentuieren. Aus diesem Grund sind große, etablierte Hedge-Fonds in Bezug auf den Fremdmitteleinsatz in der Regel konservativ. Dies wirft die Frage auf, ob die Anleger nicht eher die teilweise Fremdfinanzierung eines diversifizierten Hedge-Fonds-Portfolios prüfen sollten.

1.15. Drei Schlüsselaspekte gilt es in diesem Kontext zu beachten:

- Wie beeinflusst Leverage die Performance von Indizes aus einzelnen Hedge-Fonds?
- Was bedeutet die Steigerung der Renditen mittels Leverage für den Kapitalerhalt? Die Credit Suisse hat jeden Hedge-Fonds-Stil und Leverage-Grad auf die Wahrscheinlichkeit eines Drawdown von mehr als 10% des Startkapitals zu Beginn jedes Jahres getestet (d.h. die Shortfall-Risiken mit einem jährlichen Floor von 90% kalkuliert).
- Welches wäre der optimale Leverage-Faktor für Anleger, die in ein über sämtliche Hedge-Fonds-Strategien hinweg diversifiziertes Portfolio investiert sind?

1. Leverage-Belastungstest anhand diversifizierter Hedge-Fonds-Indizes

Der HFR Composite Index und die HFR-Subindizes der einzelnen Stile wurden für den Zeitraum 1990 bis 2007 mittels Leverage Faktoren von 1,5, 2,0 und 3,0 einem Belastungstest unterzogen.

Es wird davon ausgegangen, dass Anleger die betreffende Hedge-Fonds-Strategie zu folgenden Konditionen mit Fremdkapital unterlegen können: 1-Monats-USD-Libor plus fixer Spread von 240 Basispunkten plus (variabler) Renditespread zwischen Commercial Papers und Treasurys.
Die dritte Kostenkomponente trägt der Einschätzung Rechnung, dass sich die Kosten für die Kapitalaufnahme in Liquiditätskrisen, wie wir sie zurzeit durchlaufen, dramatisch erhöhen können und die Anleger einen Zinsaufschlag bezahlen müssen, um ihre bestehenden Broker-Kreditlinien offen zu halten.

Höhere Renditen mit Leverage, aber niedrigere Sharpe Ratios

Zeitraum 1990 bis 2007 **Annualisierte Renditen**

Hedge-Fonds	Ohne Leverage	Leverage-Faktor 1.5	Leverage-Faktor 2.0	Leverage-Faktor 3.0
HFR Composite	13.81%	16.87%	19.87%	25.65%
HFR Fund of Hedge Funds	9.94%	11.02%	12.02%	13.80%
Convertible Arbitrage	9.81%	10.90%	11.97%	14.02%
Event-Driven	14.32%	17.65%	20.95%	27.39%
Global Macro	15.15%	18.86%	22.49%	29.50%
Equity-Market Neutral	8.89%	10.07%	11.24%	13.54%
Equity Hedge (Long/Short)	16.75%	21.30%	25.79%	34.58%
Fixed-Income Arbitrage	7.90%	7.99%	8.04%	7.98%

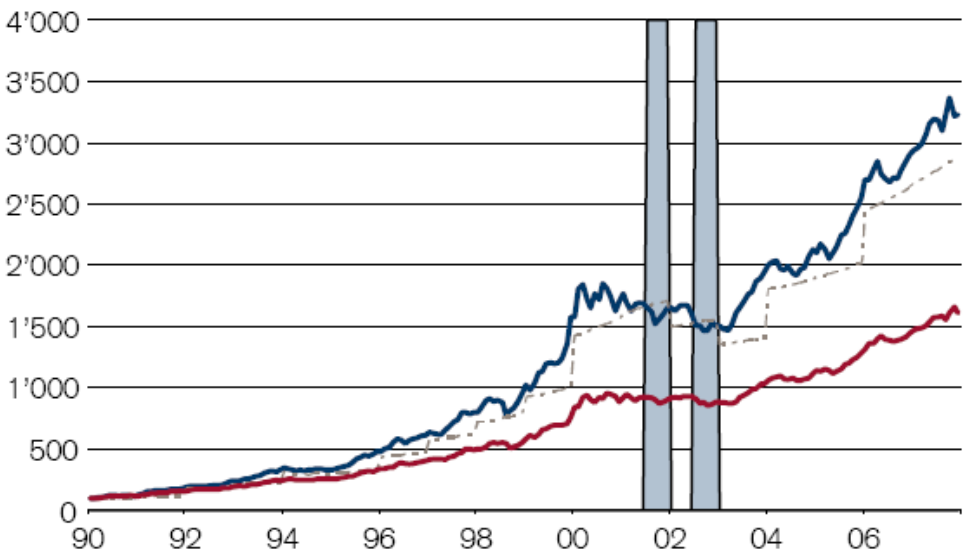
Quelle: Credit Suisse

Die obige Tabelle zeigt, dass die annualisierte Rendite zwischen 1990 und 2007 unabhängig vom gewählten Leverage (1,5, 2,0, 3,0) oder von der Strategie zunahm.

Die Performanceverbesserung infolge des Einsatzes von Fremdkapital fiel bei volatilen und renditestärkeren Stilen wie

- Long / Short Equity

Performance HFR Equity Hedge (Long/Short) 1,5

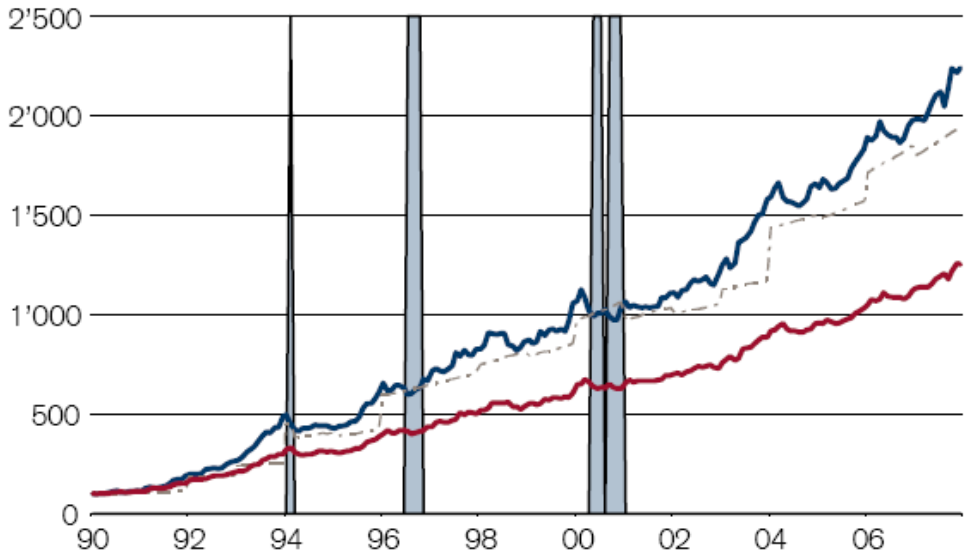


Quelle: Bloomberg, Datastream, market maker (vwd)

- Phasen unterhalb Ziel-Floor
- HFR Equity Hedge, Leverage 1,5 — Floor
- HFR Equity Hedge (Long/Short)

- Global Macro

Performance HFR Global Macro mit Leverage 1,5



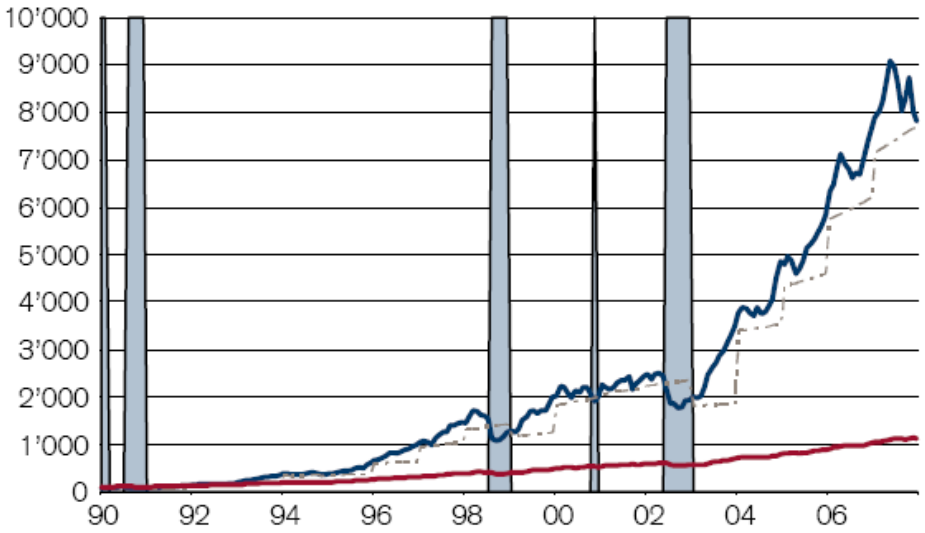
Quelle: Bloomberg, Datastream, market maker (vwd)

- Phasen unterhalb Ziel-Floor
- HFR Global Macro, Leverage 1.5 — Floor
- HFR Global Macro

und

- Event-Driven

Performance HFR Event-Driven mit Leverage 3,0

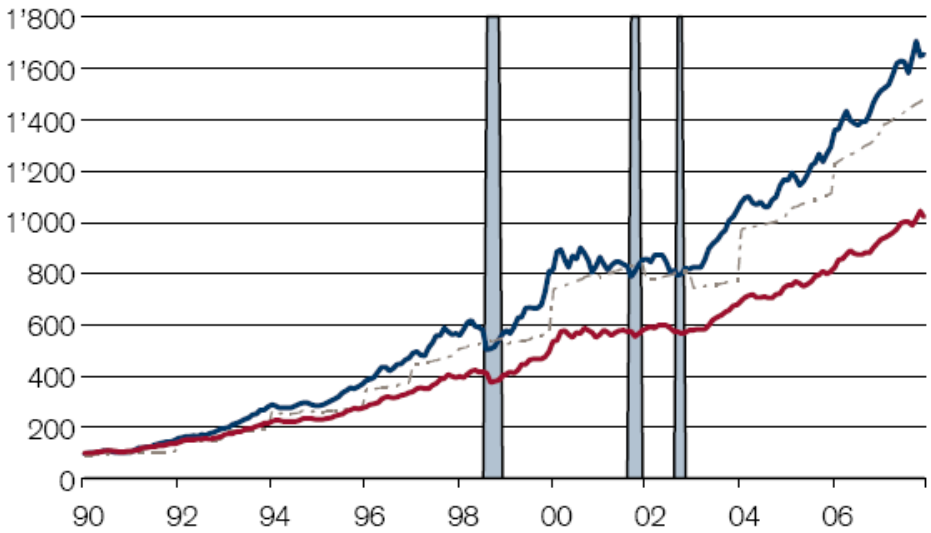


Quelle: Bloomberg, Datastream, market maker (vwd)

- Phasen unterhalb Ziel-Floor
- HFR Event-Driven, Leverage 3.0 — Floor
- HFR Event-Driven

sowie dem - HFR Composite Index am stärksten aus.

Performance HFR Composite mit Leverage 1,5

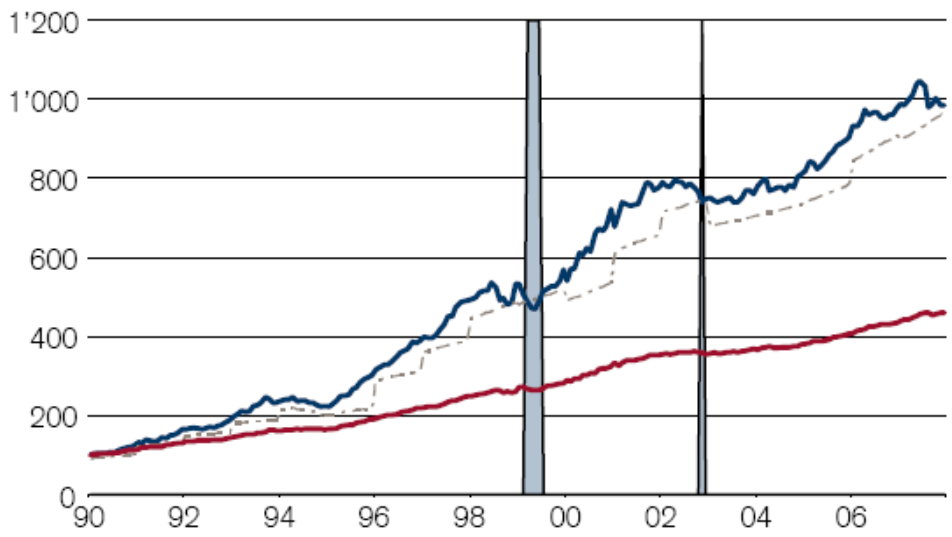


Quelle: Bloomberg, Datastream, market maker (vwd)

- Phasen unterhalb Ziel-Floor
- HFR Composite, Leverage 1,5 — Floor
- HFR Composite Index

Equity-Market Neutral

Performance HFR Equity-Market Neutral mit Leverage 3,0

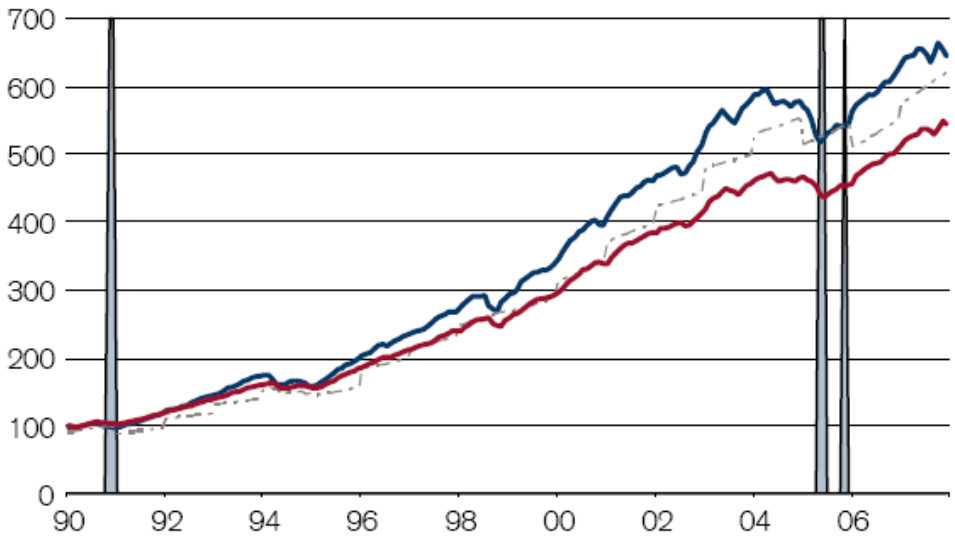


Quelle: Bloomberg, Datastream, market maker (vwd)

- Phasen unterhalb Ziel-Floor
- HFR Equity-Market Neutral, Leverage 3,0 — Floor
- HFR Equity- Market Neutral

Convertible Arbitrage

Performance HFR Convertible Arbitrage mit Leverage 1,5



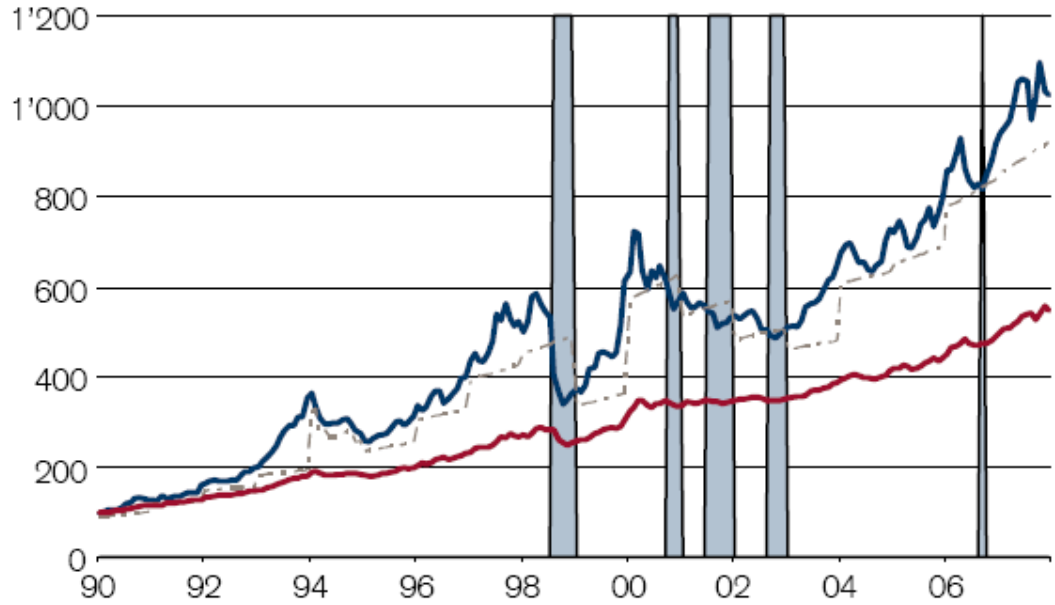
Quelle: Bloomberg, Datastream, market maker (vwd)

- Phasen unterhalb Ziel-Floor
- HFR Convertible Arbitrage, Leverage 1,5 — Floor
- HFR Convertible Arbitrage

und

Fund of Hedge Funds

Performance HFR Fund of Funds mit Leverage 1,5

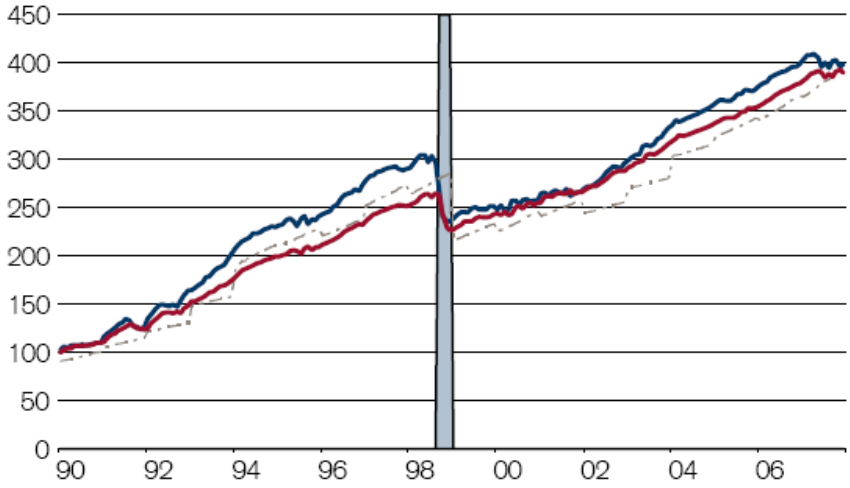


Quelle: Bloomberg, Datastream, market maker (vwd)

- Phasen unterhalb Ziel-Floor
- HFR Fund of Funds, Leverage 1.5 — Floor
- HFR Fund of Funds

generierten bei der Unterlegung mit Fremdkapital nur moderat höhere Renditen (+1.5 Prozentpunkte jährlich). Für Fixed-Income Arbitrage war der Effekt praktisch vernachlässigbar.

Performance HFR Fixed-Income Arbitrage mit Leverage 1,5



Quelle: Bloomberg, Datastream, market maker (vwd)

- Phasen unterhalb Ziel-Floor
- HFR Fixed-Income Arbitrage, Leverage 1.5 — Floor
- HFR Fixed Income Arbitrage

Meine persönlichen Annahmen und der meiner Recherchen (u.a. Allianz Global Investors und Credit Suisse) hinsichtlich der Leverage-Kosten könnten sich für diesen Stil, der tendenziell stabile, wenig volatile Erträge generiert, leicht nachteilig ausgewirkt haben.

Interessanterweise erbrachte keine dieser Strategien höhere risikobereinigte Renditen.

Sharpe Ratio

Hedge-Fonds	Ohne Leverage	Leverage-Faktor 1.5	Leverage-Faktor 2.0	Leverage-Faktor 3.0
HFR Composite	1.479	1.295	1.198	1.090
HFR Fund of Hedge Funds	1.083	0.856	0.735	0.599
Convertible Arbitrage	1.667	1.326	1.150	0.965
Event-Driven	1.628	1.436	1.336	1.228
Global Macro	1.402	1.246	1.164	1.070
Equity-Market Neutral	1.619	1.367	1.234	1.093
Equity Hedge (Long/Short)	1.501	1.362	1.289	1.208
Fixed-Income Arbitrage	0.958	0.653	0.494	0.324

Quelle: Credit Suisse

Die Sharpe Ratio, ein Maß der um das Risiko (in diesem Fall um die Volatilität) bereinigten Erträge ging mit steigendem Leverage kontinuierlich zurück.

Dies bedeutet, dass mit Leverage erzielte Erträge mit einem höheren Risiko einhergehen - keine überraschende Erkenntnis, die man aber dennoch nicht vergessen sollte.

2.1. Shortfall-Risiko nimmt mit Leverage zu

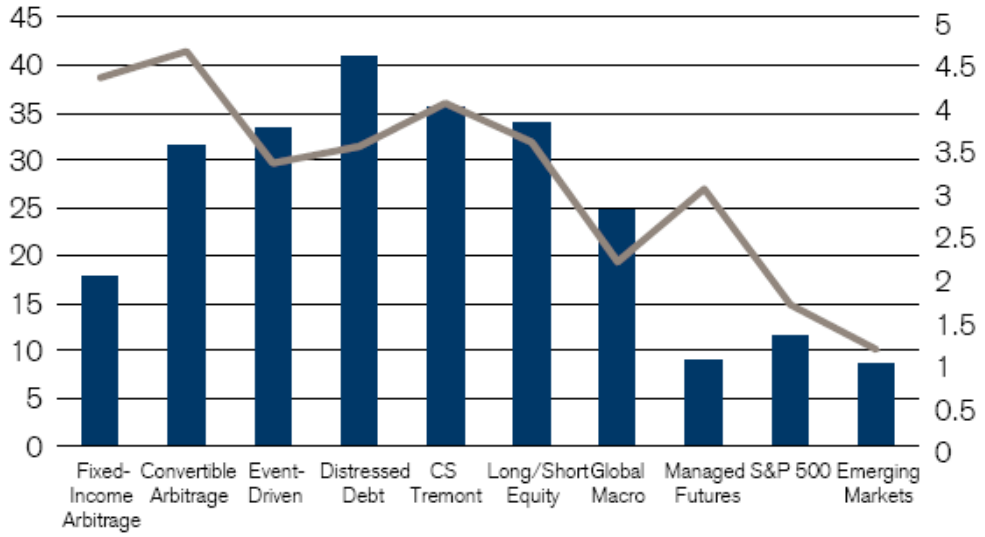
Das Risiko-Ertrags-Profil von mit Fremdkapital unterlegten Hedge-Fonds-Anlagen verändert sich merklich.

Leverage steigert die Renditen auf lange Sicht, verstärkt aber auch die kurzfristigen Auswirkungen negativer Schocks und sorgt damit für eine höhere Volatilität.

Es wurden die verschiedenen Hedge-Fonds-Stile einem Belastungstest unterzogen, indem die Leverage-Grade schrittweise erhöht wurden. Gemäß der Ausgangsidee sollte der Wert der Investitionen mit Leverage im Idealfall nicht unter einen Floor fallen, der bei 90% des Nettoinventarwerts (NAV) zu Beginn eines jeden Jahres festsetzt worden ist.

Der maximale Leverage, mit dem sich der jährliche max. Drawdown eines Hedge-Fonds auf 50% beschränken lässt, unterscheidet sich von Strategie zu Strategie.

Annualisierte Rendite in % (1994.2007) Maximaler Leverage



- Maximaler Leverage (Gesamt-Assets/Eigenkapital)
- Maximaler Leverage zur Beschränkung des maximalen jährlichen Drawdown um 50%

Mit einem höheren Leverage stieg offensichtlich auch das Risiko, dass der NAV unter unseren Ziel-Floor von 90% einbrechen würde (Shortfall-Risk). Die folgenden Beispiele und die obigen Darstellungen illustrieren den unterschiedlichen Grad der Anfälligkeit verschiedener Hedge-Fonds-Strategien.

Gemäß dem Backtesting für die Jahre 1990 bis 2007 fiel der Wert der Anlagen bei einem Leverage-Faktor von 1.5 (1/3 Fremdkapital, 2/3 Eigenkapital) nur selten unter unseren 90%-Floor. Der Wert der Anlage erholte sich in der
Regel innerhalb von 3 bis 6 Monaten nach seinem Einbruch unter das Floor-Niveau. Die Ausnahme von dieser Regel war der Fixed-Income-Arbitrage-Ansatz, der 1998 (im Zuge des LTCM-Debakels) einen substanziellen Drawdown erlitt.

Diese im Nachhinein erhobenen Fakten scheinen die Ansicht zu untermauern, dass ein moderater Leverage selbst unter Berücksichtigung von Ereignis- und Tail-Risiken mit einer verkraftbaren Zunahme des Risikos einhergeht.

Eine leichte Verschärfung des Szenarios, d.h. eine Erhöhung des Leverage auf 2.0 (50% Fremdkapital, 50% Eigenkapital) ergab keine dramatische Veränderung der Situation.

Die durchschnittliche Zahl der Monate, in denen der Anlagewert unter unserem Floor von 90% verharrte, sowie das Ausmaß der Drawdowns nahmen zwar marginal zu, aber der durchschnittliche Shortfall blieb für risikobewusste Anleger weiterhin tolerierbar und bewegte sich zwischen 3,8% (für den HFR Composite Index) und 7,5% (für den Fund of Hedge Funds Index).

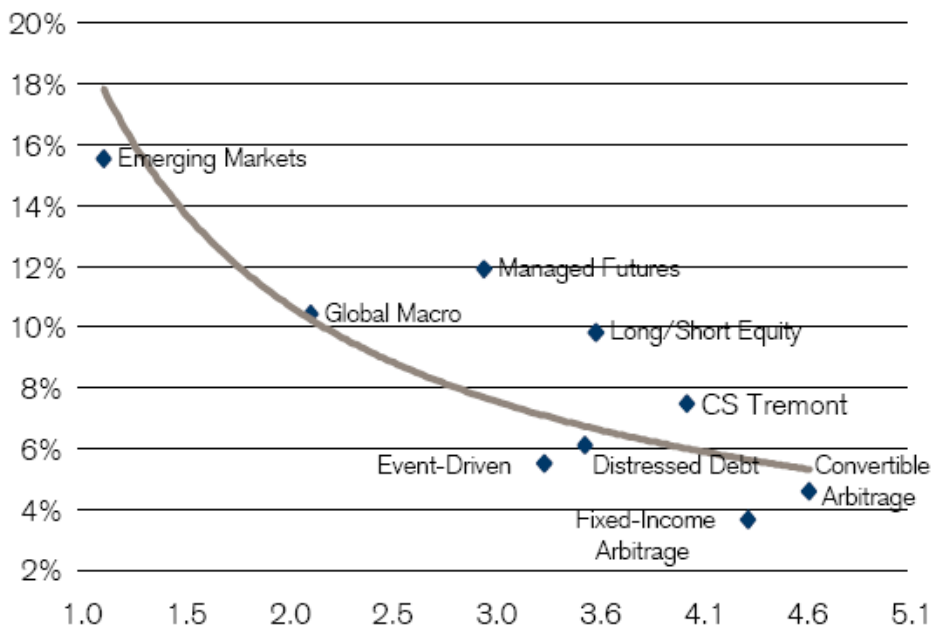
Die Zeit unter dem 90%-Floor betrug 4. 5 Monate, was für den Leverage-bereiten Anleger ebenfalls akzeptabel ist.

Ein gänzlich anderes Bild ergab sich schließlich, wenn der Leverage auf 3.0 (1/3 Eigenkapital, 2/3 Fremdkapital) erhöht wurde. Shortfalls nahmen markant zu, und der durchschnittliche Shortfall unter dem 90%-Floor erreichte für den Fund of Hedge Funds Index .10%.

Mit anderen Worten: Jedes Mal, wenn der Wert des Investments unter den 90%-Floor einbrach (der jedes Jahr neu festgelegt wurde), büssten der Anlagewert nochmals 10% ein, bevor dieser die Talsohle erreichte und sich wieder erholte.

Volatilitätsarme Hedge-Fonds-Strategien haben ein Interesse an der Maximierung des Leverage, während volatilere Strategien ohnehin riskant genug sind und daher nur zu moderaten Leverage-Faktoren tendieren.

Volatilität



Quelle: T. Schneeweis, G. Martin, H. Kazemi und V. Karavas (2004)

Welche Hedge-Fonds-Strategie-Indizes wiesen das geringste durchschnittliche Shortfall-Risiko auf?

Einerseits verzeichneten die Stile Equity-Market Neutral und - in etwas geringerem Ausmaß - Convertible Arbitrage bei Leverage-Faktoren von 1.5 und 2.0 nur wenige Shortfalls.

Für den HFR Composite Index verhielt sich dies ähnlich, während der Fund of Hedge Funds Index mehr Federn lassen musste.

Andererseits hatte eine Fremdkapitalunterlegung bei Portfolios in den Strategien Event-Driven, Global Macro, Long/Short Equity und Fixed-Income Arbitrage einen deutlichen Anstieg des Shortfall-Risikos zur Folge.

2.2. Bestimmung des optimalen Leverage für diversifizierte Hedge-Fonds-Anlagen

Diese empirische Analyse legt den Schluss nahe, dass für diversifizierte Hedge-Fonds-Anlagen eine Fremdkapitalisierung von maximal einem Drittel auf lange Sicht eine vertretbare Maßnahme darstellt.

- Leverage kann sowohl vom Hedge-Fonds-Manager als auch von Hedge-Fonds-Investoren auf ihren Anlagen eingesetzt werden. Diese Analyse lässt es für Anleger höchst ratsam erscheinen, ihre Investments auf Portfolio-Ebene mit Fremdkapital zu unterlegen.

- Sofern (große) Hedge-Fonds keine Vorzugsbehandlung hinsichtlich der von Prime-Brokern gewährten Finanzierungskosten nachweisen können, sind Hedge-Fonds unserer Meinung nach nicht unbedingt der richtige Ort für den Einsatz von Leverage, zumal ihr Leverage-Faktor nur allzu oft unangemessen (d.h. entweder zu hoch oder zu tief) sein könnte.

- Grosse, ältere Hedge-Fonds sind eventuell in Bezug auf ihren Leverage-Einsatz manchmal zu konservativ, da das Geschäftsmodell schwergewichtig auf die Stabilität der Erträge und das Überleben der Organi-

sation (an welcher die Partner üblicherweise hohe Beteiligungen halten, die wiederum einen Grossteil ihres persönlichen Nettovermögens ausmachen) ausgerichtet ist.

- Kleinere, jüngere Fonds könnten versucht sein, aggressiv auf Leverage zu setzen, um Performance und einen entsprechend guten Leistungsausweis zu generieren. Aber Investitionen in derart stark mit Fremdkapital unterlegte Vehikel würde auf die Bezahlung hoher Performancegebühren für ein fremdfinanziertes Beta-Exposure hinauslaufen - was wohl kaum als attraktive Value Proposition bezeichnet werden kann.

- Folglich sollten sich Hedge-Fonds in erster Linie darauf konzentrieren, hervorragende Trading- und Investmentfähigkeiten zu entwickeln und ein solides Risikomanagement zu pflegen.

- Wir raten nicht zur Fremdkapitalunterlegung einzelner Hedge-Fonds-Anlagen. Allerdings zeigt unsere Analyse, dass risikobewusste Anleger für diversifizierte Hedge-Fonds-Investments, wie z.B. breit abgestützte Hedge-Fonds-Portfolios oder einen Fund of Hedge Funds, Leverage in Betracht ziehen sollten.

- Leverage-Niveau: Angesichts der keiner Zeit zu unterschätzenden Risiken der Kreditmarktrisiken muss die Empfehlung lauten, eher zu vorsichtig zu sein. Der Vermeidung von Liquiditätsrisiken kommt vor allem im aktuellen Umfeld der Subprime-Krise höchste Priorität zu. Der Einsatz von bis zu 30% Fremdkapital zur Finanzierung von Positionen in Hedge-Fonds-Portfolios scheint angemessen.

Zeitraum 1990 bis 2007

		HFR Composite	HFR Fund of Hedge Funds	Convertible Arbitrage	Event-Driven	Global Macro	Equity-Market Neutral	Equity Hedge (Long/Short)	Fixed-Income Arbitrage
Durchschnittlicher Shortfall	Leverage 1.5	-2.9%	-5.5%	-1.3%	-5.5%	-2.7%	0.0%	-4.2%	-13.3%
	Leverage 2.0	-3.8%	-7.5%	-3.5%	-8.9%	-5.1%	-0.2%	-6.9%	-14.1%
	Leverage 3.0	-9.5%	-10.1%	-6.3%	-17.7%	-8.6%	-2.1%	-10.1%	-9.7%
Maximale Zahl der Monate bis Erreichung des Floor	Leverage 1.5	4	5	2	6	4	0	6	4
	Leverage 2.0	6	5	9	7	8	1	7	4
	Leverage 3.0	10	6	9	7	9	4	11	5
Gesamtzahl der Einbrüche unter den Floor	Leverage 1.5	3	3	3	3	4	0	2	1
	Leverage 2.0	6	3	3	4	6	1	4	2
	Leverage 3.0	5	5	7	5	9	2	7	7

Literaturhinweise

N. Chan, M. Getmansky, S.M. Haas und A.W. Lo (2005), “Systemic Risk and Hedge Funds”, NBER, Arbeitspapier 11200.

R. Bertelli (2007), “Financial Leverage: Risk and Performance in Hedge Fund Strategies”, Arbeitspapier.

B.G. Malkiel und A. Saha (2005), “Hedge Fund: Risk and Return”, Financial Analyst Journal, Jahrgang 61, Nr. 5.

T. Schneeweis (2007), “Where Academics and Practitioners Got It Wrong”, Arbeitspapier, Center for International Securities and Derivatives Markets.

A. Grecu, B.G. Malkiel und A. Saha (2006), “Why Do Hedge Funds Stop Reporting Their Performance?”, Arbeitspapier.

M. Getmansky (2005), “The Life Cycle of Hedge Funds: Fund Flows, Size and Performance”, Arbeitspapier.

M. Getmansky, A. Lo und S. Mei (2004), “Sifting Through the Wreckage: Lessons from Recent Hedge Fund Liquidations”.

B. Liang und H. Park (2007), “Predicting Hedge Fund Failure: A Comparison of Risk Measures”, Arbeitspapier.

McKinsey & Company (2007), “The New Power Brokers: How Oil, Asia, Hedge Funds and Private Equity Are Shaping Global Capital Markets”, McKinsey Global Institute.

T. Schneeweis, G. Martin, H. Kazemi und V. Karavas (2004), “The Impact of Leverage on Hedge Fund Risk and Return”, Isenberg School of Management, University of Massachusetts.

5. GI/ITG KuVS Fachgespräch Ortsbezogene Anwendungen und Dienste 4.-5. September 2008, Nürnberg

Jörg Roth (Hrsg.)

Georg-Simon-Ohm-Hochschule Nürnberg
90489 Nürnberg
Joerg.Roth@Ohm-hochschule.de

Abstract

Der Aufenthaltsort eines mobilen Benutzers stellt eine wichtige Information für viele zukünftige Anwendungen dar – man spricht allgemein von *ortsbezogenen Anwendungen* oder *ortsbezogenen Diensten*. Ziel dieses Fachgespräches ist, Wissenschaftlerinnen und Wissenschaftlern aber auch Vertretern der Industrie die Möglichkeit zu einem intensiven Gedankenaustausch zu Themen rund um den Ortsbezug zu geben. Nach den erfolgreichen Treffen in Hagen, Stuttgart, Berlin und München fand das fünfte Fachgespräch im Jahr 2008 in Nürnberg an der Georg-Simon-Ohm-Hochschule statt.
Die Tagung umfasste in diesem Jahr eine erfreulich hohe Zahl von 24 Beiträgen. Neben Beiträgen von Universitäten gab es viele Einreichungen von Forschungseinrichtungen, Fachhochschulen und aus der Industrie. Darüber hinaus beteiligten sich Lehrgebiete außerhalb der Informatik (Verkehrs- und Geowissenschaften). All dies zeigt, welche Aufmerksamkeit das Thema in Industrie und Forschung weiterhin genießt.
Die Komplexität des Gebiets spiegelt sich auch in der Vielfalt der behandelten Themen wider. Diese stammten aus den Themengruppen:
- Ortsbezogene Dienste im realen Einsatz
- Sicherheit ortsbezogener Dienste
- Geodaten, Umgebungsmodelle, Kontext
- Industrieller Einsatz ortsbezogener Dienste
- Netzwerkaspekte
- Ortsbezogene Dienste aus Benutzersicht

Nach der Tagung wurden die Teilnehmer aufgerufen, ihre vorgestellten Beiträge anhand der vielfältigen Diskussionen zu überarbeiten und als Papier einzureichen. Das Resultat liegt in Form dieses Tagungsbandes vor.

Prof. Dr. Jörg Roth Nürnberg, 8.10.2008

Supporting Mobility in Next Generation Internet with a Decentralized P2P LBS

Amine M. Houyou

Chair Prof. De Meer - Faculty for Computer Science and Mathematics, University of Passau,

Innstr. 43, 94032 Passau

Abstract

The recent emergence of a whole plethora of new wireless technologies, such as IEEE802.11, IEEE802.16, and UMTS, etc, also offers mobile users more diversity and possibilities for cheaper and opportunistic access to the Internet. In next generation Internet, media independence such as that proposed in IEEE 802.21, requires vertical handover between co-located heterogeneous wireless networks. Handover is, however, triggered through costly beaconing mechanisms which allow both end-device and networks to discover each other, and to detect movement. If the mobile device is made location-aware (e.g. GPS-equipped mobile phones with navigation systems), mobility could be supported by the location-awareness, and a vertical handover could be triggered without relying on frequent beacons on multiple wireless interfaces. As a result, less energy is required by the mobile node to discover wireless diversity. Instead, the mobile user discovers the network coverage via a decentralized LBS, which is designed in this work. The LBS manages location-based meta-data describing network topologies and their functionality. The description templates are managed on an overlay system connecting distributed location servers. The overlay network, which connects distributed LBS systems, is structured in way to limit the overhead introduced by the query. The structure of the network is mapped to the data structure. A design methodology is developed to ensure the localized query overhead, while taking mobility into account.

1. Context-Aware Mobility Management

The research in the domain of mobility management is now facing the challenges of wireless diversity offered in next generation Internet or 4G scenarios. The way handover should be triggered between heterogeneous networks is a challenging and complex problem.

In recent years, context-aware vertical handover has been proposed [5]. For mobile scenarios, context has been modeled with the help of ontologies that divide context information for both user and network into dynamic and static components. In [8], a centralized architecture is used, where user context and network context are retrieved and collected for a given user, a handover selection process based on a context-aware algorithm is applied to trigger the handover decision.

In fact, three components to context aware mobility management type have been proposed.

1. Context modeling dealing with context ontology [8].
2. A suitable architecture for context sensing and context retrieval, dealing with context management, sensing, and storage [5,8].
3. Decision algorithms which have been centered around fuzzy logic, and multiple criteria optimization algorithms [9,1].

In [3,4] I propose to achieve context management with the help of a distributed mobile LBS. Mobile LBS require service provision which takes user movement into account. The service delivered changes with the movement and location of the user. Movement tracking is usually done with the help of a tracking hardware like GPS, which makes the user aware of its location. Assuming that the user can use the movement tracking capabilities context sensing is adapted to the user movement [3]. Based on the user's context, the network context could be queried. A query is started to search for available heterogeneous wireless cells belonging to various operators. The discovery of layout of wireless cells allows the mobile node to select a wireless node that most suits its context. This latter problem has been approached in [1] as multiple objective optimization decision process. Vertical handover is seen as the process of selecting the cell that suits the QoS needs of the user among several alternatives. Each alternative cell is evaluated according to several objectives like *(i)* minimizing interruption time, *(ii)* maximizing bandwidth, *(iii)* longest connectivity possible, *(iv)* minimizing price of communication, *(v)* minimizing jitter, etc. The score each prospective cell achieves is combined in a utility function, which is used to select the best handover alternative between overlapping cells or next in line cell.

In another simpler approach, the mobile node simply looks up nearest free of charge wireless LAN access point (similar to war driving) the user can move towards that place using a handheld navigation system.

2. Network Context Sensing and Meta Data Abstraction

The discovery of network context has to occur across heterogeneous system. Therefore, a middleware bridging between the different networks is needed. The middleware offers a common interface to represent the heterogeneous domains and their capabilities through a service description. The network context has to be discovered and retrieved while adapting to the discovery and retrieval process to the user context. Here both the discovery and retrieval effort of network context are of interest. An important characteristic of the network context is that it is scattered along autonomous domains.

The goal of the discovery effort is to gather the network context across the domains, while caring for the dynamic nature of this information. How a network context is generated depends on the explicit and implicit descriptive information of a given wireless domain. The characteristics of a domain can include:

- Each autonomous domain should reflect geographic proximity of wireless attachment points (i.e. a continued island of wireless coverage is guaranteed via a group of access elements).
- Given a set of link technology used within each domain, some mobility management capabilities are required to describe the connectivity and the intra domain mobility support offered. An example would be to describe a UMTS domain with maximum guaranteed bandwidth, maximum handover velocity, blocking probability, dependability (in terms of probability) of the connectivity and bandwidth guarantee). In comparison a WMN based on IEEE 802.11f might offer higher bandwidth, but support a lower handover velocity, but require a much lower cost per data flow.
- Some notion of load and available resources needs to be derived for each network organization. For instance, the network load could be a measurement of the number of admitted users connected to a given available resources. The network resources are those represented by the QoS parameters, such as available bandwidth average queuing delay, or loss rate. It could also include the number of users a given mobility protocol would still scale.
- In heterogeneous wireless networks, each domain should include only those base stations implementing the same link layer protocols.

- Separating domains, including within a single link technology, apply the following heuristic: handover latency between base stations within a given domain should be considerably smaller than that between two domains.

The storage and retrieval of the data description is done using a decentralized overlay system.

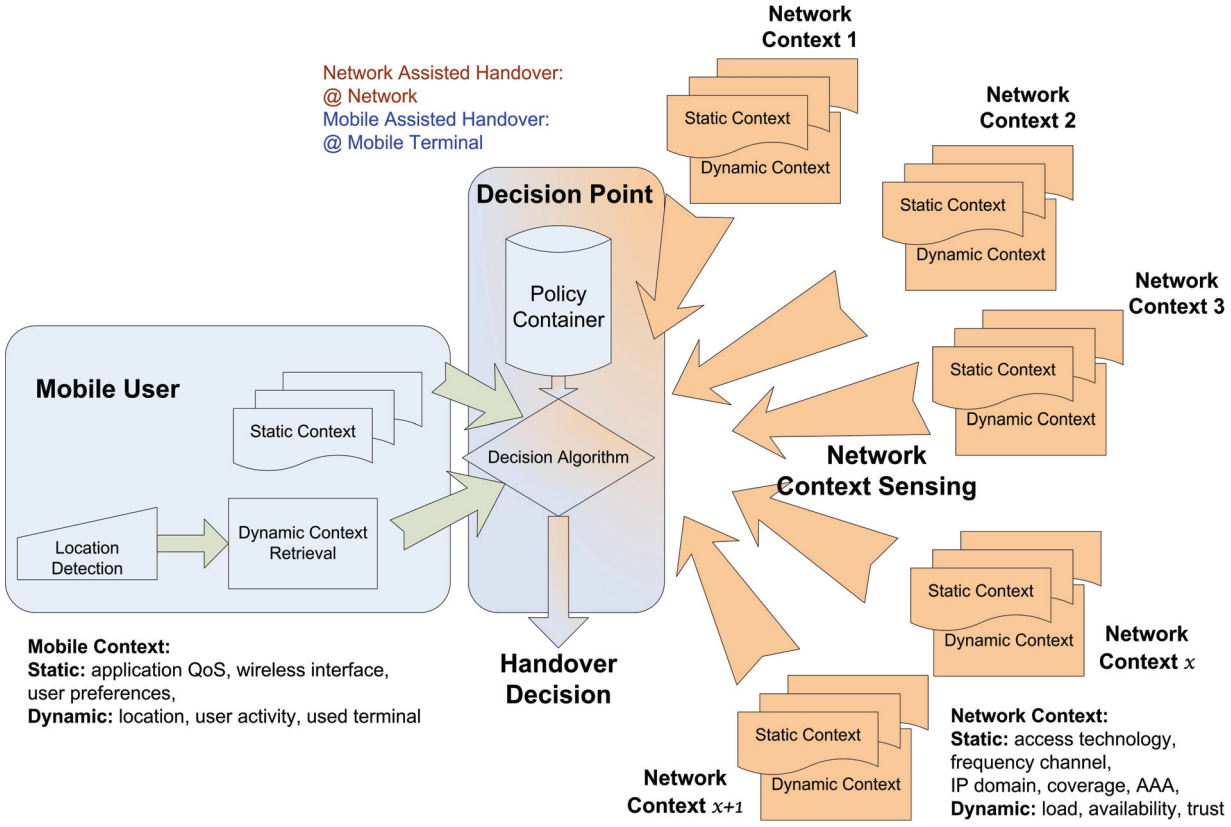


Figure 1. Context aware vertical handover – mobile and network context collection and sensing – decision point (either in network or in mobile terminal)

The architecture described in Figure 1 demonstrates the scalability and complexity problems faced by context aware mobility management. Whether a network assisted handover or mobile assisted handover is used, the decision process requires a retrieval of information from separate management domains, which has also to be adapted on one side to the user's static context (such as user network preferences, application QoS requirements, and wireless capabilities of end-terminal). An on the hand, there is also the problem of dealing with an ever changing location of the user, which requires network context adapted to the user's position and situation.

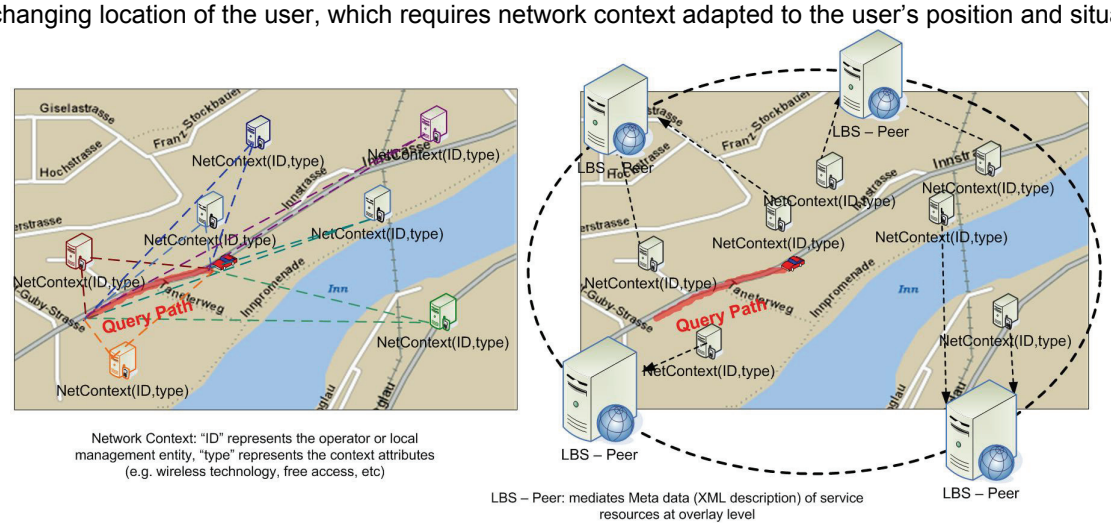


Figure 2. LBS overlay structure for providing network context to mobile users

Sensing and retrieving network context from scattered domains could be redefined as a context-rich location based service (LBS). Figure 2 shows on the left hand side a naïve approach which requires a separate query/or update process to each separate domain or context management entity. On the right hand side an overlay constructs a virtual network capable of routing to reach LBS peers (overlay nodes) as well as content. This architecture differs to classical LBS, in that a P2P relationship exists between backend severs. The overlay is used to restructure data, and route not only to pre-registered nodes, but also to scattered LBS content. The type network needed to offer efficient communication between LBS nodes is a semantic overlay.

3. Context Management in the Form of a P2P LBS

3.1. Semantic Overlays for Efficient Queries

Crespo et. al. define a semantic overlay as "a flexible network organization that improves query performance while maintaining a high degree of node autonomy" [2]. A semantic overlay aims at improving the querying process in decentralized systems while constructing overlay structures based on the semantics of the managed resources.

The type of queries that are supported by a semantic overlay are constrained by the following system requirements:

- Look up heterogeneous wireless cells among different operators.
- Look up only those wireless cells near the movement path of the user.
- Look up those wireless cells that the end device is capable to support.
- Limit the query path in the network for timeliness considerations (i.e. a result of a query is only relevant if it is sent back before the user has moved out of that cell).
- A wireless resource status could change which means that the query result is more likely to be invalid if the query occurred too long before the position has changed.

Selecting the right semantic attribute to cluster the nodes is also important. For network context, the chosen attribute has to satisfy several requirements:

- The common attribute might be described with the elements of a space *S* that could itself be split into subspaces $S_0 \cup S_1 \cup S_2 \cup \ldots S_n = S$, where *n* is a large integer representing the partition of *n* subspaces between *n* possible overlay nodes.
- Each subspace is fully independent of other subspaces $S_0 \cap S_1 \cap S_2 \cap \ldots S_n = \varnothing$, and therefore each overlay node from *n* nodes can be assigned a subspace that does not overlap with other spaces.
- The chosen attribute is common to all possible objects, but its value is assigned per object from the space *S*.

The choice of the semantic attribute could be for instance the wireless technology, so that each object representing a given technology is clustered together. Although this would satisfy the conditions above, the number of subspaces that result is quite limited. In addition to that, clustered objects representing the same technology would not scale, since this means a mobile user moving through the city of Berlin has to contact the same overlay node as a user from New York. Another possibility would be to achieve an aggregation along the lines of operator boundaries. However, the query with the same geographic scope has to be repeated per operator domain, similar to the naïve approach in Figure 2.

The more intuitive aggregating attribute is location. Those objects belonging to any operator and using any type of technology are clustered together if they are geographically close to each other or co-located together. For range queries, further DHT routing is required to efficiently structure the overlay network and query objects.

3.2. Distributed Hash Table for Location-based Range Queries

Chord [7] is the most efficient DHT protocol when looking at routing effort. In order to route to any object or peer, a *O(log(N))* effort is required. Each node has to store pointers to *O(log(N))* other nodes (where *N* is the total number of peers). The problem with Chord is that its query language requires a precise known object ID that can generate a precise *(key, value)* pair, making range queries difficult. Semantic clustering dependent of location, replaces the Chord hash addressing scheme, so clusters the *O(log(N))* search effort is repeated for each cluster instead of each object. For instance for *W* the number of objects, *O(W·log(N))* search effort is required with classical Chord. With a range gathering *w* objects, the search effort is reduced to *O((W/w)·log(N))* messages. If a range is split among several direct neighbouring peers *M,* then the overhead becomes *O((W/w)·[log(N)+M]).* The next problem with Chord is how to address geographically different types of information with a one-dimensional key system.

3.3. Hilbert Space Filling Curve for Addressing Geographic Information

In [3,4] space filling curve based addressing is shown to achieve the geocoding [6] required for addressing. The Hilbert based transformation of 2-D geographic IDs into 1-D integer space allows for instance to model the earth surface as a closed integer space $S \subset N$, with elements starting at 0 and finishing with 2^m (represented by a ring of modulo 2^m), where *m* the number of bits used in the Chord ID space.
The continuous ring representing the ID space is mapped to a Hilbert curve filled grid covering the earth surface. The clustering property of the Hilbert curve is taken into account so that neighbourhoods are preserved by the addressing scheme. In other words, two objects geographically close to each other are also close in their 1-D Hilbert address.

3.4. Fractal Design Optimization of Space Model

The problem with the Hilbert curve is the need for a large amount of keys to describe the same geographic content, which is depends on the granularity of the space filling curve. The more granular the Hilbert curve is, the larger the overhead becomes. This is partly due to the fact that each object (in our case a wireless cell) of let say *500m* × *500m* size requires *2500* Chord keys when using *38*-bit keys to model the earth, and only *one* Chord key when using 20-bit keys [4]. The problem with *20*-bit though, is that a smaller cell of let say *50m* × *50m* must be encoded as well in as *500m* × *500m* cell reflecting a significant loss of information about the stored data. The added complexity of using a higher granularity can be reduced by introducing a hierarchy in addressing both nodes and objects (example shown in Figure 3 (left)).
The compromise between the accuracy of geographic data is for instance a major reason for increased communication complexity. The way to compensate this is to take advantage of the fractal nature of the space filling curve used in addressing. Nodes are then addressed with a lower resolution than objects.
Each peer is assumed to manage a considerable number of keys, and the objects corresponding to those keys. This translates into assigning several subspaces ($S_i \cup S_{i+1} \cup S_{i+2} \cup \ldots$) to a single peer. Assuming a $2^k \times 2^k$ mapping of a 2D space is filled by the k^{th}-approximation of a Hilbert curve (resulting in *2k*-bit long IDs). If a given cell in that grid is divided into 4 subsquares, each resulting cell requires 2-bit longer ID, while sharing the same 2^{2k}-bit long prefix. The fractal nature of the Hilbert curve can be utilized to distribute the 2^{2k} possible keys in a k^{th}-approximation of a *2D* space to $N=2^p$ peers (and *p* is a binary index reflecting the number of peers). Therefore, each peer is responsible for 2^{2k-p} keys, where all objects managed by the same peer share the prefix which is 2^{2k-p} long (see Figure 3 (left)). The addressing scheme does not rely on a centralized name service neither does it require an external negotiation process. The given the geographic data items offered by a GIS are addressed locally. The bootstrapping stage allows the content of the GIS servers to be placed along the Chord ring.
Based on the number of nodes that discover each other due to predecessor and successor rules (as defined in Chord [7]), it could be said that some clusters with closely stacked data items will emerge. These clusters are quasi continuous, since they are allocating objects to keys which are separated by small or no gaps in the object Chord ring. Based on this, some further optimization of constructing the overlay could be achieved.
The fractal property of the Hilbert curve could be further applied to tree structure node scope as follows:

- Geographic urban groups require well meshed peers, whereas far away nodes are less important in a finger table.
- Transition between large geographic zones need to be supported, to allow ubiquity of the system and support moving between cities or in rural subspaces

Figure 3 (right) demonstrates the addressing principles and scope of each node in the overlay. Objects or data items can described with a higher granularity which suits the geographic information. During bootstrapping iterations, zones could be identified in the sense that those nodes sharing the longest prefix can be grouped to form a subspace representing an urban agglomeration. This subspace is a single Hilbert cluster starting from a min ID and finishing at a max ID.. In the urban subspace, a fully meshed Chord ring is built. The participating peers within that zone can elect a single node which is used as a gateway to access another ring at a higher hierarchical level. This selected node takes as an address with the longest common prefix shared by all the nodes and objects at a lower level and can be part of a smaller sized Chord ring (at a higher level). Each hierarchy is addressed by a different resolution depending on its geographic scope. The efficiency of the multiple Chord hierarchy is proved in [4].

4. Conclusion

A context management architecture has been proposed in this paper in order to build decentralized mobile LBS between context containers. The geographic-centered semantic structure of the information and data is used to address a semantic overlay network. A summary of the requirements for geographic information coding which cares for efficient communication has been listed. The overlay is also structured to limit the scope seen by

each node is a global network. The routing aspect is therefore optimized thanks to the self-similarity and fractal addressing scheme offered by the Hilbert space filling curve. The curve can address geographically dense environments with higher granularity, while keeping a less granular knowledge about less dense areas.

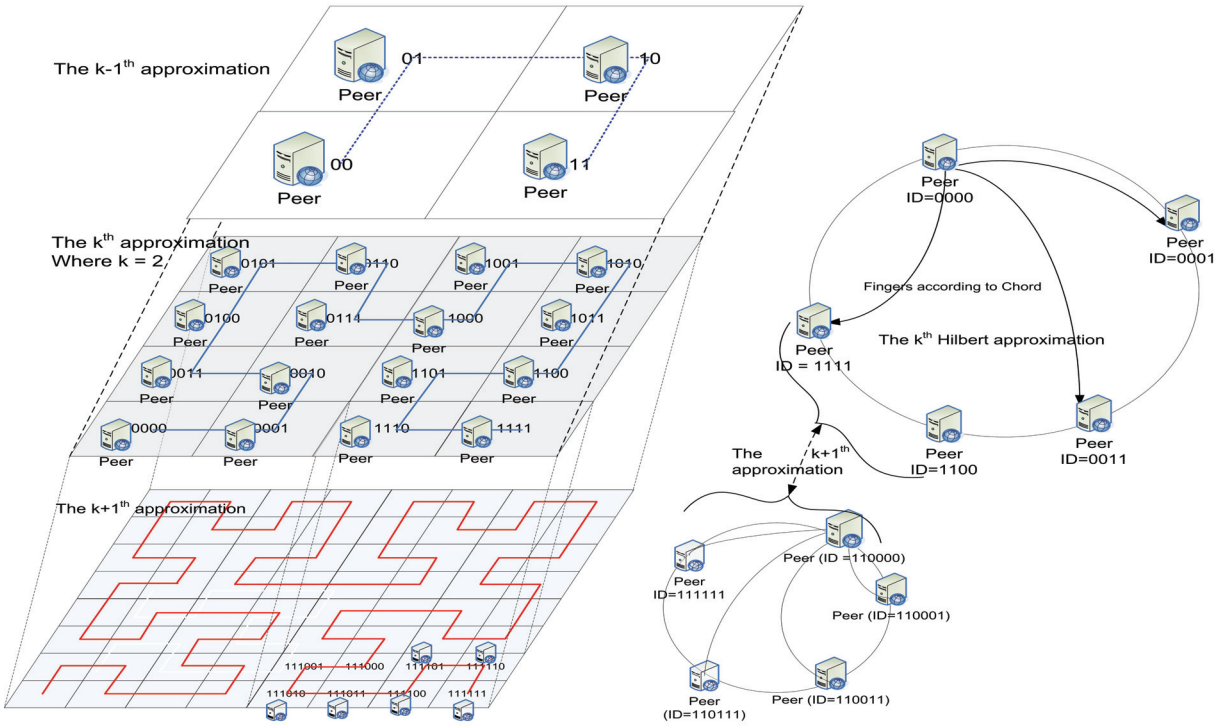


Figure 3. Left: Hierarchical partition of peer IDs following the Hilbert 1st, 2nd, and 3rd approximations; Right: Hierarchical Approach to Partition of Geographically dense spaces to more granular Chord rings

Acknowledgements

This work has been supported by the Euro-NF network of excellence, EU grant IST 216366, and specifically the specific project CAMYM.

References

1. Balasubramaniam, S., Indulska, J.: Vertical Handover Supporting Pervasive Computing in Future Wireless Networks. Computer Communications. 27(8), pp. 708-719 (May 2004)
2. Crespo, A., Garcia-Molina, H.: Semantic Overlay Networks for P2P Systems, In Agents and Peer-to-Peer Computing (invited talk), LNCS Vol. 3601, Springer, Heidelberg (2005)
3. Houyou, A. M., Stenzer A., De Meer, H.: Performance Evaluation of Overlay-based Range Queries for Mobile Systems. In Llorenç Cerdà-Alabern(eds.) Wireless Systems and Mobility in Next Generation Internet, Barcelona, Spain, January 16 – 18, 2008. LNCS, vol. 5122 – Springer, Heidelberg (2008)
4. Houyou, A. M., De Meer, H.: Efficient Overlay Mediation for Mobile Location-Based Services, LBS 2008 - 5th International Symposium on LBS & TeleCartography, Salzburg, Austria, November 26-28, (2008)
5. Prehofer, C.; Nafisi, N.; Wei, Q., "A framework for context-aware handover decisions," Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003. 14th IEEE Proceedings on , vol.3, no., pp. 2794-2798 vol.3, 7-10 Sept. 2003
6. Schiller, J., Voisard, A.: Location-Based Services. Morgen Kaufmann/Elsevier, San Fransisco, CA, USA, (2004)
7. Stoica, I., Morris, R., Liben-Nowell, D., Karger, D. R., Frans Kaashoek, M., Dabek, F., Balakrishnan, H.: Chord: a Scalable Peer-to-Peer Lookup Protocol for Internet Applications. IEEE/ACM Transactions on Networking. 11(1), pp 17—32. (2003)
8. Wei, Q., Prehofer, C.: Context management in mobile environments. In *Proceedings of ANwire Workshop, Paris*, (2003)
9. Zhu, F. and McNair, J. 2006. Multiservice vertical handoff decision algorithms. *EURASIP J. Wirel. Commun. Netw.* 2006, 2 (Apr. 2006), 52-52

Privacy in Location-Based Services: Case for an End-To-End Solution

Carsten Kleiner

Fachhochschule Hannover
Fakultät IV, Abt. Informatik
Ricklinger Stadtweg 120
30459 Hannover

Abstract

In this article we motivate why location privacy is a very important aspect when offering location-based services for mobile devices. We review the many approaches to provide user privacy in LBS that have been proposed in the literature and classify them into certain categories. After discussing the advantages and drawbacks of each of the categories we show that none of the approaches suggested so far may be used for a large class of real-world services. That is due to the fact that they do not support optimal service quality, leaving the exact position on the device, applicability for pull services, arbitrary number of users and chargeable services at the same time. We therefore suggest an approach based on end-to-end encryption of the position information between mobile device and backend system of the service provider. We discuss the advantages of this solution and show that it even has bigger advantages when applied in service mashups. Mashups seem to be very promising for future classes of services where the mobile device carries its own positioning technology. We also show issues remaining to be solved before our solution may be used in real-world scenarios.

1. Introduction and Motivation

1.1. Location-Based Services

Location-Based Services (LBS) are increasingly becoming mainstream and are already being deployed in production environments. LBS are specifically appealing for users of mobile devices such as smartphones and PDAs, since such devices are carried by the user anyway and so location information may easily be obtained. In addition nowadays usually such devices contain some kind of positioning technology in their hardware, e.g. GPS receiver or a WLAN-based positioning method. Also more classical positioning technologies such as GSM-based positioning may be used for using LBS and they also require a mobile device on client side. Since the overall system architecture is different for such passive positioning technologies we will focus on active positioning such as GPS in this article.
In addition the same or similar services as for stationary devices may be used by using web services on server side and mobile devices on client side. Therefore mobile devices nowadays provide access to a very wide range of LBS. While one of the main goals of a service-oriented architecture is the independence of clients from the service provider's infrastructure (which we make use of by allowing mobile as well as stationary devices as clients in principle), it is beneficial to make some assumption on the components in the service backend to better understand our proposed solution in section 3. It is thus not absolutely necessary to have the suggested backend architecture (cf. figures 1-3) for our approach to work, but the assumptions will probably hold for most of the service back ends anyway..

1.2. Privacy in LBS

Especially with increasing usage potential of LBS by personal mobile devices but also in general privacy concerns play an ever increasing role in any aspect of communication due to several serious crimes performed on user privacy lately. The combination of location at a given time and a specific mobile device has to be considered an extremely sensitive piece of information, since continuous recording of such data facilitates generation of detailed movement profiles of individuals, probably even without their knowledge. Therefore a lot of research has already been performed on how to improve privacy protection in LBS. An extensive discussion on the proposed approaches and also a simple classification can be found in section 2.
The reason for this article is that even given all these approaches there is still no completely satisfying solution to protect privacy in LBS that addresses all of the following issues together:

- Optimal quality of service
- Independence of number of service users
- Exact position of the mobile device does not leave the device
- Applicability for chargable services/services with registration
- Applicability for service mashups
- Extension to protecting the service result possible
- Support for active positioning (i.e. user-initiated)
- Applicability for poll services

Since none of the approaches discussed in section 2 addresses all the above issues, we give a sketch of a potentially improved solution addressing all issues in section 3. In section 3 we also discuss this new approach and analyze its effectiveness as well as its drawbacks. We conclude with ideas for future improvements as well as open issues in the area of privacy in LBS in section 4.

2. Related Approaches

Due to the great importance of user privacy in LBS there have already been suggestions by several research groups on how to improve location privacy. In general the approaches suggested in the literature roughly fall into three different categories: (1) k-anonymity, (2) spatial and/or temporal cloaking and (3) usage of a trusted third party. There are also several approaches combining some of the ideas from different categories. These approaches usually inherit advantages and drawbacks from the different categories and thus do not solve all the issues discussed in section 1 as well.
One of the classes of suggested approaches is k-anonymity (e.g. [1], [9]), where a possible attacker is not able to distinguish the location of a client from at least k-1 other clients using the same service (cf. figure 1). The exact location of the mobile device is transmitted to some intermediary anonymizer component which uses the exact position of the device; thus service quality is not affected. Furthermore no details on how to implement the anonymizer are fixed thus even a P2P-like implementation without a physical implementation of the anony-

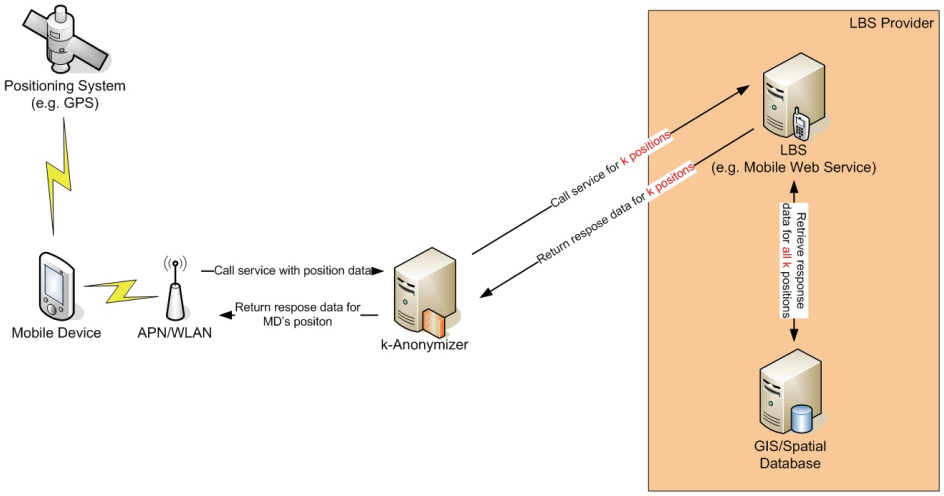


Figure 1: Privacy preservation by using a k-Anonymizer

mizer component is possible. A defined level of privacy can be achieved since k is usually dynamic and may even be chosen by the service user. Drawbacks include potentially leaking at least approximate location information from the device (depending on the particular implementation) and the fact that this approach only works for services with comparatively high number of users in a certain area. In addition due to the high number of fake service calls this model is not suitable for services requiring a charge per use.
The cloaking approach (e.g. [7], [12]) does not send the exact position of the mobile client but rather a certain kind of approximation (either a randomly placed location close to the original one or a buffer object containing the original location, cf. figure 2). The advantage is obviously that no one can determine the exact position since it is not revealed by the client. Also the degree of privacy may be chosen by the client by defining an approximation accordingly. Major drawbacks of this approach are a lack of quality of the LBS due to the approximate location information provided and leaking at least approximate location information to the service provider. There is a classical privacy vs. quality trade off in this approach since usually a service may only be of a sufficient quality, if the approximation is not too rough. It is possible to combine this approach with the k-anonymity model by e.g. defining a cloaked region as one where at least k-1 other users are located. Such combinations inherit the drawbacks of both approaches to a certain degree. Especially leaking at least approximate location information and the inability to use it for chargeable services seem important.

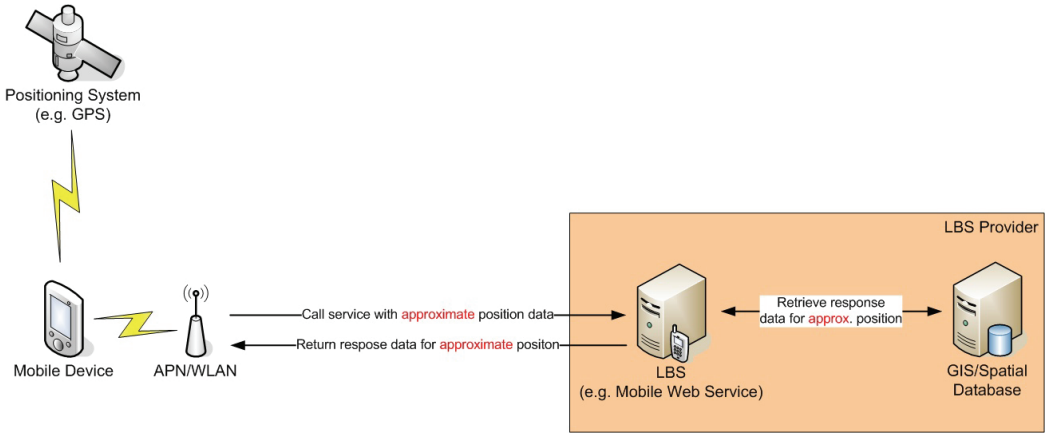


Figure 2: Privacy preservation by using spatial cloaking

In using a trusted third party (TTP) as proxy (e.g. [13], [16]), which may also be necessary to provide the previously discussed k-anonymity depending on the approach, the mobile client contacts the TTP instead of the LBS with the service request and location information (cf. figure 3). The TTP then queries the LBS on behalf of the end user and forwards the requested data on to the mobile client. This approach works well as long as the TTP can really be trusted. In the case where the TTP is the mobile service operator this assumption probably holds, since the service operator knows (at least the approximate) location of the mobile device anyway, because it is registered with its infrastructure (be it a WLAN access point or the radio cell). So user privacy is pre-

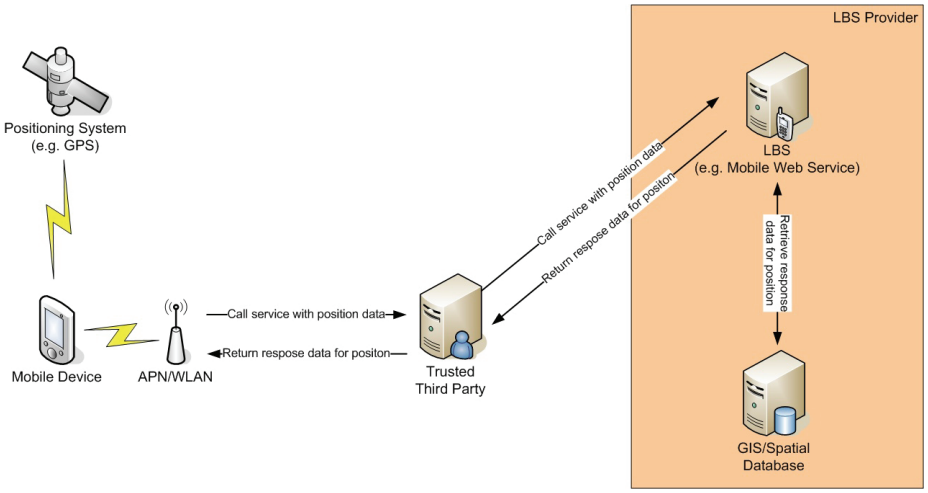


Figure 3: Privacy preservation by using a trusted third party (TTP)

served except for the TTP. Drawbacks of this approach are services where a registration is required (e.g. commercial services) since there a request has to be bound to the mobile client in some way. Moreover LBS where the response allows determining the location of the user (e.g. a map service where the location of the user is probably the center of the requested map) cannot use this approach without extending the TTP's capabilities. Such services would require specific encryption of the service response for the particular mobile unit, which is difficult to achieve without a direct connection between the two. Thus the TTP would have to handle that. Also the TTP appears as a single point of attack.

A very interesting approach which is also worked out to a very high level of detail regarding components and protocols required for implementation is the one in [15]. In that paper the components of mobile operator, application provider and location intermediary (similar to the TTP in the previous category) are introduced. The approach guarantees location privacy for push services as long as the components are really independent and do not cooperate to record a location profile. The approach is technically sound but also requires a third party as location intermediary. Moreover it is only suitable for push services (i.e. services where information is sent to the client based on passive positioning) and finally also requires some means of anonymization between client and application provider. Most solutions for this problem operate based on mixes (cf. [3], [5]) and thus require high number of service users to work properly.

3. Concept for an End-to-End Solution

In summary of section 2 each of the previously proposed solutions has drawbacks for certain types of services. Therefore we suggest using a different solution where the location of the mobile client is transmitted in encrypted form (cf. figure 4). The encryption must be in such a way that only the LBS backend database (or whichever system is used in the backend to service a request) is able to decrypt it during computation of the requested service response. The decryption may never be materialized but may only exist in main memory during computation of the response. It would also be possible to encrypt the response for the specific device using the service so that the content of the response in addition cannot be used for obtaining information about the location of the device. End-to-end privacy may be achieved by this approach, since the position (not even approximately) is revealed to no one in the process. Moreover issues such as applicability for charged/registration services, optimal service quality and independence of number of service users are addressed, since a direct flow of information between client and service provider is used. Problems of the suggested approach include:

- If the operator of the spatial database or backend component cannot be trusted to log information accessible during service computation, a volatile session key is required for encrypting the location information. This in turn makes introducing a more complex transaction protocol necessary (which is required for charged/registration services anyway).
- If encrypting the service response is necessary in order to avoid disclosing location information of the client which may be derived from a service response, the transaction protocol becomes more complex.
- The GIS/spatial database component must be able to perform the decryption at run time. In addition it must be executed pretty fast in order to obtain a low response time and thus good user experience.
- The architecture makes assumptions about capabilities of the backend system and thus introduces additional requirements for the service provider. Also it conflicts to some degree with the service concept.

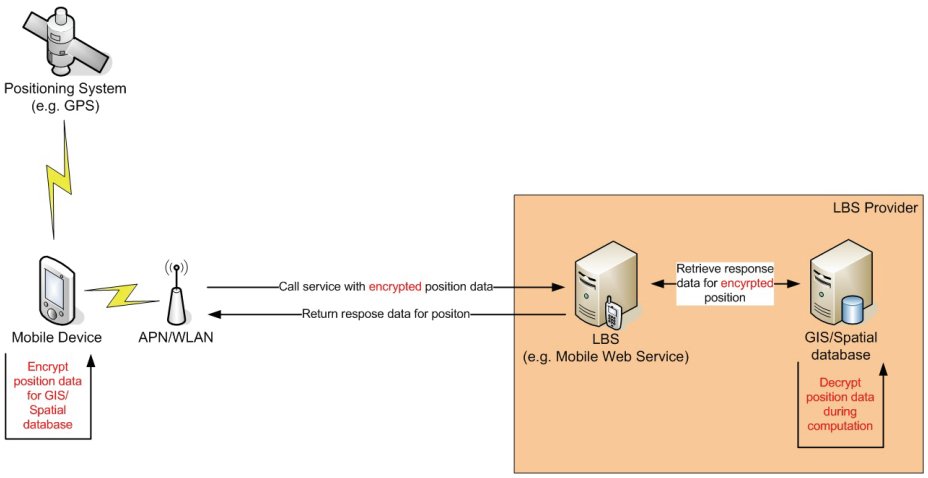


Figure 4: Privacy preservation using end-to-end encryption of location information

To be able to trust the LBS provider that such a spatial database (or other component) is really used, a certification model could be used where independent organizations certify a service to comply with the regulations.
But the key benefit of this solution is, that it works even better for mashed up LBSs, which become increasingly popular. In those services results from multiple LBS providers are combined into a single end user response. The mashup LBS acts as a proxy shielding the actual requester from the targeted LBSs; thus the advantages of the TTP approach are gained automatically and the certification as above would not be required at all. Similar to the analysis without mashup the end-to.end security of the location information is achieved, since it is only revealed in encrypted form. On the other hand the suggested approach also has some issues that need to be investigated further:

- There is a need for a pretty complex transaction protocol in the mashup scenario, especially in cases where session keys for all basic LBS and encryption of the service responses is required.
- Is a mashup service with such a complex transaction protocol worth the implementation effort for a service provider? What can be the application areas and business models?
- If the mashup really aggregates the results of the individual services to a comprehensive response, then only segment-wise encryption of the responses is possible. This is due to the fact that the mashup needs the responses in clear text in order to be able to perform the aggregation in general.
- Due to the many encryption/decryption steps required during protocol execution and the slow hardware of mobile devices there is need for extensive evaluation and testing. This is required to make sure that an acceptable response time of the services is possible in this scenario.

The suggested solution is only a blueprint architecture so far. Details of the required protocols for the different possible scenarios still remain to be defined.

4. Conclusion and Future Work

In this article we have discussed existing work on how to protect user privacy of LBS for mobile devices. All the solutions presented so far lack the combination of satisfying the following properties:

- Optimal quality of service, esp. with respect to the quality of the response, is achieved.
- The current location of the mobile device is known only by the client.
- The model is applicable to chargeable services and/or services with registration.
- The model should work independently of the number of service users.

Therefore in this article we sketched a different model, which works based on an end-to-end encryption of the location information. It is only decrypted at runtime by the component computing the service response. The advantage of that model is that it provides all the aforementioned properties together. In addition it is suitable for service mashups and even achieves advantages of the TTP model in that case.
This article presents only a study of related work and a rough sketch of our proposed solution. Before the proposed solution could be used for real-world services there is a lot that remains to be done. Firstly all the required components and protocols have to be specified in detail, both for the basic and the mashup case. Thereafter a prototypical implementation would be necessary in order to proof that the concept may in fact be implemented on current devices. Moreover benchmark tests with recent mobile devices and also backend systems would be required in order to guarantee that user-friendly services could really be achieved. Also a detailed security analysis would be required in order to proof the desired properties of the system. An extension to

push services or integration with other approaches for push services such as in [15] would also be an interesting topic for future work. Finally in the case where the whole system has been shown to be applicable extensions to other services that do not only require the current position but also other input as well as general geo services should be developed.

5. References

[1] BAMBA, BHUVAN, LING LIU, PETER PESTI und TING WANG: Supporting anonymous location queries in mobile environments with privacy grid. In: WWW'08: Proceeding of the 17th international conference on World Wide Web, p. 237–246, New York, NY, USA, 2008. ACM.
[2] BARKHUUS, LOUISE und ANIND K. DEY: Location-Based Services for Mobile Telephony: a Study of Users' Privacy Concerns. In: RAUTERBERG, MATTHIAS, MARINO MENOZZI und JANET WESSON (Herausgeber): INTERACT. IOS Press, 2003.
[3] BERESFORD, ALASTAIR R. und FRANK STAJANO: Mix Zones: User Privacy in Location-aware Services. In: PerCom Workshops, p. 127–131. IEEE Computer Society, 2004.
[4] BETTINI, CLAUDIO, XIAOYANG SEAN WANG und SUSHIL JAJODIA: Protecting Privacy Against Location-Based Personal Identification. In: JONKER, WILLEM und MILAN PETKOVIC (Herausgeber): Secure Data Management, Band 3674 der Reihe Lecture Notes in Computer Science, p. 185–199. Springer, 2005.
[5] CAMENISCH, JAN und ELS VAN HERREWEGHEN: Design and implementation of the idemix anonymous credential system. In: ATLURI, VIJAYALAKSHMI (Herausgeber): ACM Conference on Computer and Communications Security, p. 21–30. ACM, 2002.
[6] CANDEBAT, THIBAULT, CAMERON ROSS DUNNE und DAVID T. GRAY: Pseudonym management using mediated identity-based cryptography. In: DIM '05: Proceedings of the 2005 workshop on Digital identity management, p. 1–10, New York, NY, USA, 2005. ACM.
[7] CHOW, CHI-YIN, MOHAMED F. MOKBEL und XUAN LIU: A peer-to-peer spatial cloaking algorithm for anonymous location-based service. In: GIS '06: Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems, p. 171–178, New York, NY, USA, 2006. ACM.
[8] DECKER, MICHAEL: Location Privacy – An Overview –. In: Proceedings of the 7th Int. Conference on Mobile Business (ICMB), 2008.
[9] GEDIK, BUGRA und LING LIU: Location Privacy in Mobile Systems: A Personalized Anonymization Model. In: ICDCS '05: Proceedings of the 25th IEEE International Conference on Distributed Computing Systems, p. 620–629, Washington, DC, USA, 2005. IEEE Computer Society.
[10] GHINITA, GABRIEL, PANOS KALNIS, ALI KHOSHGOZARAN, CYRUS SHAHABI und KIAN-LEE TAN: Private queries in location based services: anonymizers are not necessary. In: SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, p. 121–132, New York, NY,
USA, 2008. ACM.
[11] GHINITA, GABRIEL, PANOS KALNIS und SPIROS SKIADOPOULOS: PRIVE: anonymous location-based queries in distributed mobile systems. In: WWW '07: Proceedings of the 16th international conference onWorldWideWeb, p. 371– 380, New York, NY, USA, 2007. ACM.
[12] GRUTESER, MARCO und DIRK GRUNWALD: Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking. In: MobiSys '03: Proceedings of the 1st international conference on Mobile systems, applications and services, p. 31–42, New York, NY, USA, 2003. ACM.
[13] JORNS, OLIVER, GERALD QUIRCHMAYR und OLIVER JUNG: A privacy enhancing mechanism based on pseudonyms for identity protection in location-based services. In: ACSW '07: Proceedings of the fifth Australasian symposium on ACSW frontiers, p. 133–142, Darlinghurst, Australia, 2007. Australian Computer Society, Inc.
[14] JUNGLAS, IRIS A. und CHRISTIANE SPITZMÜLLER: A Research Model for Studying Privacy Concerns Pertaining to Location-Based Services. In: HICSS. IEEE Computer Society, 2005.
[15] KÖLSCH, TOBIAS, LOTHAR FRITSCH, MARKULF KOHLWEISS und DOGAN KESDOGAN: Privacy for Profitable Location Based Services. In: HUTTER, DIETER und MARKUS ULLMANN: Security in Pervasive Computing, Second International Conference, SPC 2005, p. 164–178, LNCS 3450, Springer, 2005.
[16] MOKBEL, MOHAMED F., CHI-YIN CHOW und WALID G. AREF: The new Casper: query processing for location services without compromising privacy. In: VLDB '06: Proceedings of the 32nd international conference on Very large databases, p. 763–774. VLDB Endowment, 2006.
[17] XU, TOBY und YING CAI: Location anonymity in continuous location-based services. In: GIS '07: Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems, p. 1–8, ACM 2007.
[18] ZIBUSCHKA, JAN, LOTHAR FRITSCH, MIKE RADMACHER, TOBIAS SCHERNER und KAI RANNENBERG: Enabling Privacy of Real-Life LBS. In: VENTER, HEIN S., MARIKI M. ELOFF, LES LABUSCHAGNE, JAN H. P. ELOFF und ROSSOUW VON SOLMS (Herausgeber): SEC, Band 232 der Reihe IFIP, p. 325– 336. Springer, 2007.

Automated User Feedback Generation in the Software Development of Mobile Applications

Christian Schueller

UnternehmerTUM GmbH

Lichtenbergstr. 8, 85748 Garching

Wolfgang Woerndl

Technische Universität München

Boltzmannstr. 3, 85748 Garching

Abstract

In addition to software quality and tariff structure, user experience has a decisive influence on the market success of mobile applications. Thus, it is necessary to get valid feedback from future users during the software development process. However, studying the generation of feedback regarding usability and user experience is still at the beginning and is currently an immature discipline in computer science.

Therefore, we present a concept for the generation of feedback in the software life cycle of mobile applications. A fundamental approach is the implicit and explicit collection of user feedback. This is achieved by an automated recording of user activities (e.g. key-events) in the appropriate context of use, as well as an *in situ* collection of explicit feedback through questionnaires. This way collected data is made available to the developers on a webpage for further development and enhancement of mobile applications. For evaluating and assessing this concept, a prototype implementation has been realized, followed by an evaluation study to analyze the possibilities and limitations of the developed model and the associated tools.

1. Introduction

As shown in previous studies, user experience and usability have, in addition to the software quality and tariff structure, a decisive influence on the market success of mobile applications [1]. Therefore it is necessary to verify mobile applications during the development with regard to objective achievement. The necessity to generate feedback from future application users arises from this requirement, rather than carrying out expert evaluations. Unlike stationary or web-based applications, only a few examples and empirical values exist on how mobile applications could be designed successfully. However, feedback generation to evaluate mobile usability and the experience still represents an immature discipline in computer science [2] and is highly time-consuming and costly [3].

As part of this work we generated deeper insights how feedback generation can be made systematically available to the application developers in a user-centered process model, as well as how suitable tools should look like. On the basis of existing concepts and tools, we developed our own approach to generate feedback in the software life cycle of mobile applications. A fundamental approach is the automated collection of user feedback in the appropriate context of use. This was realized by automated recording of user activities (e.g. key-events) and collection of explicit feedback through questionnaires. To reduce expenditure of time and costs, the feedback generation component was integrated into the already existing framework *play.tools* [4].

First, we introduce our "*play.tools*" framework for mobile applications. In section 3 we discuss how to integrate feedback generation into the development of mobile applications from a more theoretical perspective. Then, we describe the design and implementation of our automated feedback generation system. In section 5, we finally discuss related work and conclude our paper.

2. Background

2.1 Developing mobile applications

More and more humans cannot imagine life without a mobile phone, whether on a private or business level. These phones are not only used for voice transmission, but also mobile applications like location-based services or m-commerce applications are being deployed at an increasing rate. The development of marketable and innovative mobile applications represents a huge challenge for software developers. The causes for this are, among other things, unknown user/customer needs, insufficient standards, incorrect programming interfaces and extremely short product life cycles regarding mobile devices and the corresponding software [5]. Particularly the access to system near functions of mobile devices, the positioning and the development of intuitively operated user interfaces are still very complex due to the absence of appropriate software frameworks, which encapsulate these functionalities suitably.

Therefore, the framework *play.tools* was created to reduce the complexity during prototype implementation and subsequent market tests for context-sensitive mobile applications. Thereby, significant generic components of a context-sensitive mobile application (user interface, position determination, supply of geographical information, persistence, etc.) are entirely enclosed in a modular architecture. The developers do not have to worry about technical details of generic functions of context-sensitive mobile applications. They can completely concentrate on the development of the software components.

A further problem of mobile applications is the acquisition of feedback by market tests and finding interesting users – with regard to their context – for software evaluation. Existing platforms for mobile applications such as http://www.handango.com supply hundreds and thousands of different programs, but offer no functionality to select test users for developers. For that purpose, a distribution platform with a recommender based deployment server was developed. By this means, interested users can find relevant applications more easily [6].

2.2 The *play.tools* framework

Figure 1 depicts a high level overview of the distribution platform and the interaction of the main components in our existing *play.tools* framework. Application programmers design and implement client modules and server applications (services) by using the provided generic modules. These applications may communicate with other services and/or use a database to store items. Developers then register their applications with the deployment server (step "1" in Fig. 1). Users can access the deployment server directly and download client modules on

their mobile devices or they use the special client application of the deployment server. The deployment client recommends interesting applications to users depending on their profile and context. Typically, the client program interacts with the server to provide a service, e.g. a mobile tourist guide. Information about users (profiles) is kept separately from the services; thereby personal information can be reused for different applications and is kept under the control of the user (privacy).

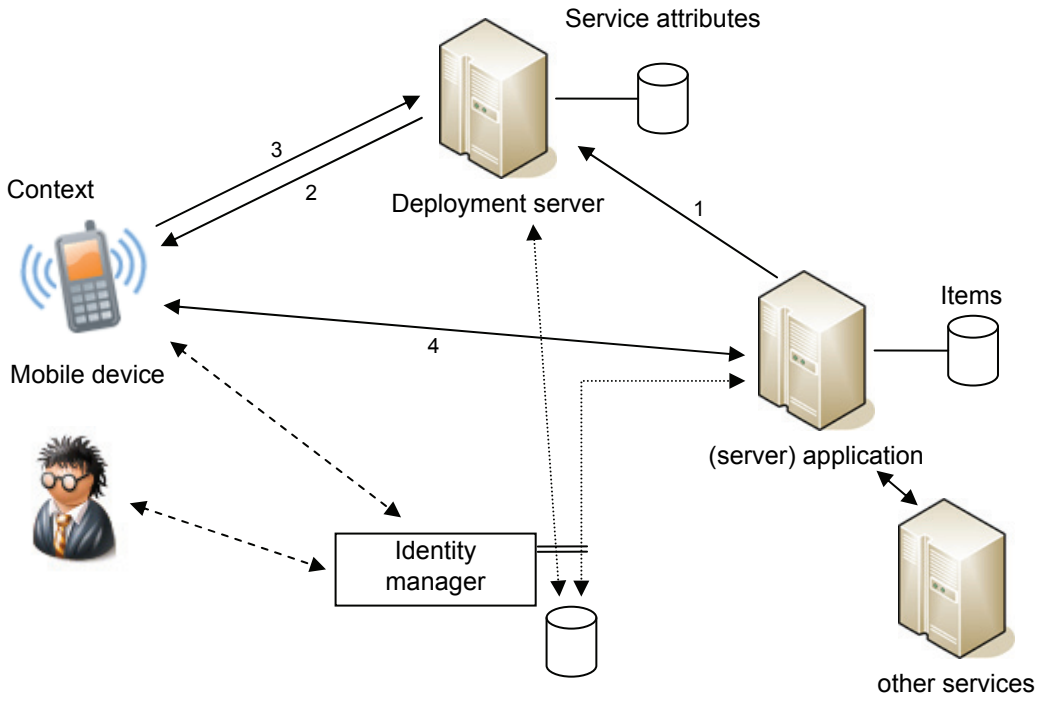


Figure 1: *play.tools* architecture overview

As motivated by the introduction and background, the goal of this work is to integrate an automated feedback generation into the software development process of mobile applications to improve usability and user experience.

3. Feedback generation

Before describing the design and implementation of our feedback system in section 4, we will discuss the automated feedback generation more generally. Evaluation plays an important role in user-centered development of mobile applications [7]. However, there are no systematic approaches [8] and guidelines [9] on how to conduct field evaluations within tight budget limitations to generate user-feedback in the context of use. In addition, hardly any tools exist to facilitate the complex *in situ* data collection [2] with the involvement of users.

Hence, previous work had the goal to adapt the usability lab to the peculiarities of the evaluation of mobile applications. Special cameras record the screen as well as the facial expressions of test subjects to capture their emotions. The remaining problem is the lack of context. Kjeldskov and Stage tried to simulate the context in a laboratory environment with their "*new techniques for usability evaluation of mobile systems"* [10] by using treadmills. The test person has to walk on a treadmill while interacting with a mobile application, to emulate a sightseeing walk. Nielson *et al*. showed that context simulation is not adequate to generate valid user feedback despite inevitable high costs and effort field tests. Their concluding summary was: "It's definitely worth the hassle" [2].

For field testing extensive equipment like cameras and batteries, plus an accompanying person are until now necessary [11]. As in the usability lab, cameras record the interaction and emotions of the subject. The accompanying person documents noteworthy observations. Back in the office, the generated data is synchronized and evaluated. Reichl *et al.* tried to reduce the effort and the impact of the additional equipment by minimizing the used technology [12]. They created the wearable lab environment: a special hat, where the required mini cameras are mounted. Through an air interface the video data is transferred to the accompanying person and

saved on a laptop. Additionally, special events are reported manually. In doing so, problems like the influence of the test person through the monitoring process and the extra attention remains unsolved [13].

Therefore, the procedure presented in this work does not use any additional equipment and the test person is not accompanied by an observer. Instead, the feedback is implicitly generated by the use of the application itself, and explicitly with the help of questionnaires. Hardware sensors as well as software sensors contribute to the implicit feedback generation. Embedded hardware sensors in the mobile devices record the system status, mainly in which context the application is used. The software sensors are implemented into the source code and register all pressed keys and triggered events by the user. All data generated through the use of mobile devices and applications are saved in log-files. In the explicit feedback generation, the user is encouraged to answer questions directly after a special event occurs. For example, a user will be asked after completing an activity – like searching a restaurant – if the function met their expectations. Altogether, we use the following methods for feedback generation:

- Implicit feedback, especially for the detection of usability problems
  - Hardware sensors (GPS/GSM-position, speed, signal strength, acceleration, device information, etc.)
  - Software sensors (Key- and event-sensors, user actions, action-based screenshots, etc.)
- Explicit feedback, particularly to identify the user experience
  - Questionnaires to detect intentions, feelings, thinking of users
  - Context-triggered questions

The collected data is processed and made available as information to the developers in a web interface, enabling a continued improvement of the mobile application. In addition to a simple visualization of the recorded data – using tables and charts –, the data is also interpreted. Dependencies found, can be visualized with heat maps as in the following example (figure 2).

Figure 2: Heat maps: graphical representation of a two-dimensional data

Finally, the developer can interpret the information obtained from user behavior analysis and the identified problems users had while using the application, to enhance the mobile application in the course of the development process.

Once a sufficient amount of data is collected with the automated feedback generation, heuristics and style guides can be established to avoid mistakes in the concept development of future applications.

4. System design, implementation and test

In this section we will show how the conceptual ideas of the last chapter can be realized in concrete software architecture. Thereby, we must map the process of generating feedback to a set of software components.

4.1 Design and implementation

The design of the feedback system should fulfill the following requirements: (A) Developers should be able to collect a rich variety of information on the use of their mobile applications and functions. (B) The system should be scalable to be able to collect large amounts of data over long periods. (C) Allow a simple addition of new data collection capabilities. (D) Information should be collected using actual data on the natural environment of

the user. (E) Protection and backup of the collected data in the field is also important for the feedback generation system. (F) And last but not least, the utilization of the CPU should be kept low.

To enable the use of the automated feedback generation for developers in a fast and simple way, the feedback generation was added as an additional generic component to the *play.tools* framework. Like the other *play.tools* components, the feedback generation component can be configured by using an XML-file. If a developer builds an application with the help of *play.tools* (see sec. 2), the user interaction and the context of use is automatically logged with no extra effort. The developer can also choose from a predefined set of questions or additionally create their own questions for explicit feedback generation.

On a higher level of detail, the real test begins when the developer registers their application at the *play.tools* deployment server, after having implemented it with generic *play.tools* components. Users, whose profiles and context fit the application's deployment parameters, get a recommendation to test it. Once the user accepts the recommendation, the application is installed (through the Java based OTA-mechanism) on their mobile device. When the user starts the application, a Context Logger [14] helps to capture the context of use, by recording data from the hardware sensors. An Event Logger logs the user pressed buttons, selected options and used functions. In addition, the Event Logger attaches a timestamp, measures duration and percentage of completed entry fields.

The recorded data will be temporarily stored on the device until the next synchronization with the server. After the data is transmitted to the server, it will be analyzed and interpreted with the goal to detect usage pattern. Geographical data about the location of use will be visualized in maps. The graphic treatment includes the generation of high-level context (e.g. the tester is sitting or running), with the help of the Context Analyzer [14]. Even the session length and the variance of the process time were edited. The design and implementation of our system is highly extensible, so it is easily possible to integrate additional sensors and new learning classifiers, for example.

4.2 Tests

For evaluating and assessing this concept, a partial implementation has been realized and finally evaluated in a study to allow statements on the possibilities and limitations of the developed model and the associated tools. A first evaluation with ten test users has shown that logging alone is not sufficient for a meaningful assessment of usability.

Compared to the conventional usability test (thinking-aloud), only one third of the errors were found with the help of logging alone. A critical issue was to record what the user could not do in addition to the information about what he or she actually has done. At this point, explicit feedback could improve the results. Since this test was performed in a usability lab, explicit feedback was not collected in the first study. So we will continue to test the system in future evaluations in order to identify weaknesses and to maximize the benefit to the process of innovative mobile applications creation.

However, for the current version detailed domain knowledge is still needed from the developers, to indicate the presented information appropriately. Altogether, automation is a useful addition to existing evaluation methods – like heuristic evaluation – to make valid statements on the usability and user experience. Furthermore applications can be evaluated in spite of a tight budget. It is also possible to do further analysis, such as target group analyses, on the collected data.

5. Conclusion

In this paper, we have presented the generation of automated user feedback in the domain of mobile devices. Firstly we have discussed automated feedback generation more generally, and then described the design and implementation of our ideas. The evaluation showed that our approach can produce meaningful results which can be utilized in mobile applications. We plan to do further iterations and evaluations, in order to maximize the assistance for software developers.

A similar approach is used by Froelich *et al*. [15]. Contextual usage data is collected to learn more about the user's behavior in general and not on the usability of a mobile application in specific. For example, they explore situations in which mobile phones are charged. There are some tools, like the ContextLogger [16], that capture

the context in which an application is running. However, this does not record the user interaction with the application itself. Moreover, a range of computerized experience-sampling tools have been developed to elicit user response in the field [17], [18], [19] but none provide an extensible framework for combining automatic logging and user sampling on a participant's personal device.

6. References

[1] Schmidt, A. *Entwicklung mobiler Anwendungssysteme - Grundlagen, Konzepte, Modelle.* Saarbrücken: VDM Verlag Dr. Müller, 2007. (in German)
[2] Nielsen, C. M.; Overgaard, M.; Pedersen, M. B.; Stage, J. and Stenild, J. "It's worth the hassle!: the added value of evaluating the usability of mobile systems in the field." *In: Proceedings of the 4th Nordic conference on Human-computer interaction.* pp. 272-280. ACM Press, 2006.
[3] McDonald, S.; Monahan, K.; and Cockton, G. "Modified contextual design as a field evaluation method." *In: Proceedings of the 4th Nordic conference on Human-computer interaction.* pp. 437-440. ACM Press, 2006.
[5] Okhrin, I. and Richter, K. "Mobile Business: Framework, Business Applications and Practical Implementation in Logistics Companies", *In: Tech Report Mobile Internet Business*, no. 1, November 2005.
[4] Schueller, C.; Doll, B.; Woerndl, W. "Play.Tools: Ein Software-Framework zur prototypischen Umsetzung kontextsensitiver mobiler Anwendung als Unterstützung von Innovationsprozessen." *In: Proceedings of the 2nd conference on Mobilitaet und Mobile Informationssysteme*, 2007. (in German)
[6] Woerndl, W.; Schueller, C.; Wojtech, R. "A Hybrid Recommender System for Context-aware Recommendations of Mobile Applications." *In: Proceedings of the IEEE 3rd International Workshop on Web Personalisation, Recommender Systems and Intelligent User Interfaces*. pp. 871-878. IEEE Computer Science, 2007.
[7] Rogers, Y., et al. "Why It's Worth the Hassle: The Value of In-Situ Studies When Designing Ubicomp." *In: Proceedings of the 9th International Conference Ubiquitous Computing*. pp. 336-353. LNCS 4717. Springer, 2007.
[8] Krauß, M. and Krannich, D. "ripcord: rapid interface prototyping for cordless devices." *In: Proceedings of the 8th International Conference on Human Computer Interaction with Mobile Devices and Services.* pp. 187-190. ACM Press, 2006.
[9] Pousttchi, K. and Thurnher, B. "Understanding Effects and Determinants of Mobile Support Tools: A Usability-Centered Field Study on IT Service Technicians." *In: Proceedings of the 4th International Conference on Mobile Business.* pp. 10-10. IEEE Computer Society, 2006.
[10] Kjeldskov, J. and Stage, J. "New Techniques for Usability Evaluation of Mobile Systems." *In: International Journal of Human Computer Studies*. vol. 60, no. 5-6, pp. 599-620. Academic Press, 2004.
[11] Roto et al. "Examining Mobile Phone Use in the Wild with Quasi-Experimentation." *In: HIIT Tech Report.* no. 1, August 13, 2004 .
[12] Reichl et al. "The LiLiPUT Prototype: A Wearable Lab Environment for User Tests of Mobile Telecommunication Applications." *In: Proceedings of Computer Human Interaction Conference.* pp. 1833-1838. ACM Press, 2007.
[13] Kjeldskov, J. and Skov, M. B. "Creating a Realistic Laboratory Setting: A Comparative Study of Three Think-Aloud Usability Evaluations of a Mobile System." *In: Proceedings of the 9th IFIP TC13 International Conference on Human Computer Interaction.* pp. 663-670. IOS Press, 2003.
[14] Woerndl, W.; Schueller, C.; Rottach, T. "Generating high level context from sensor data for mobile applications." *In: Proceedings of the International Conference Wireless Applications and Computing.* 2007.
[15] Froehlich, J.; Chen, M. Y.; Consolvo, S.; Harrison, B. and Landay, J. A. "MyExperience: A System for In situ Tracing and Capturing of User Feedback on Mobile Phones." *In: Proceedings of the International Conference On Mobile Systems, Applications And Services.* pp. 57-70. ACM Press, 2007.
[16] Raento, M.; Oulasvirta, A.; Petit, R. and Toivonen, H. "ContextPhone: a prototyping platform for context-aware mobile applications." *In: Pervasive Computing, IEEE, vol. 4, no. 2,* pp. 51- 59, Jan.-March. 2005.
[17] Barrett, L.F. and Barrett, D.J. "An Introduction to Computerized Experience Sampling in Psychology." *In: Social Science Computer Review, vol. 19, no. 2, S01, pp. 175-185.* Sage Publications, 2001.
[18] Carter, S.; Mankoff, J. and Heer, J. "Momento: Support for Situated Ubicomp Experimentation." *In: Proceedings of the ACM Conference on Human Factors in Computing Systems.* ACM Press, 2007.
[19] Consolvo, S. and Walker, M. "Using the Experience Sampling Method to Evaluate Ubicomp Applications." *In: IEEE Pervasive Computing Mobile and Ubiquitous Systems: The Human Experience, vol. 2, no. 2,* pp. 24-31. Apr-Jun. 2003.

Ein kontextbezogener Instant-Messaging-Dienst auf Basis des XMPP-Protokolls

F. Dürr, J. Palauro, L. Geiger, R. Lange, K. Rothermel

Institut für Parallele und Verteilte Systeme
Universitätsstraße 38
70569 Stuttgart
GERMANY

E-Mail: {nachname}@ipvs.uni-stuttgart.de

Abstract

Dieser Beitrag beschreibt die Verteilung kontextbezogener Informationen mittels eines erweiterten Instant-Messaging-Dienstes. Dieser Dienst ermöglicht das Senden von Nachrichten an alle Teilnehmer, die einen bestimmten Kontext besitzen und sich z.B. an einem bestimmten Ort aufhalten oder bestimmte Interessen besitzen. Als Basis dient das Extensible Messaging and Presence Protokoll (XMPP) sowie ein XMPP-basierter Instant-Messaging-Dienst. Wir beschreiben Protokoll- und Architekturerweiterungen für die Integration von Kontextinformationen in das XMPP-Protokoll und die Server-Infrastruktur. Der erweiterte Dienst unterstützt insbesondere die Server-seitige Filterung von Nachrichten aufgrund von Kontextinformationen und ermöglicht dadurch die effiziente Nachrichtenverteilung.

1. Einleitung

In diesem Artikel wird die Verteilung kontextbezogener Informationen mit Hilfe eines neuartigen kontextbezogenen Instant-Messaging-Dienstes betrachtet. Ziel dieses Dienstes ist die Implementierung des Contextcast-Kommunikationsparadigmas. Contextcast ermöglicht die Zustellung von Nachrichten an alle Nutzer, die einen bestimmten Kontext besitzen. Das heißt, im Gegensatz zu aktuellen Instant-Messaging-Diensten, bei denen die Empfänger einer Nachricht explizit durch ihren Benutzer-ID adressiert werden, soll es dieser kontextbezogene Instant-Messaging-Dienst ermöglichen, Benutzer implizit aufgrund ihres Kontexts zu adressieren. Der Sender muss dabei die Identität der Empfänger nicht kennen – im Unterschied zu herkömmlichen Instant-Messaging-Diensten, bei denen die Empfänger den Sendern in Form so genannter Buddy-Listen bekannt sind.

Der Kontext, der zur Adressierung eines Benutzers verwendet werden kann, ist dabei vielfältiger Natur. Eine wichtige Kontextinformation ist beispielsweise die aktuelle Position des Benutzers, an der er sich aufhält. Durch Berücksichtigung dieser Information können Nachrichten an alle Personen zugestellt werden, die sich an einem bestimmten Ort aufhalten. Die reine ortsbezogene Adressierung als eine Unterart der allgemeinen kontextbezogenen Adressierung wird dabei auch als Geocast bezeichnet. Neben dem Ort sind aber eine Vielzahl weiterer Benutzerkontexte für die Adressierung nutzbar, beispielsweise das verwendete Fortbewegungsmittel (Autofahrer, Fußgänger, usw.), sein Alter, sein Geschlecht, seine Interessen oder allgemein die Situation, in der sich der Benutzer gerade befindet.

Die Anwendungsmöglichkeiten von Contextcast sind dabei so vielfältig wie die Kontextinformationen, die zur Adressierung verwendet werden können. Typische Beispiele umfassen die Verteilung von Warnmeldungen an alle Gefährdeten in einem bestimmten geographischen Gebiet, oder die Zustellung von Informationen zu Sehenswürdigkeiten an alle Touristen mit bestimmten Interessen. Denkbar ist auch die selektive Verteilung von Veranstaltungsinformationen oder von Produktinformationen.

Die Motivation für die Verwendung von Instant-Messaging-Diensten für die Implementierung der Contextcast-Kommunikation ergibt sich aus zwei aktuellen Trends. Zum einen erfreuen sich entsprechende Dienste sehr hoher Beliebtheit. Sie sind somit praktisch universell verfügbar und bieten bekannte und akzeptierte Benutzungsschnittstellen. Zum anderen ermöglichen aktuelle Technologien zur mobilen Kommunikation die Möglichkeit, Instant-Messaging-Dienste auch auf mobilen Endgeräten zur Verfügung zu stellen. Entsprechende Kommunikationstechnologien wie drahtlose lokale Netze (WLAN) oder UMTS ermöglichen hohe Übertragungsraten und werden durch Flatrate-Tarife immer attraktiver.

Neben diesen Technologietrends wird die Verwendung von Instant-Messaging für die Implementierung von Contexctast aber auch durch die technischen Eigenschaften der Instant-Messaging-Protokoll motiviert. Wie später noch im Detail ausgeführt wird, ermöglichen diese Protokolle die für Contextcast typische Push-Kommunikation. Die Klienten müssen also nicht wie bei diversen anderen Protokollen wie RSS über HTTP ständig aktiv Änderungen abfragen (engl. polling), sondern sie bleiben passiv und werden bei Verfügbarkeit von entsprechenden Informationen durch die Instant-Messaging-Infrastruktur informiert. Ferner steht eine Infrastruktur von Instant-Messaging-Servern zur Verfügung, die Informationen aufgrund der Nutzerkontexts filtern können bevor sie an den Klienten zugestellt wird, im Gegensatz beispielsweise zu RSS wo ein substantieller Anteil der Filterung beim Endgerät erfolgt. Die infrastrukturbasierte Filterung entlastet somit die Endgeräte und die (drahtlose) Kommunikationsschnittstelle.

Der Rest dieses Beitrages ist wie folgt strukturiert. In Abschnitt 2 gehen wir auf die Erweiterung des XMPP-Instant-Messaging-Protokolls um Contextcast-Mechanismen ein. Abschnitt 3 stellt die Arbeit in den Kontext verwandter Arbeiten, bevor dieser Beitrag in Abschnitt 4 mit einer Zusammenfassung und einem Ausblick auf zukünftige Arbeiten abgeschlossen wird.

2. Erweiterung des XMPP-Protokolls und der XMPP-Infrastruktur um kontextbezogene Adressierungskonzepte

In diesem Abschnitt beschreiben wir die Erweiterungen des Extensible Messaging and Presence Protocols (XMPP) um kontextbezogene Adressierungsmechanismen. XMPP ist ein offenes Standardprotokoll für Instant-Messaging. Die Offenheit und das XMPP zugrunde liegende XML-basierte Nachrichtenformat ermöglichen eine einfache Erweiterung dieses Protokolls. Ferner wird dieses Protokoll von vielen Dienstanbietern unterstützt und es existieren diverse offene XMPP-Klienten und Server, die als Grundlage für die Erweiterung um kontextbezogene Konzepte dienen können. In dieser Arbeit verwenden wir den XMPP-Klienten Spark [1] und den XMPP-Server Openfire [2] als Grundlage für unsere Erweiterungen.

Zunächst geben wir in diesem Abschnitt einen Überblick über die erweiterte Funktionalität der kontextbezogenen Nachrichtenzustellung und das zugrunde liegende Konzept zur kontextbezogenen Adressierung von Nachrichten und zur Definition von Benutzerkontexten. Gefolgt wird dieser Abschnitt durch die Beschreibung der erweiterten verteilten Systemarchitektur der XMPP-Server-Infrastruktur und der lokalen Erweiterungen des XMPP-Servers Openfire.

2.1. Erweiterte Funktionalität und kontextbezogenes Adressierungskonzept

Die Möglichkeit zur kontextbezogene Adressierung von Teilnehmern soll für die Implementierung der folgenden Funktionen genutzt werden:

Kontextbezogene Nachrichten: Hierbei handelt es sich um einzelne Nachrichten, die an Teilnehmer zugestellt werden, die einen bestimmten Kontext besitzen. Der Sender kennt hierbei die Identitäten der Empfänger nicht. Nachrichten besitzen dabei eine bestimmte Lebenszeit. Besitzt eine Teilnehmer zum Sendezeitpunkt der Nachricht den adressierten Kontext, oder ändert er während der Lebenszeit der Nachricht seinen Kontext, so dass er dem adressierten Kontext entspricht, so wird die Nachricht an ihn zugestellt.

Kontextbezogener Chat: Im Gegensatz zu einzelnen kontextbezogenen Nachrichten, die nur vom Sender zum Empfänger vermittelt werden, besteht beim kontextbezogenen Chat die Möglichkeit, Chat-Sitzungen mit Benutzern zu initiieren, die einen bestimmten Kontext besitzen. Die Teilnehmer, welche den durch die Chat-Einladung des Initiators adressieren Kontext besitzen, können dem Chat beitrete und bidirektional Nachrichten mit dem Initiator austauschen.

Kontextbezogene Newsletter: Beim kontextbezogenen Newsletter abonnieren Teilnehmer aufgrund ihrer Interessen bestimmte existierende Themengebiete, beispielsweise Nachrichten zum Thema Rock-Musik. Nachrichten werden aufgrund der Abonnements und weitere Kontextinformationen an die Teilnehmer zugestellt. So kann ein Sender z.B. eine Nachricht über ein Rock-Konzert an alle Rock-Musik-Newsletter-Abonnenten versenden, die sich gerade im Umkreis des Veranstaltungsorts aufhalten.

Zur Adressierung von kontextbezogenen Nachrichten, Chat-Einladungen und Newslettern sowie zur Definition der Teilnehmerkontexte können verschiedene Kontextattribute eingesetzt werden. Grundsätzlich handelt es sich dabei um Schlüssel-Wert-Paare, die vom Empfänger und dem Sender frei definiert werden können. So definieren beispielsweise die Paare „alter = 30“ und „position = /de/stuttgart“ das Alter und den aktuellen Aufenthaltsort einer Person. Hierdurch können beliebige Kontextinformationen spezifiziert werden, so dass jeder Anwendungsentwickler den Kontext wählen kann, der für die entsprechende Anwendung relevant ist. Andererseits hat eine völlig freie Definition den Nachteil, dass die Kompatibilität verschiedener Anwendungen schwer zu garantiert ist. Wir gehen daher davon aus, dass die verwendbaren Kontextinformationen in entsprechenden Schemata definiert sind [3]. Allgemeine, anwendungsübergreifende Kontextattribute werden von einem Standardschema festgelegt, das jede Anwendung interpretieren kann; anwendungsspezifische Erweiterungen werden in erweiterten Schemata definiert, die u.U. nur von bestimmten Anwendungen verstanden werden.

Insbesondere Ortsinformationen sind für mobile Anwendungen von großer Bedeutung. Aufgrund der für den Benutzer intuitiven Verständlichkeit werden Ortsinformationen in Form hierarchischer symbolischer Lokationen spezifiziert. So steht beispielsweise die symbolische Lokation /de/stuttgart für die Stadt Stuttgart in Deutschland. Wie in Abschnitt 2.2 beschrieben wird, werden intern auch geometrische Koordinaten verwendet, die über ein entsprechendes hybrides Lokationsmodell [4] in symbolische Lokationen übersetzt werden.

2.2. Erweiterte XMPP-Systemarchitektur

Das Instant-Messaging-System besteht im Wesentlichen aus drei Hauptkomponenten (siehe Abbildung 1).

Der *Instant-Messaging-Klient* stellt die Benutzungsschnittstelle dar und wird für das Senden und Empfangen von Nachrichten verwendet. Wir verwenden als Grundlage den XMPP-Klienten Spark [1], der durch entsprechende Komponenten erweitert wurde (diese Komponenten sind in der Abbildung durch „Conny“ – dem Namen des Projekts, das diese Erweiterungen entworfen hat – bezeichnet). Jeder mobile Klient ist mit einem oder mehreren Positionierungssystemen ausgestattet. Integriert wurden das satellitengestützte GPS-System und das WLAN-basierte Positionierungssystem der Firma Ekahau. Beide Systeme geben Positionen in Form von geometrischen Koordinaten aus (Längen-Breitengrad bzw. Koordinaten in einem kartesischen Referenzsystem).

Jeder Klient ist mit dem *Instant-Messaging-Server* seines Anbieters verbunden. Die Server haben die Aufgabe, Nachrichten vom Sender zu den Empfängern der Nachricht weiterzuleiten. Entsprechend dem XMPP-Protokoll übergibt der Klient des Senders hierzu die Nachricht an seinen lokalen Server. Dieser Server leitet die Nach-

richt an alle Server weiter, bei denen Empfänger mit entsprechenden Kontexten registriert sind, die wiederum die Nachricht an ihre lokalen Klienten mit dem adressierten Kontext zustellen. Wir setzen als Grundlage für unsere Implementierung den XMPP-Server Openfire ein [2].

Die Server besitzen Verbindungen zu einem Dienst, der das Kontextmodell verwaltet. Insbesondere verwaltet dieser Dienst das oben angesprochene hybride Lokationsmodell. Die Instant-Messaging-Komponenten nutzen dieses Modell zur Ermittlung der symbolischen Lokationshierarchie, die dann dem Benutzer im Klienten angezeigt wird, sowie zur Umrechnung geometrischer Koordinaten in symbolische Lokationen. Als *Umgebungsmodelldienst* setzen wir auf der Nexus-Infrastruktur auf [5]. Diese Infrastruktur ermöglicht die verteilte und skalierbare Verwaltung von Umgebungsmodelldaten. Einzelne Teilmodelle von verschiedenen Anbietern werden jeweils durch so genannten Spatial-Model-Server bereitgestellt, wobei es sich hierbei um ein offenes System handelt, bei dem jeder Anbieter seine Daten in das Gesamtsystem integrieren kann. Somit wird es möglich, globale Umgebungsmodelle bereitzustellen. Durch die Föderation der Teilmodelle werden die Teilmodelle den Instant-Messaging-Servern als ein durchgängiges Modell präsentiert.

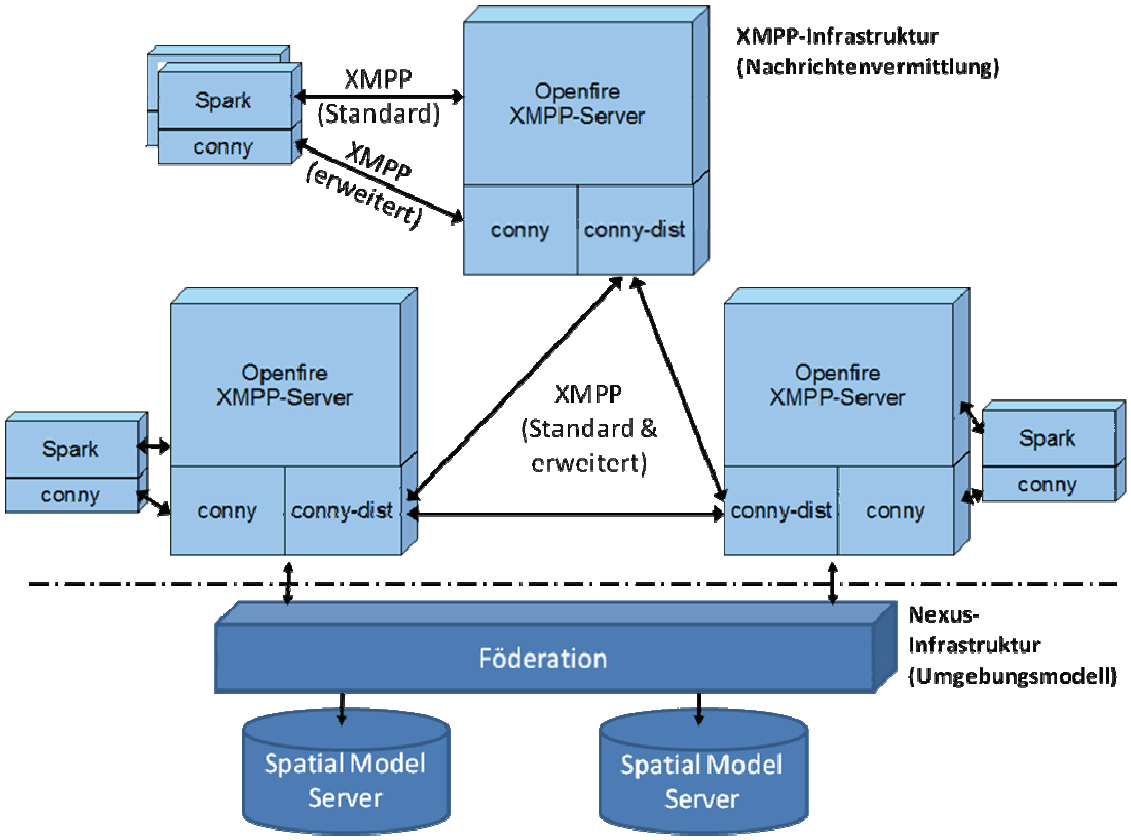


Abbildung 1: Systemarchitektur

2.3. Kontextbezogene Nachrichtenvermittlung

Die kontextbezogene Nachrichtenvermittlung erweitert die Standard-XMPP-Kommunikation. Das heißt, das System unterstützt neben der Weiterleitung kontextbezogener Nachrichten auch weiterhin herkömmliches Instant-Messaging. Die Erweiterungen gliedern sich in zwei Teile: erstens die erweiterte Klient/Server-Kommunikation, zweitens die Server/Server-Kommunikation.

Die Klient/Server-Kommunikation wurde um verschiedene Nachrichtentypen erweitert. So wurden zum einen Nachrichten eingeführt, um den Kontext eines Klienten, d.h. seine Kontextattribute einschließlich Ort, sowie seine Abonnements von kontextbezogenen Newslettern, an den Server zu übermitteln. Der Server verwaltet diese Informationen und verwendet sie für die Server-seitige Nachrichtenfilterung. Das heißt, wenn ein Server eine kontextbezogene Nachricht erhält, dann stellt er diese nur an diejenigen Klienten zu, deren Kontext dem adressierten Kontext entspricht. Für die verschiedenen Funktionen (kontextbezogene Nachricht, Newsletter und Chat-Einladung) wurden entsprechende XMPP-Nachrichtenformate definiert. Beispielhaft soll hier eine einfache kontextbezogene Nachricht dargestellt werden:

```
<message id="B97rz-72" to="conny.bar.com" from="foo@bar.com">
  <connymessage type="message" sender="foo@bar.com"
   creationdate="1206455896030" deliverdate="1206455896030"
   lifetime="2100000">
    <context>
      <location>/europa/deutschland/bw/stuttgart</location>
      <profile>
        <value key="Studiengang" value="Softwaretechnik"/>
      </profile>
    </context>
    <content>
      <subject>A Test Message</subject>
      <text>A simple test message.</text>
    </content>
  </connymessage>
</message>
```

Diese Nachricht definiert zum einen die bereits angesprochene Zeitspanne der Nachrichtauslieferung durch die Attribute creationdate und lifetime. Ferner wird hier ein einfacher Kontext adressiert, bestehend aus der Position der Empfänger (location) und einem weiteren Kontextattribut (Studiengang). Die XMPP-Nachricht ist an die Komponente conny des XMPP-Servers mit der Adresse bar.com adressiert. Ein entsprechend erweiterter XMPP-Server wird die Nachricht anstatt sie direkt an einen bestimmten Benutzer weiterzuleiten an diese Komponente übergeben, welche die kontextbezogene Nachrichtenfilterung und –weiterleitung implementiert.

Die Server/Server-Kommunikation wird primär für die Vermittlung von Nachrichten an die Server entfernter Empfänger eingesetzt, die nicht beim Server des Senders registriert sind. In der aktuellen Implementierung bilden die Server hierzu ein unstrukturiertes Netz. Nachrichten werden in diesem Netz durch Fluten (engl. flooding) weitergeleitet, d.h. der Server sendet eine weiterzuleitende Nachricht einfach an alle mit ihm verbundenen Server. Es findet also aktuell keine Filterung bei der Weiterleitung im Netz statt, sondern nur eine Server-seitige Filterung am „Rande“ des Netzes. Obwohl diese Server-seitige Filterung die Klient/Server-Verbindung entlastet, so wird doch das Netz u.U. stark belastet. Dieses einfache Verfahren erfüllt aber eine wichtige Anforderung an das System. Die Kontextinformationen der Teilnehmer werden nicht an andere fremde Server weitergeleitet, sondern verbleiben auf den als vertrauenswürdig angenommenen Servern des jeweiligen Anbieters des Teilnehmers, was einen wichtigen Vorteil bezüglich der Privatheit von Teilnehmerkontexten darstellt.

3. Verwandte Arbeiten

Heutige Web-Standards konzentrieren sich auf den Zugriff auf ortsbezogene Informationen. Hervorzuheben ist hier insbesondere der GeoRSS-Standard [6], mit dem es möglich ist, Informationen mit Orten zu verknüpfen. Obwohl RSS grundsätzlich die Idee einer Push-Kommunikation verfolgt – der Benutzer wird informiert, sobald sich relevante Änderungen ergeben – so ist die technische Umsetzung der Kommunikation über HTTP oft Pull-basiert realisiert. Wir argumentieren daher, dass ein Push-basiertes Protokoll wie XMPP für die Verbreitung orts- und kontextbezogener Informationen deutlich besser geeigneter ist.

Die orts- und kontextbezogene Vermittlung von Informationen mittels spezieller Overlay-Netze ist derzeit aktiver Forschungsgegenstand [7, 8, 9]. Diese Arbeit kann von diesen Ansätzen profitieren, indem entsprechende Vermittlungsprotokolle im Netz aus XMPP-Servern implementiert werden. Hierdurch könnte die präsentierte Server-seitige Filterung am Rande des Netzes zu einer netzseitigen Filterung erweitert werden.

4. Zusammenfassung und zukünftige Arbeiten

In diesem Beitrag wurde Erweiterungen eines Instant-Messaging-Dienstes zur kontextbezogenen Vermittlung von Nachrichten vorgeschlagen. Durch diese Erweiterungen ist es möglich, Nachrichten an Benutzer zu senden, die sich in einem bestimmten Kontext, z.B. an einem bestimmten Ort, befinden, mit diesen Chat-Sitzungen aufzubauen oder kontextbezogene Newsletter zu abonnieren. Ermöglicht wurde diese kontextbezogene Kommunikation durch Erweiterung des XMPP-Protokolls, die in die XMPP-Klienten- und Server-Software Spark und Openfire integriert wurden. Durch die für XMPP typische Push-basierte Kommunikation und die Server-seitige Filterung in der XMPP-Infrastruktur konnte eine effiziente Nachrichtenverteilung implementiert werden, die insbesondere die Klienten und deren meist drahtlosen Kommunikationsverbindungen entlastet. Gleichzeitig stellt das Vermittlungsverfahren sicher, dass teilnehmerspezifische private Kontextinformationen

beim vertrauenswürdigen Server des Teilnehmers verbleiben und nicht an andere Server und Teilnehmer weitergeleitet werden.

In zukünftigen Arbeiten soll vor allem die Effizienz der Nachrichtenverteilung weiter gesteigert werden. Hierzu sind spezielle kontextbasierte Vermittlungsalgorithmen in das XMPP-Server-Netz zu integrieren, wodurch sich eine Nachrichtenfilterung im Netz anstatt an eine Server-seitige Filterung an den Rändern des Netzes ergibt. Hierzu können beispielsweise existierende Geocast-Protokolle integriert werden. Da diese Protokolle allerdings nur eine effiziente Vermittlung auf reinen Ortsinformationen ermöglichen, die nur einen Teil der relevanten Empfängerkontexte darstellen, ist es unser Ziel, spezielle Contextcast-Protokolle zu entwerfen und zu integrieren, um eine feingranulare Filterung aufgrund allgemeiner Kontextinformationen zu ermöglichen. Eine wesentliche Herausforderung besteht dabei darin, einerseits die Privatheit der Kontextdaten sicherzustellen und teilnehmerspezifische Kontextdaten nicht beliebig im Netz zu verteilen, andererseits aber auch mit den weitergegebenen Informationen einen effiziente Nachrichtenfilterung zu ermöglichen.

Denkbar sind auch Erweiterungen der zur Adressierung eingesetzten Kontextinformationen beispielsweise um zeitbezogene Kontexte, um hiermit gezielt Teilnehmer anzusprechen, die sich z.B. zu einem bestimmten Zeitpunkt an einem bestimmten Ort aufgehalten haben.

Danksagungen

Unser Dank gilt vor allem den Teilnehmern des Studienprojekts Advanced Instant Messaging Service – Sören Brunk, Hannes Mühleisen, Jonas Palauro, Andreas Poszlovszki, Michael Schäfer, Johannes Schneider, Katharina Wiesner und Alexander Wobser –, die das vorgestellte Systems umgesetzt haben.

Diese Arbeit wurde teilweise durch die Deutsche Forschungsgemeinschaft (DFG) innerhalb des Sonderforschungsbereichs 627 unterstützt.

Referenzen

[1] http://www.igniterealtime.org/projects/spark/index.jsp
[2] http://www.igniterealtime.org/projects/openfire/index.jsp
[3] M. Bauer, F. Dürr, J. Geiger, M. Grossmann, N. Hönle, J. Joswig, D. Nicklas, T. Schwarz: „Information Management and Exchange in the Nexus Platform“, Technischer Bericht Nr. 2004/04.
[4] F. Dürr, K. Rothermel: „On a location model for fine-grained geocast“. UbiComp 2003: 5th International conference on ubiquitous computing, 2003.
[5] D. Nicklas, M. Großmann, T. Schwarz, B. Mitschang: „A Model-Based, Open Architecture for Mobile, Spatially Aware Applications“. Proceedings of the 7th International Symposium on Spatial and Temporal Databases, 2001.
[6] C. Reed, R. Singh, R. Lake, J. Lieberman, M. Maron: „An introduction to georss: A standards based approach for geo-enabling rss feeds“. Open Geospatial Consortium Inc., White Paper OGC 06-050r3, 2006. http://www.opengeospatial.org/pt/06-050r3
[7] F. Dürr, C. Becker, K. Rothermel: „An Overlay Network for Forwarding Symbolically Addressed Geocast Messages“. Proceedings of the 15th International Conference on Computer Communications and Networks (ICCCN ’06), 2006.
[8] L. Geiger, F. Dürr: „Kontextbezogene Kommunikation“. 4. GI/ITG KuVS Fachgespräch Ortsbezogene Anwendungen und Dienste, 2007.
[9] G. Cugola, J. E. M. de Cote: „On Introducing Location Awareness in Publish-Subscribe Middleware“. ICDCSW '05: Proceedings of the 4th International Workshop on Distributed Event-Based Systems (DEBS) (ICDCSW'05), 2005.

Ortsbezogene, probabilistische Aufgabenverteilung am Beispiel von Sensornetzen

Gerhard Fuchs

Universität Erlangen Nürnberg
Lehrstuhl Informatik 7

Martensstr. 3
91058 Erlangen

Kurzfassung

Die Komplexität heutiger Systeme nimmt zu. Einzelne Systeme bestehen aus immer mehr Teilsystemen, die zusammen für einen Anwender eine Aufgabe lösen sollen. Wie können komplexe zusammengesetzte Systeme koordiniert werden? Ein Aspekt, der in diesem Kontext behandelt werden muss ist die Aufgabenverteilung, also die Frage „Wer übernimmt eine Aufgabe?“.
Roboter Sensornetze (RSNW) bestehen aus sehr vielen einzelnen Robotern und Sensorknoten, die miteinander Aufgaben für einen Anwender ausführen sollen. Abstrakt gesehen sind sie ein Gesamtsystem, das aus mehren Teilsystemen (Knoten) besteht.
In diesem Artikel wird am Beispiel der RSNW die Ortsbezogene Probabilistische Aufgabenverteilung (OPA) vorgestellt. Ein System, welches für einen Anwender eine Aufgabe lösen soll, erhält so lange vom Anwender einen ortsabhängigen Stimulus, bis die Aufgabe erledigt ist, oder der Anwender aufgibt. Jeder Knoten, der diesen Stimulus erhält entscheidet autonom, wie er mit der Anfrage umgeht. Er kann die Aufgabe je nach Begabung (deterministisch), Motivation (zufällig) und Kommunikationsfreudigkeit (zufällig), ausführen, weiterleiten, oder verwerfen. Bei diesem Mechanismus kann nicht garantiert werden, dass eine Aufgabe überhaupt von einem Knoten angenommen wird.
Das Prinzip ist durch Studien von zell- und entwicklungsbiologischen Vorgängen inspiriert. Die Forschungen stehen am Anfang, ein kurzer Ausblick auf weitere Forschungsaspekte wird gegeben.

1. Einleitung

Ein sich derzeit abzeichnender Trend ist, dass die durchschnittliche Anzahl der Computer pro Person steigt. Computer sind hierbei nicht nur der PC unter dem Schreibtisch, auch andere alltägliche Gegenstände wie z.B. das Auto, die Wachmaschine oder der Herd sind mit Prozessoren versehen. Hinzu kommt, dass diese Geräte, um zusätzliche Funktionalität zu erzielen, miteinander vernetzt werden – Stichwort „Internet der Dinge“ [1], [2]. Eine Person ist je nach Aufenthaltsort von immer anderen Computern umgeben so dass ein weiterer Trend dahin geht, den Ort und den Kontext bei Diensten mit zu berücksichtigen. Die Komplexität der Systeme, die für eine Person zuständig sind nimmt zu und es gilt sicher zu stellen, dass die Geräte auch im Sinne des Anwenders arbeiten.
Am Lehrstuhl Informatik 7 (Rechnernetze und Kommunikationssysteme) beschäftigen wir uns u.a. mit heterogenen Roboter Sensornetzen (RSNW), die ein Verbund aus mehreren Robotern und Sensorknoten sind. Ein Sensorknoten enthält in der Regel die vier Hauptbaugruppen CPU, Funk, Batterie, und Sensor [3]. Ist ein Aktuator (z.B. Fahrwerk) angeschlossen sprechen wir von einem Roboter.
Wie werden komplexe Systeme koordiniert? In diesem Zusammenhang müssen drei wesentliche Fragestellungen geklärt werden (Abbildung 1): **Wer** übernimmt eine Aufgabe (**Aufgabenverteilung**)? **Wie** ist eine Aufgabe zu lösen (**Ablaufbeschreibung** z.B. [4], [5]). **Was** ist für eine Aufgabe zu lösen (**Aufgabebeschreibung**)?

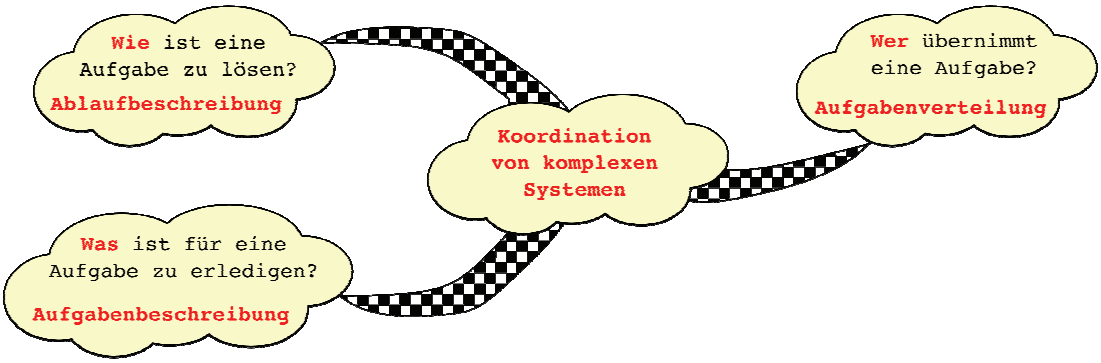


Abbildung 1: Fragestellungen bei der Koordination von komplexen Systemen.

Der Nachfolgende Artikel beschreibt die ortsbezogene, probabilistische Aufgabenverteilung OPA und gliedert sich wie folgt: Abschnitt 2 schildert am Beispiel eines RSNW ein mögliches Szenario bei dem die OPA verwendet werden kann. Das Prinzip der OPA steht in Abschnitt 3, die konkrete theoretische Umsetzung in Abschnitt 4. Variationsmöglichkeiten, die bei der OPA in Hinblick auf Ortsbezogenheit berücksichtigt werden müssen, stehen in Abschnitt 5. In Abschnitt 6 wird der Artikel zusammengefasst und ein Ausblick auf weitere mögliche Forschungsarbeiten gegeben. Die Forschungen sind am Anfang, erste grundlegende studentische Arbeiten auf diesem Gebiet [6], [7] sind bereits abgeschlossen.

2. Mögliches Szenario

Gegeben ist ein heterogenes RSNW, welches auf einer Fläche ausgebracht worden ist. Die Roboter und Sensorknoten sind unterschiedlich mit Sensoren bestückt, einige besitzen Temperaturfühler. Ein Anwender betritt mit seinem Handy dieses Gebiet und will über eine Funkschnittstelle (z.B. Bluetooth) Kontakt zu dem RSNW aufnehmen, um die aktuelle Temperatur zu erfragen. Diese Anfrage („Gib mir die aktuelle Temperatur“) erreicht eine gewisse Teilmenge an Robotern und Sensorknoten (nachfolgend „Knoten“ genannt) und ist für diese ein **Stimulus**. Ein Knoten, der die Anfrage erhält entscheidet als Erstes, ob er die Aufgabe prinzipiell lösen kann, hier z.B. ob ein Temperaturfühler angeschlossen ist. Ist dies der Fall – der Knoten ist **begabt** – hat der Knoten die Möglichkeit zu entscheiden, ob er die Aufgabe **ausführt**, oder nicht. Ist der Knoten **motiviert**, misst er die Temperatur und schickt das Ergebnis an den Anwender zurück. Ist der Knoten nicht **begabt**, oder nicht **motiviert**, muss er entschieden, ob er die Anfrage an seine Nachbarn **weiterleitet**, oder komplett **verwirft** (ist der Knoten **kommunikativ**?). Bei der OPA entscheidet der Knoten zufällig, ob er **motiviert / kommunikativ** ist.
Da eine Antwort des RSNW nicht garantiert werden kann, muss der **Stimulus** ggf. wiederholt werden. Kommen mehrere Antworten zurück, kann der Anwender z.B. einen Mittelwert bilden und das Ergebnis hat eine höhere Qualität. Kommt keine Antwort zurück, muss der Anwender entscheiden ob er **aufgibt**.

3. Prinzip

In diesem Abschnitt ist das grundlegende Prinzip der OPA stickpunktartig wiedergegeben:

- Ein System kann sich aus mehreren Teilsystemen (ebenfalls Systeme) zusammensetzen.
- Ein System ist prinzipiell inaktiv.
- Jedes System befindet sich an einer adressierbaren Position im Raum (von Außen betrachtet).
- Die Position eines Systems entscheidet darüber, welche Aufgabe es ausführen soll.
- Ein System erschließt die Position aus den vorhandenen Signalen.
- Die Signale sind nur eine gewisse Zeit gültig, d.h. sie haben eine Halbwertszeit und müssen bei Bedarf erneuert werden.
- Ist eine spezielle Signalmischung an der Position des Systems vorhanden, reagiert das System.
- Die Reaktion des Signals erfolgt probabilistisch.

Bei diesem Mechanismus kann nicht garantiert werden, dass eine Aufgabe überhaupt von einem Knoten angenommen wird. Im Vergleich zu MURDOCH [8] oder OOA [9] wird auf einen ausgezeichneten Koordinator verzichtet. Die Idee zu dem in diesem Artikel vorgestellten Prinzip ist durch das Studium von zellbiologischen, speziell entwicklungsbiologischen Vorgängen entstanden.

4. Umsetzung

3.1. Abstraktion

Abbildung 2 zeigt am Beispiel eines Sensornetzes die für die OPA verwendete Abstraktion. Weitere Informationen zu der nachfolgend beschrieben Abstraktion, sowie weiterführende Definitionen sind in [10] beschrieben. Ein Sensornetz besteht aus einzelnen Sensorknoten, die sich für einen außenstehenden Betrachter an einer eindeutig adressierbaren Position im Raum befinden. Die Abstraktion hiervon ist das Gesamtsystem (GeS), das als Gitternetz in dem sich die (Teil-) Systeme (Knoten) befinden dargestellt wird. Ein Knoten ist als Ellipse eingezeichnet und ist beim Beispiel die Abstraktion eines Sensorknotens, die Adressierung erfolgt über Koordinaten.

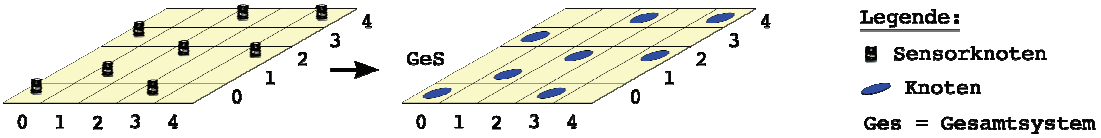


Abbildung 2: Abstrakte Darstellung eines Gesamtsystems am Beispiel eines Sensornetzwerks.

3.2. Ablauf

Abbildung 3 zeigt den Ablauf der OPA aus der Perspektive eines außenstehenden Beobachters. Eine Anfrage muss nicht immer erfolgreich verlaufen.

1) Start eines Aufgabenzyklus: Das Gesamtsystem ist bereit und wartet auf einen Stimulus.
2) Stimulus: Es gibt eine Aufgabe zu erledigen (z.B.: Anwender fragt das RSNW nach der Temperatur).
3) Reaktion: Das System arbeitet.
4) Resultat: Das Ergebnis wird ggf. sichtbar.
5) Test: Ist das Ergebnis OK? Wenn „ja“, dann weiter mit (6); Wenn „nein“, dann weiter mit (2). Der Aufgabensteller kann die Anfrage unverrichteter Dinge „aufgeben“ müssen, da das System unzureichend reagiert, oder er das Ergebnis nicht mehr benötigt, d.h. weiter mit (6).
6) Ende eines Aufgabenzyklus: Die Aufgabe ist erfolgreich durchgeführt, oder aufgegeben worden.

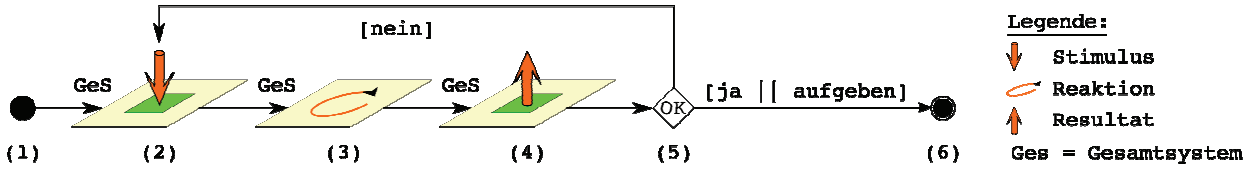


Abbildung 3: Ablauf der ortsbezogenen, probabilistischen Aufgabenverteilung (Sicht von Außen).

3.3. Konkretisierung

Ablauf einer Abstrakten Aktion: Der im Abschnitt 3.2 dargestellte Ablauf ist von Außen betrachtet eine Aktion, die zur Erledigung einer Aufgabe ausgeführt werden muss. Diese Aktion kann in zwei Teilaktionen aufgeteilt werden (Code in Abbildung 4). Der erste Schritt beim Ausführen (Aufruf von `go()`, Z. 10) ist das Festsetzen der Grenzen (Z. 11). Bei der OPA ist das die Ausbreitung des Stimulus, beim RSNW z.B. die Sende-Reichweite der Bluetooth-Schnittstelle des Handys des Anwenders. Anschließend nehmen die empfangenden Geräte die Arbeit auf (Z. 12). Eine abstrakte Aktion kann auch selbst-organisierend realisiert werden.

```
01   public abstract class Aktion {
02      private Aktion begrenzen;      // begrenzende Aktion
03      private Aktion arbeiten;       // kontrollierte Aktion
04
05      public Aktion(Aktion begrenzen, Aktion arbeiten) {
06         this.begrenzen = begrenzen;
07         this.arbeiten  = arbeiten;
08      }
09
10      public final void go() {
11         if(begrenzen != null)  begrenzen.go();  // ggf. festlegen der Grenzen
12         if(arbeiten  != null)  arbeiten.go();   // ggf. ausfuehren der Aktion
13      }
14   }
```

Abbildung 4: Formulierung des Ablaufs einer abstrakten Aktion in JAVA.

Aufteilung des Gesamtsystems: Ein Gesamtsystem (Ges) wird für die analytischen Betrachtungen wie in Abbildung 5.1 dargestellt in ein Knoten- (KnS) und ein Signalsystem (SiS) aufgeteilt. Zwischen den beiden Teilsystemen gibt es einen Zusammenhang der Positionen. Beide Teilsysteme können ein eigenständiges Verhalten haben. An den Positionen können sich, wie in Abbildung 5.2 dargestellt Knoten (Ellipsen) und Signale (Kreuze) befinden. Knoten und Signale sind Systeme. Ein Stimulus ist bei der hier vorgestellten Abstrahierung eine Ansammlung von Signalen, d.h. Kreuzen im Signalsystem (Abbildung 5.3). Das KnS und das SiS beeinflussen sich gegenseitig (Abbildung 5.4). Das KnS kann das SiS beeinflussen, indem es neue Signale generiert (a), das SiS wird vom KnS ausgewertet (b), so dass eine Art Zyklus entsteht.

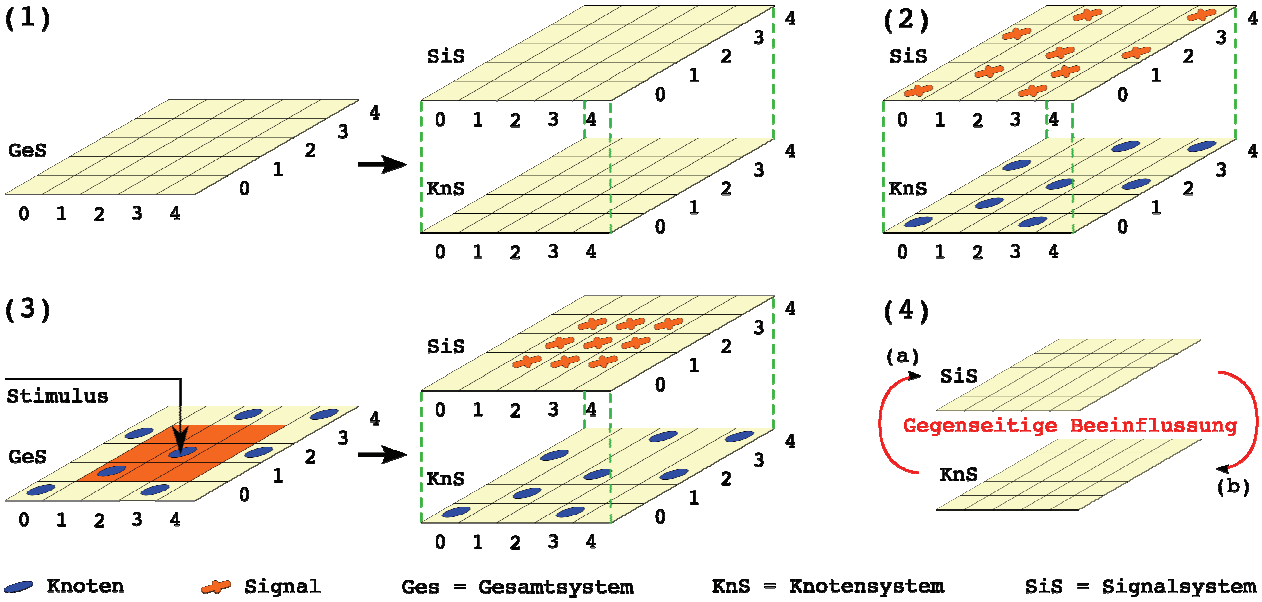


Abbildung 5: Grundlegendes Modell bei der OPA: 1) Aufteilung des GeS in ein KnS und ein SiS; 2) Darstellung von Knoten und Signalen in ihren Teilsystemen; 3) Darstellung eines Stimulus bei der Einwirkung auf das GeS; 4) Zyklische gegenseitige Beeinflussung zwischen KnS und SiS; 4a) KnS beeinflusst das SiS; 4b) SiS beeinflusst das KnS.

Probabilistische Reaktion: Erkennt ein Knoten, dass er eine Aufgabe zu erledigen hat (an der Position des Knotens ist ein entsprechender Stimulus), muss der Knoten darauf reagieren. Die probabilistische Reaktion eines Systems bei Ankunft eines Stimulus wird in Abbildung 6 gezeigt. Für den Knoten gibt es 3 Möglichkeiten: er kann die Aufgabe ausführen, weiterleiten, oder verwerfen. Wie er sich verhält hängt von 3 Entscheidungen ab, die er treffen muss: ist der Knoten begabt, motiviert, und kommunikativ? Die erste Entscheidung hängt von den Fähigkeiten des Knotens und den Anforderungen der Aufgaben an den Ausführenden ab, die letzten beiden sind zufällig, hängen also von Wahrscheinlichkeiten (M, K) ab.

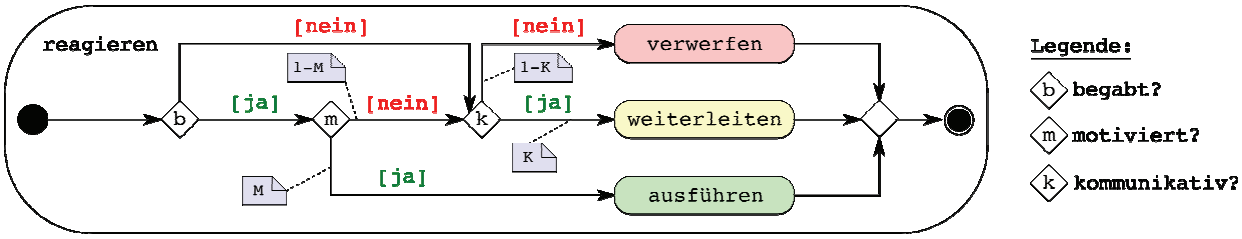


Abbildung 6: Probabilistische Reaktion eines Systems auf einen Stimulus. ***M*** ist die Wahrscheinlichkeit, dass das System motiviert ist die im Stimulus bekannt gegebene Aufgabe abzuarbeiten. ***K*** ist die Wahrscheinlichkeit, dass das System den Stimulus weiter kommuniziert.

5. Ortsabhängige Variationsmöglichkeiten

In Abbildung 7 sind 4 wesentliche ortsabhängige Variationsmöglichkeiten bei der OPA dargestellt. Auf die Art der Knoten und den Inhalt der Signale wird an dieser Stelle nicht eingegangen. Interessante Bereiche (Ausdehnungen) sind grün eingefärbt:

1) Variation des Einzugsbereichs des Knotens. Von welchen Positionen kann ein Knoten Signale empfangen? Von welchen Positionen wird ein Knoten beeinflusst? Betrachtet der Knoten ausschließlich die Signale an seiner Position, oder auch die von anderen?
2) Variation des Wirkungsbereichs des Knotens. An welche Positionen sendet ein Knoten Signale. Welche Positionen kann ein Knoten beeinflussen? Agiert er nur an seiner lokalen Position, oder hat er einen ausgedehnten Wirkungsbereich?
3) Variation der Ausdehnung des Stimulus. An welchen Positionen wird das Gesamtsystem von Außen stimuliert? Gibt es eine einzige, ausgezeichnete Position, oder eine Bereich?
4) Größe des bei der Adressierung verwendeten Rasters. Wie groß ist bei der Betrachtung eine Position im Gesamtsystem? Kann sich ein Knoten oder Signal auf mehreren Positionen gleichzeitig befinden? Können sich mehrere Knoten oder Signale auf einer Position befinden? Diese Aufteilung ist von den Wahrnehmungsmöglichkeiten (Auflösung) des Beobachters abhängig.

Die Beobachtung des zeitlichen Verlaufs bei den oben aufgezählten Punkten ergibt ein weiteres, orthogonales Kriterium. Sind die Kriterien zu jedem Zeitpunkt gültig (statisch), oder veränderbar (dynamisch)? Weiterhin ist zu berücksichtigen, ob die Konten oder Signale die Position wechseln können, d.h. mobil sind.

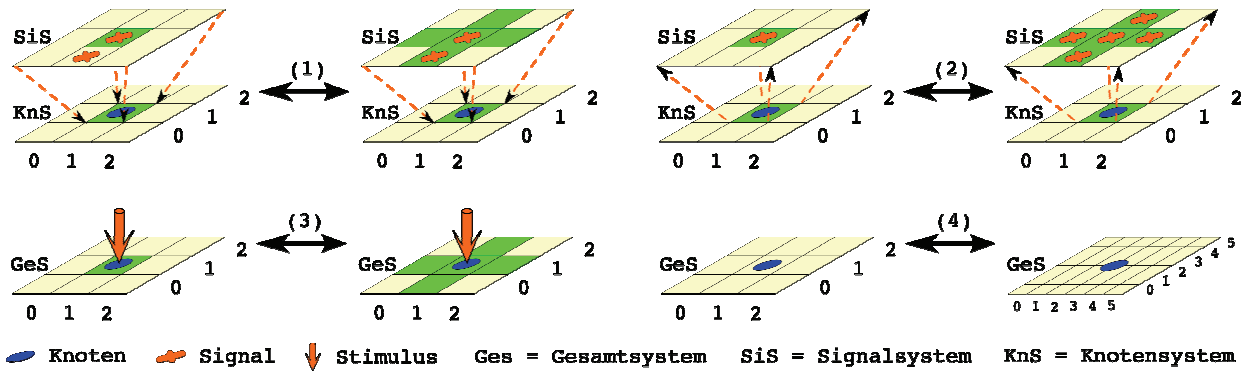


Abbildung 7: Variationsmöglichkeiten bei der OPA: 1) Einzugsbereich des Knotens; 2) Wirkungsbereich des Knotens; 3) Ausdehnung des Stimulus; 4) Größe des bei der Adressierung verwendeten Rasters.

6. Zusammenfassung und Ausblick

In diesem Artikel habe ich, ausgehend von einer kurzen Einordnung in den Kontext der Koordination von komplexen Systemen und einem möglichen Szenario im Bereich Roboter Sensornetze (RSNW), die Grundzüge der Ortsbezogene, Probabilistische Aufgabenverteilung (OPA) vorgestellt. Das Prinzip ist zell- und entwicklungsbiologisch inspiriert und basiert im Kern darin, dass ein System seine auszuführende Aufgabe an Hand der an seiner Position vorhandenen Signalmischung ermittelt. Ist klar, was ausgeführt werden soll, entscheidet das System autonom, wie es mit dieser Information um geht. Der Mechanismus kommt ohne Koordinator aus, die Reaktion ist probabilistisch.
Für die Umsetzung habe ich eine auf [10] basierende Abstraktion vorgestellt. Das Gebiet in dem sich das RSNW befindet, wird in diskrete Bereiche unterteilt, die eindeutig adressierbar sind. Von Außen betrachtet wird ein Stimulus an das System gerichtet, worauf dieses reagiert. Nach der Auswertung des Resultats wird entschieden, ob ein erneuter Stimulus nötig ist. Das System wird so lange „genervt“ bis das Ergebnis den Ansprüchen des Anwenders entspricht, oder dieser aufgibt. Dieses Prinzip ist durch die Beschreibungen des Ablaufs einer abstrakten Aktion, der Aufteilung des Gesamtsystems in ein Knoten- und ein Signalsystem und der probabilistischen Reaktion konkretisiert worden. Im letzten Abschnitt habe ich die ortsabhängigen vier Variationsmöglichkeiten beschrieben. Es handelt sich um den Einzugsbereich / Wirkungsbereich des Knotens, die Ausdehnung des Stimulus, und der Größe des bei der Adressierung verwendeten Rasters.
Gernot Roth hat in seiner Studienarbeit [6] bereits einen ersten prototypischen Entwurf eines ähnlichen Mechanismus mit OMNeT++ (http://www.omnetpp.org) realisiert. Bei dieser Implementierung sind der hohe Kommunikationsaufwand der Konten und die geringe Performance negativ aufgefallen. Positiv waren bei diesem Ansatz die gleichmäßige Verteilung der Aufgaben im Gesamtsystem und die gute Skalierbarkeit.
Hakan Calim hat in seiner Diplomarbeit [7] die Zelldifferenzierung der Fruchtfliege behandelt. Er hat theoretisch untersucht wie ein Sensornetzwerk mit Hilfe der gewonnenen Erkenntnisse ausgebracht werden kann, und die Sensorknoten an Hand der empfangenen Signale ihre Position im Sensornetz und darauf aufbauend eine eindeutige Adressierung herleiten können. An Hand dieser Position kann den Knoten eine Aufgabe zugewiesen werden. Er neue Leistungsmerkmale wie die „Knotendichte“ und darauf aufbauend Dichte basierende Aufgaben definiert.
Weiterer Forschungsbedarf besteht darin, diesen Mechanismus in der Simulationsumgebung SelOnEx [10] und im realen auf SunSPOTs (http://www.sunspotworld.com) basierenden RSNW zu realisieren. Darauf aufbauend sind Kriterien festzulegen und an Hand derer die Vor- und Nachteile gegenüber bereits existierenden Mechanismen zu untersuchen. Mittelfristig erhoffe ich mir eine Möglichkeit zur Koordination von Systemen, die aus sehr vielen Teilsystemen bestehen.

Literatur:

[1] ITU: „The Internet of Things – Executive Summary“; ITU Internet Reports 2005; WSIS; Tunis, Tunesien; 2005.

[2] Bullinger, H.-J. und ten Hompel, M. (Hersg.): „**Internet der Dinge**“; Springer; Berlin, Heidelberg; 2008.

[3] Akyildiz, I. und Su, W. und Sankarasubramaniam, Y. und Cayirci, E.: „**Wireless sensor networks: a survey**“; Elsevier Computer Networks; vol. 38; S. 393-422; 2002.

[4] Fuchs, G.: „**Aufgabenbeschreibung mit UML2-Aktivitätsdiagrammen am Beispiel von Roboter-Sensornetzen**“; Lehrstuhl Informatik 7, Universität Erlangen-Nürnberg, Technischer Bericht 02/2008.

[5] Damm, C.: „**Implementierung und Bewertung eines RDF-basierten Frameworks zur Interpretierung und Ausführung von UML2-Aktivitätsdiagrammen auf Sensorknoten**“; Lehrstuhl Informatik 7, Universität Erlangen-Nürnberg; Diplomarbeit; 2008.

[6] Roth, G.: „**Probabilistische Aufgabenverteilung in autonomen Systemen**“; Lehrstuhl Informatik 7, Universität Erlangen-Nürnberg; Studienarbeit; 2007.

[7] Calim, H.: „**Selbstorganisiertes Sensornetzwerk-Deployment basierend auf biologischer Zelldifferenzierung**“; Lehrstuhl Informatik 7, Universität Erlangen-Nürnberg; Diplomarbeit; 2006.

[8] B. P. Gerkey, B.P.: „**On Multi-Robot Task Allocation**“; Faculty of the Graduate School, University of southern California, Dissertation; 2003.

[9] Martin, D., Cheyer, A. und Moran, D.: „**The Open Agent Architecture: a framework for building distributed software systems**“; Applied Artificial Intelligence; vol. 13; S. 91-128; 1999.

[10] Fuchs, G. und Mika, S.: „**Eine zellbiologisch inspirierte Sichtweise auf selbst-organisierende und emergente Systeme - Die Experimentierumgebung SelOnEx**“; GI/ITG KuVS Fachgespräch Selbstorganisierende, Adaptive, Kontextsensitive verteilte Systeme (SAKS 2008); Wiesbaden, Deutschland; 2008.

Ortsbezogene mobile Dienste zur Verbesserung der Sicherheit bei Großveranstaltungen

Heiko Roßnagel

Institut für Arbeitswissenschaft und Technologiemanagement (IAT)
Universität Stuttgart
Nobelstr. 12, 70569 Stuttgart

heiko.rossnagel@iao.fraunhofer.de

Wolf Engelbach

Institut für Arbeitswissenschaft und Technologiemanagement (IAT)
Universität Stuttgart
Nobelstr. 12, 70569 Stuttgart

Wolf.engelbach@iao.fraunhofer.de

Sandra Frings

Institut für Arbeitswissenschaft und Technologiemanagement (IAT)
Universität Stuttgart
Nobelstr. 12, 70569 Stuttgart

sandra.frings@iao.fraunhofer.de

Abstract

In vielen Metropolregionen finden immer häufiger Großveranstaltungen statt, bei denen enorme Besucherströme bewältigt werden müssen. Durch die steigende Anzahl an Großveranstaltungen und immer kürzer werdenden Vorlaufzeiten werden die Organisation und Durchführung immer komplexer und zeitkritischer. Ortsbezogene mobile Dienste könnten einen wertvollen Beitrag leisten, um diese Probleme zu reduzieren. Mobile Dienste können sowohl bei der Durchführung von Veranstaltungen als auch im Bereich des Notfallmanagements eingesetzt werden. Eine zusätzliche Kombination mit mobilen Mehrwertdiensten wird die Vertrautheit mit dem System verbessern und darüber hinaus eine attraktivere Gestaltung der Veranstaltung für die Teilnehmer ermöglichen.

1. Einführung

In vielen Metropolregionen finden immer häufiger Großveranstaltungen statt, bei denen enorme Besucherströme bewältigt werden müssen. Dabei treten Belastungsspitzen sowohl im öffentlichen Personennahverkehr (ÖPNV) als auch im motorisierten Individualverkehr (mIV) auf, was zu zahlreichen verkehrs- und sicherheitstechnischen Herausforderungen führen kann. Durch die steigende Anzahl an Großveranstaltungen und immer kürzer werdenden Vorlaufzeiten werden die Organisation und Durchführung immer komplexer und zeitkritischer. Schwierigkeiten entstehen beispielsweise durch unzureichenden Informationsaustausch zwischen den Verantwortlichen, fehlende Informationsweitergabe an Fahrgäste, mangelhafte Schulungen des angeworbenen Sicherheitspersonals, knappe Finanzmittel der beteiligten Institutionen, uneinheitliches Datenmanagement der einzelnen Einsatzzentralen, eingeschränkten Informationsaustausch im Krisenfall sowie einer späten Erkennung von Krisenfällen.
Ortsbezogene mobile Dienste könnten einen wertvollen Beitrag leisten, um diese Probleme zu reduzieren. Mobilfunkinfrastrukturen bieten standardisierte drahtlose Kommunikationsdienste in nahezu allen Ländern an und ermöglichen eine schnelle Verbreitung von Informationen [10]. Diese Infrastrukturen könnten sowohl zur Unterstützung bei der Durchführung von Großveranstaltungen als auch für Notfalldienstleistungen, insbesondere unter der Verwendung von ortsbasierten Diensten, genutzt werden. Im Rahmen des vom BMBF geförderten Forschungsprojekts VeRSiert [22], dessen Ziel es ist, eine bessere organisatorische und informationstechnische Vernetzung von Nahverkehrsgesellschaften, Einsatzkräften, Veranstaltern und Fahrgästen zu erreichen, um die Sicherheit im ÖPNV bei Großveranstaltungen zu erhöhen, werden daher auch ortsbezogene mobile Dienste untersucht und entwickelt.

2. Untersuchungsgegenstand

Um ein möglichst aussagekräftiges Spektrum von Veranstaltungen und ungeplanten Ereignissen abdecken zu können, wurden im Projekt mehrere Arbeitsszenarien entwickelt. Diese Szenarien sollen dazu dienen, komplexe Zusammenhänge und vielfältig miteinander verwobene Situationen verständlicher und damit auch beherrschbarer zu machen. Die Szenarien werden durchgängig innerhalb des Projektes für die verschiedenen Aufgabengebiete verwendet. Zur Entwicklung und Bewertung der Szenarien wurden Klassifikationen von Veranstaltungen und ungeplanten Ereignissen verwendet, die in [13] genauer beschrieben werden.
Klassifikationen können Forschern und Praktikern dabei helfen, komplexe Gebiete zu analysieren und zu verstehen. Daher nehmen Klassifikationstechniken wie Typologien und Taxonomien eine wichtige Rolle innerhalb der Forschung ein. Sie ermöglichen es, Komplexität zu reduzieren sowie Ähnlichkeiten und Unterschiede der Untersuchungsobjekte zu identifizieren [2]. Klassifikationen können als Vokabular eines Untersuchungsgebietes dienen und als eine Sammlung von definierten Konstrukten die Grundlage für zukünftige Forschungsaktivitäten bilden [12]. Sie verbessern die Wissensbasis, indem sie es Forschern und Praktikern ermöglichen, die inhärenten Probleme bei der Konzeption und Implementierung von Informationssystemen zu verstehen und zu adressieren [17]. Entsprechend der üblichen sozialwissenschaftlichen Praxis wird in den nächsten beiden Unterabschnitten zunächst jeweils eine Menge von Dimensionen konzeptionell definiert, die dann später anhand von empirischen Beispielen untersucht werden [1]. Dieses Vorgehen ermöglicht es, Interaktionseffekte zu berücksichtigen, die zwischen unterschiedlichen Dimensionen existieren können [25].

2.1. Großveranstaltungen

Innerhalb von VeRSiert wurden Großveranstaltungen anhand ihrer verschiedenen Eigenschaftsausprägungen klassifiziert. Tabelle 1 zeigt einen Überblick über mögliche Dimensionen von Großveranstaltungen und ihre jeweiligen Ausprägungen.

2.2. Ungeplante Ereignisse

Analog zu den Veranstaltungen können auch ungeplante Ereignisse, die während einer Veranstaltung auftreten können, klassifiziert werden. Dabei wird im Projekt VeRSiert bewusst der Begriff eines ungeplanten Ereignisses verwendet. Dies soll verdeutlichen, dass auch Ereignisse berücksichtigt werden, die zwar einen erheblichen Einfluss auf die Durchführung einer Veranstaltung haben, aber nicht durch Begriffe wie Notfall oder gar Katastrophe abgedeckt werden. Tabelle 2 gibt einen Überblick über die Dimensionen und Ausprägungen zur Klassifikation von ungeplanten Ereignissen.

Dimension	Ausprägungen			
Häufigkeit des Auftretens	Regelmäßig wöchentlich	Regelmäßig jährlich	Regelmäßig mit wechselnden Orten	Unregelmäßig
Inhalt der Veranstaltung	Sportveranstaltung (zuschauerorientiert)	Konzert	Volksfest	Kirchentag
	Sportveranstaltung (teilnehmerorientiert)	Staatsbesuch	Demonstration	Messe/ Ausstellung
Örtliche Begrenzung	Fest eingegrenzte Lokation	Mischform	Offene Lokation	
Räumliche Ausdehnung der Veranstaltung	Lokation, Straßenzug	Stadtteile	Stadtgebiet	
Auswirkungen auf den ÖPNV	Lokale Auswirkungen	Regionale Auswirkungen	Überregionale Auswirkungen	
Auswirkungen auf den mIV	Lokale Auswirkungen	Regionale Auswirkungen	Überregionale Auswirkungen	
Dauer der Veranstaltung	Stunden	Eintägig	Mehrere Tage	
Dauer der Planung	Wenige Tage	Bis 1 Monat	Mehrere Monate	1 Jahr und darüber
Anzahl der Besucher	Bis 10.000	Bis 100.000	Über 100.000	
Einzugsgebiet	Örtlich	Städtisch	Regional	National
	International			
Teilnehmerstruktur	Eher ähnlich, unfanatisch	Eher ähnlich, fanatisch	Eher unterschiedlich, unfanatisch	Eher unterschiedlich, fanatisch
Sicherheitseinstufung	Unbedenklich	Einfache Sicherheitsvorkehrungen	Große Sicherheitsvorkehrungen	Nicht vorhersehbar
Finanzierung	Zweckorientierte staatliche oder private Kostenübernahme	Ticketing eines geschlossenen Events	Übertragungsrechte, Sponsoring, Startgelder, Catering, etc.	Misch-Finanzierung
Auswirkungen auf zeitgleiche Veranstaltungen	Keine zeitgleichen Veranstaltungen	Eine zeitgleiche Veranstaltung	Mehrere zeitgleiche Veranstaltungen	
Erfahrung des Veranstalters	Keine	Erfahrungen aus ähnlichen Veranstaltungen	Erfahrungen aus früheren Durchführungen der gleichen Veranstaltung	
Genehmigung	Nicht erforderlich (Privatgelände oder Brauchtum)	Versammlung auf öffentlicher Verkehrsfläche	Kommerzielle Veranstaltung auf öffentlicher Verkehrsfläche	
An- und Abreise	Unmotorisiert	ÖPNV	Organisiert	mIV

Tabelle 1: Dimensionen und Ausprägungen zur Klassifikation von Veranstaltungen

2.3. Arbeitsszenarien

Auf Basis der zuvor beschriebenen Klassifikationen wurden für VeRSiert relevante Szenarien identifiziert, die ein möglichst breites Spektrum an Veranstaltungen und potenziellen ungeplanten Ereignissen abdecken sollen. Insgesamt wurden drei unterschiedliche Veranstaltungen ausgewählt. Bei diesen Veranstaltungen handelt es sich um ein Bundesligaspiel, die Kölner Lichter und den Deutschen Evangelischen Kirchentag. Darüber hinaus wurden mehrere ungeplante Ereignisse betrachtet, die bei diesen Veranstaltungen auftreten können. Tabelle 3 zeigt einen Überblick, welche der ungeplanten Ereignisse in den fiktiven Szenarien bei welcher der Veranstaltungen auftreten. Eine ausführlichere Beschreibung der Arbeitsszenarien findet sich in [13].

3. Ortsbezogene Mobile Dienste im Notfallmanagement

Seit mehreren Jahren gibt es eine Vielzahl von Bestrebungen, Mobilfunkinfrastrukturen für das Notfall- und Katastrophenmanagement zu nutzen. So werden diese Notfalldienstleistungen von Organisationen und Standardisierungsgremien wie dem European Telecommunications Standards Institute (ETSI) mit dem Ziel diskutiert, ein Framework für weltweit interoperable mobilfunkbasierte Notfalldienstleistungen zu schaffen [21]. Darüber hinaus zeigt sich die Europäische Kommission an diesem Thema sehr interessiert und fördert Forschung und Standardisierung in diesem Bereich [8] [15]. Dabei fokussiert sich die europäische Vorgehensweise auf das, was mit der heutigen verfügbaren Technologie erreichbar ist, anstatt Anforderungen zu definieren, die mit heutiger Technologie nicht zu erfüllen sind, wie dies im E911 Projekt geschehen ist [4].

Dimension	Ausprägungen			
Zugänglichkeit	Ungehindert zugänglich	Eingeschränkt zugänglich	Schwer zugänglich	Nahezu unzugänglich
Örtliche Ausdehnung des Ereignisses	Lokal begrenzt	Mittleres Gebiet	Großes Gebiet	
Örtliche Ausdehnung der Auswirkungen	Lokal begrenzt	Mittleres Gebiet	Großes Gebiet	
Anzahl der unmittelbar gefährdeten Personen	Bis 20	20-100	Über 100	
Anzahl betroffener Personen	Bis 100	Bis 1.000	Bis 10.000	Über 10.000
Dringlichkeit der Bedrohung	Nur Sachschäden in geringem Umfang	Personen- und Sachschäden zu erwarten	Schwerverletzte und Tote	
Art des ungeplanten Ereignisses	Vorsätzliche Gefährdung Technischer Defekt	Fahrlässige Gefährdung	Organisatorische Mängel	Höhere Gewalt
Kommunikationsinfrastruktur betroffen	Keine	Festnetzleitungen	Mobilfunkinfrastruktur	
Auswirkungen auf den ÖPNV	Keine	Zeitliche Auswirkungen	Logistische Auswirkungen	
Auswirkungen auf den mIV	Keine	Zeitliche Auswirkungen	Logistische Auswirkungen	
Zeitspanne zwischen Ereigniserkennung und Auswirkungen	Tage	Stunden	Minuten	Keine
Dauer der Auswirkungen	Tage	Stunden	Minuten	
Zeitpunkt des Ereignisses	Vorbereitung	Anreise	Durchführung	Abreise

Tabelle 2: Dimensionen und Ausprägungen zur Klassifikation von ungeplanten Ereignissen

	Bombendrohung	Personenschaden	Defekte U-Bahn / Betriebsstörung	Anschlag	Amoklauf	Fehleinschätzung (Sperrung)	Unwetter
Bundesligaspiel			X		X		
Kölner Lichter		X		X			X
Kirchentag	X					X	

Tabelle 3: Überblick über das Auftreten von ungeplanten Ereignissen bei den Großveranstaltungen

Auch die Regierung der Niederlande hat beschlossen ein mobilfunkgestütztes Katastrophenmanagementsystem landesweit einzuführen [6].
In [28] wurde eine mögliche Kategorisierung der Anforderungen an Katastrophenmanagementsysteme vorgenommen. In [24] wurde untersucht, inwieweit diese Anforderungen durch neue Technologien wie die Mobilfunktechnologie erfüllt werden können. Diese Analyse zeigt, dass Mobilkommunikationsinfrastrukturen anderen Technologien gegenüber in wichtigen Teilbereichen überlegen sind [24]. Einige dieser Vorteile sind: (1) Identifikation und Lokalisierung von Spezialisten, (2) Verschicken speziell zugeschnittener Nachrichten an unterschiedliche Orte und Benutzergruppen, (3) dynamische Benachrichtigungen an Individuen beim Wechsel in unterschiedliche Gefahrenzonen, (4) Messung der Bewegungen von Mobilfunknutzern und (5) Bereitstellung eines Rückkanals für Opfer [24].
Voraussetzungen für den Einsatz einer solchen Technologie sind eine hohe Marktpenetration und Netzabdeckung. Die Marktpenetration von mobilen Endgeräten in Westeuropa lag 2007 zwischen 117% in Schweden und 83% in Frankreich (Bevölkerung / Mobilfunkkunden) [10]. Auch die Netzabdeckung ist in den europäischen industrialisierten Ländern nahezu vollständig gegeben und somit für den Einsatz in Katastrophenschutz geeignet [7]. Selbstverständlich ist ein solches mobilfunkbasiertes Katastrophenmanagementsystem nur so lange von Nutzen, wie die darunter liegende Mobilfunkinfrastruktur funktionstüchtig ist. Viele Katastrophen haben direkte Auswirkungen auf die Kommunikationsinfrastruktur innerhalb des betroffenen Gebietes [18]. Dies ist aber ein Problem, das auf alle vernetzen Katastrophenmanagementsysteme, wie beispielsweise auch Sirenen, zu-

trifft [14]. Nicht alle Katastrophenarten haben auch Auswirkungen auf die Mobilfunkinfrastruktur. Darüber hinaus kann vor einer bevorstehenden Katastrophe gewarnt und evakuiert werden, bevor der Katastrophenfall eintritt. So hätten bei einer rechtzeitigen Warnung vor dem Tsunami in 2004 viele Menschen gerettet werden können [24]. Darüber hinaus ist eine auf dem Mobiltelefon empfangene Nachricht deutlich leichter zu verstehen als beispielsweise Sirenensignale, was einen enormen Vorteil darstellt. Mittels mobiler Benachrichtigungen können Warnungen ermöglicht werden, die ausreichend verständlich und handlungsorientiert sind, sowie eine Bestätigung der Informationen ermöglichen. Dies stellt die ideale Form einer Warnung dar [3].
In [24] und [9] wird ein Katastrophenmanagementsystem auf Basis von GSM vorgestellt. Katastrophenmanager verwenden ein geografisches Informationssystem (GIS), um Aktivitäten wie Warnungen, Ortung und Routing von Opfern zu steuern und zur Kontrolle der Einsatzkräfte [27]. Tritt ein Notfall ein, sendet der Katastrophenmanager mittels Cell Broadcast System (CBS) Warnungen über das Mobilfunknetz an die betroffenen Regionen, um sicherzustellen, dass potenzielle Opfer rechtzeitig gewarnt werden. Anschließend ist er in der Lage, Opfer und Spezialisten mittels des Katastrophenmanagementsystems zu orten. Diese Informationen ermöglichen gezielte und zeitlich optimierte Evakuierungen der betroffenen Gebiete [24]. Dies kann mittels zeitlich abgestimmter Warnungen in unterschiedlichen Gebieten erreicht werden, die eine Überfüllung der Fluchtwege verhindern können. Die Genauigkeit der Ortung kann sich aufgrund unterschiedlicher Ortungsverfahren und Zellgrößen von Funkzelle zu Funkzelle unterscheiden [27]. Dies muss bei der Planung und Ausführung der Evakuierung berücksichtigt werden.
Neben diesen neuen Chancen zu einem fein abgestimmten Katastrophenmanagement, werden ebenfalls neue Missbrauchsmöglichkeiten entstehen [16]. Zum Beispiel ist es möglich, gefälschte SMS-Nachrichten zu erzeugen, die zu gezielter Desinformation genutzt werden können [19]. Auch ist die Effektivität von Warnungen durch ihre Glaubwürdigkeit begrenzt. Wenn die Bevölkerung der Warnung nicht glaubt, wird sie auch keine entsprechenden Vorbereitungen treffen [20]. In [23] werden Sicherheitsanforderungen an ein Katastrophenmanagementsystem identifiziert und Lösungsansätze präsentiert. Selbstverständlich müssen auch Datenschutzaspekte beim Aufbau von mobilfunkbasierten DMS berücksichtigt werden. Daraus resultierende Anforderungen und mögliche Lösungsansätze werden in [9] diskutiert.

4. Ortsbezogene mobile Dienste bei Großveranstaltungen

Mobile Dienste können zwar in Notfallsituationen verantwortlichen Personen bei der Vorbereitung von Evakuierungen, Instruktion und Unterstützung von Einsatzkräften, sowie bei der Lokalisierung von Opfern unterstützen [5]. Eine wesentliche Voraussetzung, um auf Notfälle vorbereitet zu sein, ist allerdings, dass die betroffenen Personen mit dem Notfallsystem vertraut sind und vernünftig auf die Warnsignale ohne Verzögerungen reagieren können [11]. Diese Voraussetzung ist sehr schwer zu erfüllen, wenn das System ausschließlich in Notfällen verwendet wird. Der Erfolg eines Notfallmanagementsystems hängt somit sehr stark von geübten Nutzern ab, die mit den Funktionalitäten der Dienste vertraut sind [26]. Bei einem selten genutzten Notfallmanagementsystem können auch nur eingeschränkte praktische Erfahrungen erwartet werden [16]. Mobile Mehrwertdienste, die die gleiche Infrastruktur nutzen, können die Vertrautheit der Nutzer mit dem System verbessern und gleichzeitig neue Möglichkeiten schaffen, um das Veranstaltungserlebnis für die Teilnehmer attraktiver zu gestalten und die Durchführung der Veranstaltung zu unterstützen. Hierbei kann es sich um Informationsdienste wie Hinweise auf „After-Event-Parties“ oder kontextbezogene Fahrpläne, Transaktionsdienste wie Mobile Ticketing oder Mobile Payment, sowie Kommunikationsdienste wie Friend Finder Dienste oder Mobile Communities handeln. Im Rahmen des Projekts VeRSiert werden daher sowohl mobile Dienste betrachtet, die zur Unterstützung bei der Planung und Durchführung von Großveranstaltungen dienen, als auch der Einsatz von Mobilfunk für das Management von Notfallsituationen untersucht. Ziel ist es, eine einheitliche Plattform zu verwenden, die eine gemeinsame Verwendung der Infrastruktur ermöglicht.

5. Zusammenfassung

Mobile Dienste können einen wertvollen Beitrag leisten, um die immer komplexere und zeitkritischere Organisation und Durchführung von Großveranstaltungen zu erleichtern. Sie können in Notfallsituationen verantwortlichen Personen bei der Vorbereitung von Evakuierungen, Instruktion und Unterstützung von Einsatzkräften, sowie bei der Lokalisierung von Opfern unterstützen. Darüber hinaus können mobile Mehrwertdienste, die die gleiche Infrastruktur nutzen, die Vertrautheit der Nutzer mit dem System verbessern und gleichzeitig das Veranstaltungserlebnis für die Teilnehmer attraktiver gestalten.

6. Literatur

[1] K.D. Bailey, Monothetic and Polythetic Typologies and their Relation to Conceptualization: Measurement and Scaling, American Sociological Review, 38, (1) (1973) 18-33.

[2] K.D. Bailey, Typologies and Taxonomies: An Introduction to Classification Techniques, Sage, Thousand Oaks, CA, USA, 1994.
[3] A. Botterell and R. Addams-Moring, Public Warning in the Networked Age: Open Standards to the rescue? Communications of the ACM, 50, (3) (2007) 59-60.
[4] K. Burke and A. Yasinsac, The ramifications of E911, College of Arts and Science, Florida State University, Tallahassee, Florida, *http://websrv.cs.fsu.edu/research/reports/TR-030404.pdf*, 3. April, 2004.
[5] L. Carver and M. Turoff, Human Computer Interaction: The Human and Computer as a Team in Emergency Management Information Systems, Communications of the ACM, 50, (3) (2007) 33-38.
[6] Cisco Systems, Unique mobile communications solution transforms disaster and emergency response capabilities in the Netherlands, *http://www.cisco.com/web/strategy/docs/gov/Cisco_SVS.pdf*, accessed 2008-02-18.
[7] Coveragemaps.com, GSM European Coverage 2008, *http://www.coveragemaps.com/gsmposter_europe.htm*, accessed 2008-02-18.
[8] Europäische Kommission, Commision Recommendation on the processing of caller location information in electronic communication networks for the purpose of location-enhanced emergency call services, Official Journal of the European Union, Brüssel, *www.emtel.etsi.org/Docs/Recommendation_C(2003)2657.pdf*, 2003.
[9] L. Fritsch and T. Scherner, A Multilaterally Secure, Privacy-Friendly Location-based Service for Disaster Management and Civil Protection, Proceedings of the AICED/ICN 2005, Springer, Berlin, Heidelberg, New York, 2005, pp. 1130-1137.
[11] E. Gruntfest and C. Huber, Status report on flood warning systems in the United States, Environmental Management, 13, (3) (1989) 279-286.
[10] GSMworld, GSM Operators, Coverage Maps and Roaming Information, *www.gsmworld.com/roaming/gsminfo/index.shtml*, accessed 2008-02-07.
[12] A.R. Hevner, S.T. March, J. Park and S. Ram, Design Science in Information Systems Research, MIS Quaterly, 28, (1) (2004) 75-105.
[13] Institut für Arbeitswissenschaften und Technologiemanagement (IAT), VeRSiert - Sicherheit im ÖPNV bei Großveranstaltungen: Klassifikation der unterschiedlichen Aspekte und Arbeitsszenarien, 2008-09-30, 2008.
[14] R.G. Little, Toward More Robust Infrastructure: Observations on Improving the Resilience and Reliability of Critical Systems, Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS 03), IEEE, Hawaii, 2003.
[15] B. Ludden, A. Pickford, J. Medland and H. Johnson, Cgalies final report V1.0, Report on implementation issues related to access to location information by emergency services (E112) in the European Union, *http://ec.europa.eu/environment/civil/pdfdocs/cgaliesfinalreportv1_0.pdf*, 28. Januar, 2002.
[16] B.S. Manoj and A. Hubenko Baker, Communication Challenges in Emergency Response, Communications of the ACM, 50, (3) (2007) 51-53.
[17] S.T. March and G.F. Smith, Design and natural science research on information technology, Decision Support Systems, 15 (1995) 251-266.
[18] D. Mendonca, T. Jefferson and J. Harrald, Collaborative Adhocracies and Mix-and-Match Technologies in Emergency Management: Using the emergent interoperability approach to address unanticipated contingencies during emergency response, Communications of the ACM, 50, (3) (2007) 45-49.
[19] J. Muntermann and H. Roßnagel, Security Issues and Capabilities of Mobile Brokerage Services and Infrastructures, Journal of Information System Security, 2, (1) (2006) 27-43.
[20] E.J. Pinker, An Analysis of Short-Term Responses to Threats of Terrorism, Management Science, 53, (6) (2007) 865-880.
[21] Project Mesa, Mobile Broadband for Public Safety, *http://www.projectmesa.org/*, accessed 2008-07-29.
[22] Projekt VeRSiert, Versiert Homepage, *http://www.versiert.info/*, accessed 2008-08-01.
[23] H. Roßnagel and T. Scherner, Secure Mobile Notifications of Civilians in Case of a Disaster, in: H. Leitold and E. Markatos, Ed., Proceedings of the 10th IFIP Open Conference on Communication and Multimedia Security (IFIP CMS 06), Springer, Berlin Heidelberg, 2006, pp. 33-42.
[24] T. Scherner and L. Fritsch, Notifying Civilians in Time: Disaster Warning System Based on a Multilaterally Secure, Economic, and Mobile Infrastructure, Proceedings of the Eleventh Americas Conference on Information Systems (AMCIS 05), AIS, Omaha, Nebraska, 2005, 1610-1619.
[25] A.L. Stinchcombe, Constructing Social Theories, Univ. of Chicago Press, Chicago, IL, USA, 1987.
[26] M. Turoff, M. Chumer, R. Hiltz, R. Klashner, M. Alles, M. Vasarhelyi and A. Kogan, Assuring Homeland Security: Continuous Monitoring, Control & Assurance of Emergency Preparedness, Journal of Information Technology Theory and Application (JITTA), 6, (3) (2004) 1-24.
[27] R. Van der Togt, E. Beinat, S. Zlatanova and H.J. Scholten, Location Interoperability Services for Medical Emergency Operations during Disasters, in: P. van Oosterom, S. Zlatanova and E. M. Fendel, Ed., Geo-information for Disaster Management, Springer Verlag, Heidelberg, 2005, pp. 1127-1141.
[28] Y. Yuan and B. Detlor, Intelligent Mobile Crisis Response Systems: Systems to help coordinate responder communication and response efforts in order to minimize the threat to human life and damage to property., Communications of the ACM, 48, (2) (2005) 95-98.

Extracting line string features from GPS logs

Jörg Roth

Georg-Simon-Ohm-Hochschule Nürnberg
90489 Nürnberg
Germany
Joerg.Roth@Ohm-hochschule.de

Abstract

Geo data is an important foundation for any type of location-based service, but geo data often is expensive or distribution is limited by certain license restrictions. As a solution, open projects collect geo data from individuals and publish these data under a public licence. As a major drawback, the correction and integration of collected data is difficult and cost-intensive.
This paper describes an approach to automatically derive linear map data (roads, paths, highways etc.) from GPS logs. People just have to passively carry inexpensive GPS loggers whenever they drive or walk outdoors. The raw GPS logs then are automatically merged to a map. The approach contains error correction and sensor data fusion mechanisms. Our algorithm takes into account the specific measurement characteristic of GPS and is based on a probabilistic model. It performs two steps: first, track parts that represent the same paths are identified. Second, identifiable parts are fused to a single path, considering the Gaussian distribution of the input measurements.
We verified the approach with approx. 200 000 measurements in the area of Nuremberg that represents approx. 6 000 km driving distance. We show that an automatic processing of GPS logs to produce a road map is effectively possible.

1. Introduction

Geo data form the natural resource for location-based services. Geo objects describe the world in terms of natural, artificial and virtual entities that cover the Earth's surface. Important functions such as displaying maps, maintaining points of interests (POIs) and navigation strongly rely on the quality of geo data.
Official geo data often is expensive or protected by certain license agreements. As a solution, open projects collect geo data from individuals and publish these data under a public licence. For this, people measure path information with the satellite navigation system GPS using so-called *GPS loggers.* They periodically measure the current position (e.g. every 5s) that is stored in the persistant device memory. In projects such as *Open Street Maps* [OSM08], people can derive road information from GPS loggers that is integrated into a large geo database. As a major drawback, the integration of log data to maps is a manual, time-consuming task.
This paper describes an approach to *automatically* derive linear map data (roads, paths, highways etc.) from GPS logs. People just have to passively carry GPS loggers whenever they drive or walk outdoors. The raw GPS logs then are merged to a map that includes all collected data. The benefits of such an approach are:

- The collection of map data is simple and inexpensive. People just have to passively carry the GPS logger devices; further manual processing is not required. Thus, it is easier to get a higher number of contributors and even unpopular paths such as rarely used trails through forests can be detected.
- If people move on the same path multiple times, precision is improved by many measurements.
- Whereas existing geo data sources often represent a path as a single, bidirectional line, we can identify multiple lanes per road, at least different lanes for the two driving or walking directions.
- Additional statistics can be derived, e.g., the average driving speed at a certain position or statistics about driving directions. These statistics can be used for further evaluations e.g. to estimate positions of traffic lights or street sections with frequent traffic jams.

The approach does not detect the names of geo objects. Only line strings are extracted from the log data, not point-like or area-like geo objects.

2. Related Work

Related work that processes GPS logs tries to achieve three goals. First, logs can be used to detect special places that are visited more often. Second, logs can be used to get additional information about paths inside an existing map and third, logs can be used to derive line string features and build a new map from scratch. Even though, this paper presents an approach of the third kind, we briefly discuss related work of the first two types.
The problem to find special places (called *meaningful* places or *points of interest*) can be considered as a lower-dimensional variation of the line string feature extraction. Meaningful places are usually not modelled with their polygonal border but as centre points with a certain circular extension. Such points can easily be derived using clustering algorithms as proposed in [NK06] or Bayesian networks as presented in [YS07]. [JLO07] extends the clustering idea to identify moving objects.
Further approaches extend existing maps with the help of GPS logs. [JPR04] present a general framework to collect and store GPS log data to enrich existing maps, e.g. to store statistic about a car or the average speed, but no new roads are derived from the logs. [CJP05] proposed an approach to efficiently track moving objects. They take a predefined map and store the position history of an object into the given map. As a side effect, the road geometries of original maps are corrected according to the position log.
[RLW99] is one of the first approaches to derive line string features from GPS logs, but it requires an existing map that is extended. Note that the problem of modifying an existing map is much simpler because a general network of roads already exists. In [SWR+04] tracks are segmented with the help of points that are either derived from an existing map or computed using a clustering algorithm. This means, the problem to find line strings is reduced to the point clustering problem, but the line strings between the segment points are not adequately modelled. [MMB04] uses thresholds to identify corresponding linear path segments, but does not take into account the probabilistic distribution of measurements.

Discussion

Approaches that create line string maps have to face three problems as illustrated in fig. 1.

- *Line Clustering Problem*: Even though measurements may describe the same path, individual positions are not necessarily located close together. This is because the measurements are triggered using a fix clock (e.g. every 5s) or by movement (e.g. every 10m). Thus, a specific measurement randomly resides on the path line. A point clustering algorithm thus cannot be applied.
- *Segment Problem*: Users may drive on same roads, but whole tracks usually are not identical. The problem is to find out, which measurements of two tracks represent same paths and which not.

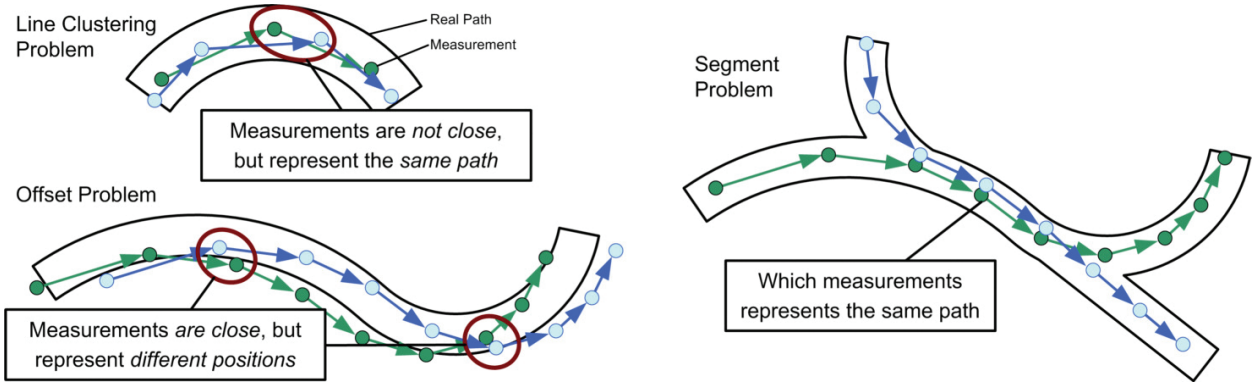


Fig. 1: Problems to fuse tracks

- *Offset Problem*: GPS logs have a high precision (i.e. subsequent measurements of the same position are close), but often have low accuracy (the difference of measured and real position often is high). If a GPS receiver receives a certain set of satellites, all measurements nearly have the same constant offset error and a track shape follows the actual road shape. Two corresponding tracks may contain similar positions that represent different path points. An approach thus has to explicitly assume a constant offset between multiple tracks and should not simply identify nearby measurements.

Many approaches consider the output of a GPS measurement as a true position without representing the error characteristic of a typical physical measurement. We strongly believe that a suitable approach has to model measurements as a probability distribution rather than a certain point in space.

3. Probabilistic Track Fusion (PTF)

We assume that GPS logs contain sequences of 2D positions and the respective time stamps. In addition, our approach has to know the variances in two dimensions for each measurement. They can be derived from information about the GPS measurement: the number of received satellites, whether *Differential GPS* was available and the so-called *Horizontal Dilution of Precision* that describes errors based the satellite constellation. If such values are not available, the standard variance of GPS is used, i.e., 78m^2 for 2dRMS(95%)=25m.
In a first step, the sequences of measurements are converted into a so-called *track model* that contains *coherent* sets of measurements:

- First, breaks are identified. Breaks are phases with zero speed for a certain time. The required time to define a break depends on the average speed nearby a break.
- Leading and trailing ends of a track (only some seconds) are removed as they often contain high measurement errors. This is because such measurement are often indoors and GPS has low accuracy before it completely fails or when it acquires a new set of satellites.
- The tracks are automatically classified according to their speed profile into classes such as pedestrian, bicycle, or car.

The *Probabilistic Track Fusion* (*PTF*) converts the track model into a *path model*. In contrast to the track model, the path model represents each real path only once. The algorithm runs in two steps:
Step 1: Identify those parts of different tracks that represent the same path.
Step 2: Fuse corresponding track parts to a single path using a probabilistic model.

The algorithm solves the problems described in fig. 1, especially the offset problem. It is an *online algorithm*, i.e. tracks are consecutively integrated into a final path model. The track ordering has no influence on the final result. Note that step 1 is based on a heuristic, whereas step 2 uses closed mathematically founded formulas.

Step 1: Identification of corresponding track parts
This step is based on histograms. Consider two tracks T_1, T_2 with measured positions P_{1i}, P_{2k}. Let $L_i(\alpha)$ denote the straight line through P_{1i} with angle α and $S_i(\alpha)$ the set of intersections of $L_i(\alpha)$ with the segments (P_{2k}, P_{2k+1}). Let further denote

$$d_i(\alpha) = \begin{cases} \min\left\{ \left| (P_{1i}, s) \right| \;\middle|\; s \in S_i(\alpha) \right\} & \text{if } S_i(\alpha) \neq \varnothing \\ \text{undef} \;\; \text{if } S_i(\alpha) = \varnothing \end{cases} \quad (1)$$

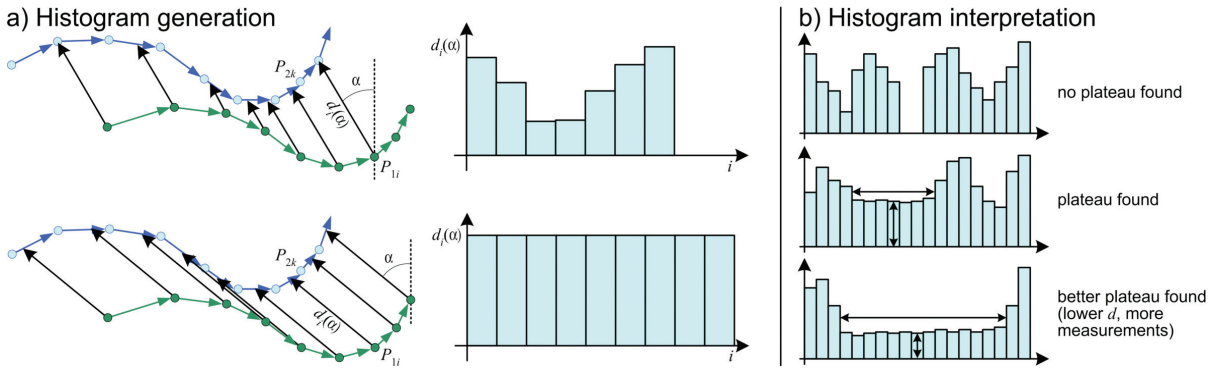


Fig. 2: Histogram generation and interpretation

A histogram for a certain angle α represents values of $d_i(\alpha)$ for all P_{1i} of T_1. Fig. 2a illustrates the histogram generation. Note that intersections in $S_i(\alpha)$ usually are not measured points P_{2k} of T_2 but are usually inside the segments (P_{2k}, P_{2k+1}). We thus build a *bidirectional* histogram that also measures the offset between points P_{2k} to segments (P_{1i}, P_{1i+1}) of T_1. For two tracks and for a set of values for α (e.g. every 10 degrees) the corresponding histogram undergoes an interpretation (fig. 2b). The goal is to find plateaus in the histogram, as they indicate a nearly constant distance between the two tracks under a certain angle α. This is typical for an offset error produced by GPS. Such plateaus can easily be detected – corresponding measurements both have a low average value d_{avg} and a low standard deviation d_{dev}. Note that according to the segment problem (fig. 1), a plateau only covers a subset $\{i_0, \ldots i_1\}$ of all histogram entries. To identify the "best" plateau, we use the formula

$$v = \frac{i_1 - i_0}{d_{avg} + d_{dev}} \tag{2}$$

where higher values represent better plateaus. After the plateau with the highest v is found, the remaining track parts again undergo a plateau search, until no further plateaus are detected. After this we know for each pair of tracks:

- whether parts of the tracks represent the same path,
- the measurements of each track that can be identified,
- the offset vector between the corresponding parts of the tracks.

Once we know the corresponding parts, we now have to fuse these parts to a single path representation.

Step 2: Fusion of corresponding track parts

Given two track segments that represent the same path, what is a *single* sequence of positions that represents both track inputs? We model measurements as a triple (x, y, var), where x, y is the measured position and var the variance of the corresponding Gaussian distribution related to the measurement. Note, even though we actually deal with two-dimensional variances, the following formulas only contain one variance value for both axes for better readability.

Our approach converts two sequences of (x, y, var) from the two input tracks again to a sequence of (x, y, var) – the output path. This is a great benefit, as the result again can be used as an input track for further fusions without to switch to another representation. From the histogram approach above we combine corresponding parts of two tracks T_1, T_2 with measured positions P_{1i}, P_{2k} as illustrated in fig. 3.

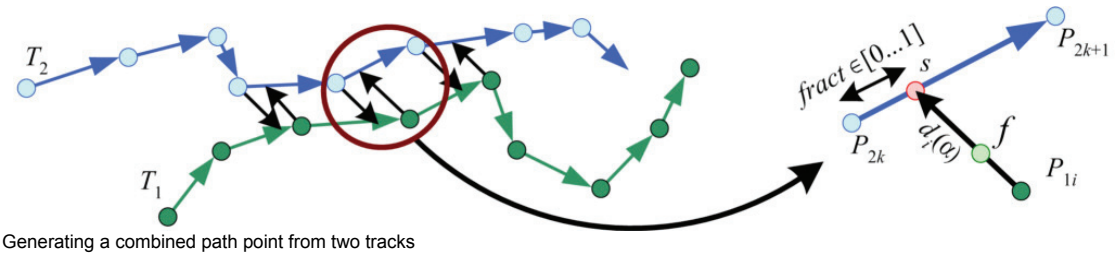


Fig. 3: Generating a combined path point from two tracks

Our model contains

- the intersection $s = (x_s, y_s, \mathrm{var}_s)$ of (P_{2k}, P_{2k+1}) and $L_i(\alpha)$ with distance $d_i(\alpha)$ to P_{1i}
- the fraction $fract = |(P_{2k}, s)| \,/\, |(P_{2k}, P_{2k+1})|$
- the position $f = (x_f, y_f, \mathrm{var}_f)$ that represents the combined path point after fusing the tracks T_1, T_2.

As $L_i(\alpha)$ is given by the histogram, the position (x_s, y_s) of s is known, but we still have to compute var_s. We use the formula

$$\mathrm{var}_S = \mathrm{var}_{P_{2k}} \cdot (1 - fract) + \mathrm{var}_{P_{2k+1}} \cdot fract \tag{3}$$

that represents the linear interpolation between the two measurement variances of P_{2k}, P_{2k+1}. We also could take into account that between two measurements we do not have definite knowledge about the positions, especially near the centre. We could reflect the missing knowledge by a higher variance, with e.g., the formula

$$\mathrm{var}_S = \left(\mathrm{var}_{P_{2k}} \cdot (1 - fract) + \mathrm{var}_{P_{2k+1}} \cdot fract\right) \cdot (2 - (2\,fract - 1)^2) \tag{4}$$

that adds the variances of P_{2k} and P_{2k+1} for *fract* = 0.5. Note that other estimations with similar characteristics are conceivable. Experiments show that formula (3) produces reasonable result, but an optimal formula is based on a motion model (see future work).
From (x_s, y_s, var_s) and (x_{P1i}, y_{P1i}, var_{P1i}) we now can compute the combined position f using Bayes conditional probabilities applied to Gaussian distributions of GPS measurements. We get

$$f = \begin{pmatrix} x_f \\ y_f \\ \mathrm{var}_f \end{pmatrix} = \begin{pmatrix} \left(x_S\,\mathrm{var}_{P_{1i}} + x_{P_{1i}}\,\mathrm{var}_S\right) / \left(\mathrm{var}_{P_{1i}} + \mathrm{var}_S\right) \\ \left(y_S\,\mathrm{var}_{P_{1i}} + y_{P_{1i}}\,\mathrm{var}_S\right) / \left(\mathrm{var}_{P_{1i}} + \mathrm{var}_S\right) \\ \left(\mathrm{var}_{P_{1i}} \cdot \mathrm{var}_S\right) / \left(\mathrm{var}_{P_{1i}} + \mathrm{var}_S\right) \end{pmatrix} \tag{5}$$

Note that the variance of f is always lower than the lowest variance of the two input positions (regardless of their distance), thus the precision always increases for each step.
Until now we only fuse two *different* tracks, but also a *single* track may represent same paths multiple times. If, e.g. a driver lost the way, he may drive on the same road twice. Such a track has to be cut into two independent parts in order to use to approach above. For this, prior to checking two tracks, each track undergoes a self-overlap test, which also uses the histogram approach above: a histogram is generated that uses the same track for both inputs. To avoid the meaningless zero-plateau with all $d_i(\alpha)$=0, the two values of i_0 have to keep a certain distance, thus only real self-overlaps are detected.
Whenever two track segments are fused, the number of resulting positions is approximately the sum of measurements of both input segments. The number permanently would increase, if no additional mechanism was applied. Thus, after a fusion step, the result path undergoes a resampling step. It removes the second position of three consecutive positions that nearly reside on a straight line. As a result, the number of positions remains nearly constant.

4. Sample Execution and Performance Discussion

We verified our approach with the help of two sample data sets mainly collected in the area on Nuremberg. Table 1 presents the statistics.

Table 1: Sample Execution Statistics

	Urban Example		**Regional Example**	
	Input	**PTF Output**	**Input**	**PTF Output**
Number of Tracks	518	649	1 252	1 378
Number of Measurements	120 501	48 747	221 700	99 302
Measured Distance	2 146 km	251 km	5 899 km	1 435 km
Processing Time (@3.4 GHz)	9.5 hours		11.0 hours	

Fig. 4 illustrates the results of the urban example. Note that even though the number of measurements (and thus the measured distance) dramatically is reduced by the PTF fusion process, the number of tracks slightly increased. This is because long tracks that cover multiple roads are split up into short paths.
This sample execution shows that the PTF approach is able to automatically produce the required output. The path model now can easily be used for further processing, e.g. to create a topology map or to generate statistics about the respective roads.
Some remarks about the complexity: Processing time mainly depends on the number of tracks t and the average number of measurements per track m. If we consider the number of checked angles (and thus the number of histograms) as constant, we can generate histograms for pairs of tracks in $O(m^2)$ steps. This means $O(tm^2)$ steps are required for the self-overlap test and $O(t^2m^2)$ to fuse all tracks. If $n = tm$ denotes the total number of measurements, PTF requires $O(n^2)$ steps.

Fig. 4: Track Input (left) and generated Path Model (right)

As PTF does have to fulfil real time constraints, this complexity is acceptable. In addition, our implementation uses a spatial index to identify pairs of distant paths where the histogram generation is futile. This is why the regional example with widely spread paths does not require significant more time than the urban example.

5. Conclusion and Future Work

Our approach significantly simplifies the production of maps from GPS logs. It includes two steps: first, it creates and interprets histograms to identify of corresponding track parts and second, it combines paths with a probabilistic model that considers the error distribution of the GPS measurements. The approach is verified with the help of large sets of test data.

In the future we want to improve some parts of the approach. Especially the position and variance of the intersection point s is worth to be discussed. In reality, the trajectory between two measurements is not a straight line, but depends on velocity and acceleration of the moving object. We thus could model s much better and get better result for f, if we knew certain motion parameters such as the maximum acceleration. Thus, the next step is to derive a motion model from the given tracks that is integrated into PTF.

6. References

[CJP05] Civilis, A.; Jensen, C.; Pakalnis, S., Techniques for Efficient Road-Network-Based Tracking of Moving Objects, IEEE Transaction on Knowledge and Data Engineering, Vol. 17, No. 5, May 2005, 698-712

[JLO07] Jensen, C.; Lin, D.; Ooi, B., Continuous Clustering of Moving Objects, IEEE Transaction on Knowledge and Data Engineering, Vol. 19, No. 9, Sept 2007, 1161-1174

[JPR04] Jensen, C.; Lahrmann, H.; Pakalnis, S.; Runge, The infati data, J., TR-79, July 28, 2004, TIMECENTER Technical Report

[MMB04] Morris, S.; Morris; A.; Barnard, K.: Digital Trail Libraries, Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries, 2004, 63-71

[NK06] Nurmi, P.; Koolwaaij, J., Identifying meaningful locations, 3rd Annual International Conference on Mobile and Ubiquitous Systems - Workshops, 2006, 1-8

[OSM08] OpenStreetMap, http://www.openstreetmap.de/

[RLW99] Rogers, S.; Langley, P.; Wilson, C., Mining GPS Data to Augment Road Models, Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 1999, 104-113

[SWR+04] Schroedl, S.; Wagstaff, K.; Rogers, S.; Langley, P.; Wilson, C., Mining GPS Traces for Map Refinement, Data Mining and Knowledge Discovery, Vol. 9, No. 1, July 2004, Springer-Verlag, 59-87

[YS07] Extracting Meaningful Contexts from Mobile Life Log, Youngseal, L., Sung-Bae C., Proc. of Intelligent Data Engineering and Automated Learning - IDEAL 2007, Springer LNCS 4881, 2007, 750-759

Sichere Ortungsverfahren

Michael Decker

Institut AIFB,
Universität Karlsruhe (TH), 76 128 Karlsruhe

Abstract

Für die zur Realisierung ortsabhängiger und mobiler Anwendungen (Location-based Services) notwendige Ortung mobiler Endgeräte gibt es zahlreiche technische Verfahren, etwa satellitengestützte Systeme wie GPS oder Zellortung in Mobilfunknetzen. Die meisten in der Literatur beschriebenen Verfahren verfügen aber über keine besonderen Mechanismen, um bewusste Manipulationsversuche durch Dritte oder den mobilen Nutzer selbst (sog. „Location-Spoofing“) abzuwehren, obwohl dies für einige Anwendungsszenarien erforderlich ist. Im vorliegenden Beitrag wird deshalb ein systematischer Überblick über Verfahren zur Verhinderung solcher Manipulationsversuche gegeben.

1. Einleitung und Motivation

Location-based Services (LBS) sind mobile Dienste, die den Aufenthaltsort zumindest eines mobilen Endgerätes auswerten, wobei es sich hierbei nicht notwendigerweise um das Endgerät des Dienstaufrufers handeln muss (etwa bei Tracking-Diensten wie „Friend-Finder“). Solche Dienste werden meist anhand von Consumer-Szenarien (z.B. POI-Finder oder Navigationsdienste) motiviert, bei denen es für keinen Akteur einen Anreiz zur Manipulation des verwendeten Ortungsverfahrens gibt. Es sind aber durchaus Szenarien denkbar, in denen ein Angreifer ein Ortungsverfahren beeinflusst, etwa um Fahrzeuge in einen Hinterhalt zu locken (z.B. Werttransporte, militärische Fahrzeuge). Darüber hinaus sind noch andere Anwendungen von LBS denkbar, bei denen die Ortung gegen bestimmte Manipulationsversuche abgesichert sein sollte:

- Bei ortsabhängiger Zugriffskontrolle wird der mobile Zugriff auf entfernte Ressourcen wie vertrauliche Kundendaten über einen stationären Server nur dann gewährt, wenn sich der Nutzer an einem bestimmten Ort aufhält. So sollen die Folgen des Verlusts von mobilen Endgeräten beschränkt oder ein Ausspähen von Daten durch Unbefugte verhindert werden.
- Eine ortsabhängige Zugriffskontrolle kann unter bestimmten Voraussetzungen auch eine herkömmliche identitätsbasierte Zugriffskontrolle (z.B. über Nutzername/Passwort) ersetzen, wenn etwa angenommen werden kann, dass nur Betriebsangehörige Zutritt zum Firmengelände haben, weil an der Pforte eine Eingangskontrolle stattfindet. Somit kann Privacy-Problemen entgegengewirkt werden, die sich aus der Zuordnung von Dienstsitzungen zu Nutzern ergeben.
- Bestimmte mobile Dienste (z.B. Internetzugang, Informations- oder Unterhaltungsangebote) sollen nur für Kunden verfügbar sein, die sich gerade in einem bestimmten Ladenlokal (z.B. Restaurant, Tankstelle) oder auf dem Firmengelände (z.B. Freizeitpark, Hochschulcampus) befinden.
- Für Digital Rights Management (DRM) sind Szenarien denkbar, in denen ein Zugriff auf lokale Ressourcen wie Multimedia-Dateien nur in den Ländern möglich sein soll, für die die jeweiligen Inhalte lizenziert sind [Mund05]. Eine einfache Form von ortsabhängigem DRM in der Praxis stellen die Region-Codes von DVD-Filmen dar. Es gibt auch Webangebote, die nur für Nutzer aus bestimmten Ländern verfügbar sind (z.B. das Filmportal Hulu.com), wobei das Ursprungsland eines Requests anhand der IP-Adresse ermittelt wird.
- LBS können auch zur Überwachung des Aufenthaltsortes von Gütertransporten, Fahrzeugen oder Personen unter Bewährungsauflagen eingesetzt werden.

Es wurden deshalb spezielle Ortungssysteme oder Erweiterungen für konventionelle Ortungssysteme entwickelt, die gegen bestimmte Manipulationsversuche (Angriffe) resistent sein sollen. Unter Manipulationen wird hierbei verstanden, dass ein anderer als der tatsächliche Aufenthaltsort als Ergebnis der Ortung ermittelt wird. Im vorliegenden Beitrag soll ein Überblick über solche Ortungsverfahren gegeben werden. Eine bloße Verhinderung der Ortung (Denial of Service etwa durch Störsender, „Jamming“) wird nicht betrachtet, da dies von der getäuschten Partei bemerkt wird und höchstens durch funktechnische Maßnahmen erschwert werden kann, z.B. Manchester-Coding, Aufspreizen des Signals über ein möglichst breites Spektrum oder Verwendung möglichst vieler verschiedener Frequenzen.

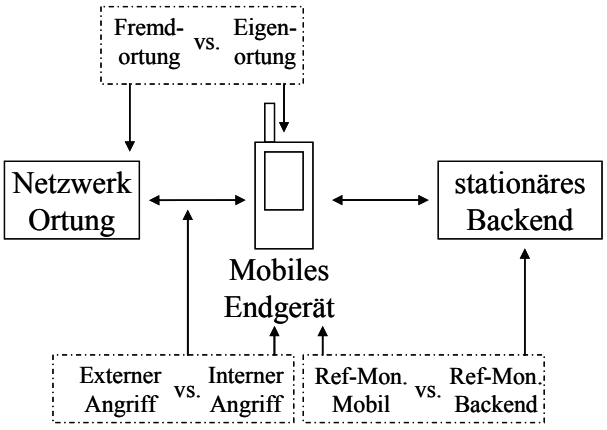


Abbildung 1: Grundarchitektur und Kriterien für sichere Ortungsverfahren

Für die Betrachtung sicherer Ortungsverfahren sind die folgenden Kriterien zur Beschreibung von Szenarien von besonderer Bedeutung (siehe auch Abbildung 1):

- Eigen-/Fremdortung: kann das mobile Endgerät selbst die Ortung errechnen oder wird die Ortung von der stationären Infrastruktur (Ortungsnetzwerk) errechnet?
- Mobiler/stationärer Referenz-Monitor: wird die die Ortung auf dem mobilem Endgerät selbst ausgewertet (z.B. GPS) oder auf einem stationären Backend-System (z.B. Zellortung)?
- Interner/externer Angreifer: wird der Manipulationsversuch durch einen Angreifer durchgeführt, der im legitimen oder illegitimen Besitz (z.B. Dieb) des Endgerätes ist oder handelt es sich um einen externen Dritten? In der Literatur wird der Begriff "Location-Spoofing" (in Anlehnung an Adressmanipulationen wie ARP-, IP-, DNS- oder MAC-Spoofing in drahtgebundenen Netzwerken) für Manipulationsversuche für Ortungssysteme durch interne (z.B. [Mund05]) als auch durch externe Angreifer (z.B. [WaJo03]) verwendet.

Der Fall eines externen Angreifers bei Eigenortung mit einem Referenzmonitor auf dem Backend ist kein ortungsspezifisches Problem, da die Manipulation dieser Kommunikation durch Verschlüsselungsverfahren (evtl. mit Zeitstempeln) verhindert werden kann.

Bevor die einzelnen Ansätze zur Verhinderung von Location-Spoofing vorgestellt werden, wird eine grobe Klassifikation solcher Ansätze eingeführt. Es werden auch Anforderungskriterien an die Fähigkeiten der mobilen Endgeräte bzw. zur Verfügung stehende Kommunikationsinfrastruktur genannt, durch die sich die einzelnen Ansätze unterscheiden.

2. Ansätze für Sicherung von Ortungsverfahren gegen Manipulationen.

1.1. Übersicht

Die in der Literatur beschriebenen Verfahren zur Absicherung von Ortungsverfahren gegen externe und interne Manipulationen lassen sich in folgende Gruppen einteilen: Plausibilitätskontrollen, Verwendung manipulationssicherer Hardware, Location-Keys, Request-Response-Verfahren und Digitale Signaturen.
Die einzelnen Verfahren unterscheiden sich auch erheblich in den Anforderungen, die sie an die Endgeräte stellen:

- Ist eine hochpräzise Zeitquelle auf dem Endgerät notwendig?
- Muss das Endgerät rechenaufwändige symmetrische oder gar asymmetrische Kryptographiealgorithmen durchführen können?
- Muss zwischen Endgerät und Backend ein Kommunikationskanal zur Verfügung stehen, der evtl. auch bestimmte Eigenschaften (max. Latenzzeit, bestimmtes Kommunikationsmedium wie Ultraschall) haben muss?

Als grundsätzliche Angriffsvariante für Ortungsverfahren ist „Rerouting“ (auch Wormhole-Angriff genannt) zu nennen: hierbei wird ein Signal an einem bestimmten Ort empfangen und an einen anderen Ort über einen vom Ortungssystem versteckten Kommunikationskanal („out-of-band“) weitergeleitet. Für GPS werden etwa von Herstellern wie Navilock oder Chronos sog. „Reradiatoren“ oder „GPS-Repeater“ angeboten, mit denen GPS-Signale an einen anderen Ort weitergeleitet werden, um auch innerhalb von Gebäuden (z.B. Entwicklungslabore, Verkaufsräume) mit GPS arbeiten zu können. Rerouting ist prinzipiell nur durch die damit einhergehende Zeitverzögerung (Latenz) erkennbar. Eine präventive Maßnahme ist die Verwendung eines möglichst breit gespreizten Signals, so dass die Weiterleitung erschwert wird. Bei Replay-Angriffen wird das Signal aufgezeichnet und später wiedergegeben.

1.2. Plausibilitätskontrollen

Einfache Plausibilitätskontrollen können — wenn nicht nur eine einzelne punktuelle Ortung vorgenommen wird — überprüfen, ob eine Folge von Orten auch plausibel ist, also insbesondere ob unrealistisch große Ortssprünge oder Geschwindigkeiten auftauchen und ob Gebiete „durchfahren“ werden, die eigentlich bebaut sind. Soll dies für eine Eigenortung durchgeführt werden kann besondere Hardware notwendig sein, z.B. Beschleunigungs- und Entfernungsmesser (Accelero- bzw. Odometer) oder ein (Kreisel-)Kompass.

Weitere Plausibilitätskontrollen sind auf der Ebene der vom Ortungssystem ausgestrahlten Funksignale möglich. In [WaJo03] werden mehrere Ansätze vorgestellt, wie bei der Verwendung von GPS der mobile Nutzer externes Spoofing durch falsche GPS-Sender (sog. Pseudoliten, also terrestrische Sender, die eigentlich von Satelliten ausgestrahlte Signale aussenden) erkennen kann: liegt die gemessene absolute Signalstärke eines Satellits deutlich über dem erwarteten Wert von ca. -160 dBw deutet dies auf einen Angriff hin. Ebenso können starke Pegelschwankungen oder gleiche Pegelstärke der Signale von verschiedenen Satelliten zur Erkennung von Angriffen genutzt werden. Auch verdächtig ist es, wenn die Signale von verschiedenen Satelliten zum glei-

chen Zeitpunkt beim Empfänger eintreffen oder die IDs der Empfangenen Satelliten nicht denen entsprechen, die laut Almanach (langfristige Satellitenbahndaten, gültig für drei Monate) in der aktuellen Region sichtbar sein sollten. Verfügt das mobile Endgerät über einen Kompass und/oder Beschleunigungsmesser kann es überprüfen, ob die ermittelte Ortung zu den zurückgelegten Strecken oder Richtungswechseln passen.

1.3. Manipulationssichere Hardware

Befindet sich der Referenzmonitor auf dem mobilem Endgerät, so ist ein manipulationssicheres Hardware-Modul (MHM) notwendig, damit der Besitzer des Endgeräts die Zugriffsentscheidung durch Manipulationen der Hardware umgehen kann (etwa direktes Auslesen gespeicherter Inhalte). Dieser Ansatz wird in [Mund06] zur Realisierung von ortsabhängigem Digital Rights Management eingesetzt. Neben dem Referenzmonitor befindet sich bei diesem Ansatz im MHM auch die Funktionalität zur Eigenortung mittels Satelliten-Signalen. Damit Rerouting- und Replay-Angriffe erkannt werden können, beinhalten diese Signale einen Zeitstempel und sind durch eine digitale Signatur vor Veränderungen geschützt. Um mögliche Zeitabweichungen aber tatsächlich zu erkennen, wird auch eine manipulationssichere Uhr benötigt. Bei Verwendung eines für Consumer-Anwendungen erschwinglichen Uhrenmoduls ist es aber notwendig, im Abstand von mehreren Stunden pro Tag eine Synchronisation durchzuführen, für die dann vorübergehend eine schnelle Rückleitung zur Verfügung stehen muss.

Manipulationssichere Hardware kann aber auch sinnvoll sein, wenn bei Eigenortung der Referenzmonitor sich auf dem stationären Backend befindet. Das MHM berechnet hierbei wiederum die Ortung, signiert sie aber und leitet sie an das Backend über drahtlose Datenkommunikation weiter. Da der zur Signatur benötigte private Schlüssel vom MHM gekapselt wird, ist es einem Angreifer nicht möglich, diesen in Erfahrung zu bringen, um selbst erstellte Ortungsinformationen zu signieren.

1.4. Locations-Keys

Location-Keys sollen internes Spoofing verhindern: der mobile Nutzer muss gegenüber dem stationären Backend seinen Aufenthalt an einem Ort dadurch beweisen, dass er eine „Location Key“ genannte Information übermittelt, die nur an diesem Ort empfangen werden kann. Es kann zwischen natürlichen und künstlichen Location-Keys unterschieden werden: Künstliche Location-Keys werden speziell für den Zweck der Verhinderung von Location-Spoofing ausgestrahlt, während natürliche Location-Keys nicht extra zur Absicherung von Ortungsverfahren erzeugt werden.

Ein Verfahren mit natürlichen Location-Keys ist der Cyber-Locator [DeMa96]. Es handelt sich hierbei um eine Erweiterung von GPS, also ein Eigenortungsverfahren. Der Referenzmonitor befindet sich auf einem stationären Backend-Server und möchte sich davon überzeugen, dass der mobile Nutzer sich auch tatsächlich an einem bestimmten Ort befindet, z.B. um davon abhängig über Zugriffsanfragen auf sensitive Daten zu entscheiden. Es ist deshalb nicht das bei GPS übliche Vorgehen ausreichend, bei dem das mobile Endgerät den anhand der GPS-Signale errechneten Ort an den Referenzmonitor weiterleitet, da hierbei beliebige Koordinaten „erfunden“ worden sein könnten. Beim Cyber-Locator übermittelt das mobile Endgeräte als Location-Key („Location Signature“) das von den GPS-Signalen empfangene Rohsignal. Dieses unterliegt zufälligen Einflüssen (z.B. atmosphärische Störungen, Abweichungen der Satelliten von Laufbahnen) und kann deshalb nicht vorhergesagt werden. Der Referenzmonitor vergleicht dieses Rohsignal mit dem von einer vertrauenswürdigen Referenzstation empfangenen, die sich in der Nähe des vermeintlichen Aufenthaltsortes befindet. Bei einer Übereinstimmung wird davon ausgegangen, dass der mobile Nutzer sich tatsächlich an dem angegebenen Ort befindet. Um hierbei Rerouting-Angriffe auszuschließen muss das Rohsignal innerhalb einer bestimmten Zeitspanne vom mobilen Endgerät an den Referenzmonitor weitergeleitet werden; die Autoren nennen hierbei einen Wert von 5 ms für diese Zeitspanne, was mit herkömmlichen ISDN- und DSL-Verbindungen nicht erreicht werden kann. Die Referenzstation sollte maximal 2.000 bis 3.000 km von mobilen Endgerät entfernt sein.

Ein Beispiel für ein Ortungssystem mit künstlichen Location-Keys ist „LAAC“ (Location-Aware Access Control) von Cho et al. [CBGo06]. Es wird für WLAN-Kommunikation beschrieben, das Prinzip ist aber unabhängig von dieser Technik anwendbar. Die einzelnen WLAN-Access-Points (APs) erzeugen hierbei unabhängig voneinander zufällige Location-Keys („Nonces“), die periodisch gewechselt werden und zusammen mit anderen Informationen ausgestrahlt werden. Wenn die APs über Sektorenantennen verfügen, kann durch Einsatz mehrerer zu einer Gruppe zusammengefassten APs ein rechteckiger Bereich als Zugriffsgebiet definiert werden; ein AP kann hierbei Mitglied in mehreren Gruppen sein. Das mobile Gerät weist sein Aufenthalt innerhalb eines

Zugriffsgebietes dadurch nach, dass es alle empfangenen Keys der APs einer bestimmten Gruppe mit der XOR-Funktion verknüpft und einen Hash-Wert vom Ergebnis berechnet. Der Hash-Wert dient gegenüber dem stationären Backend als Nachweis des Aufenthalts an einem bestimmten Ort; damit das Backend diesen Key verifizieren kann, muss es die aktuellen Location-Keys aller APs kennen. Einen Rerouting-Angriff erachten die Autoren für zu aufwändig für die Nutzer als dass dafür Gegenmaßnahmen vorgesehen werden müssten. Dies ist vielleicht auch vor dem Hintergrund zu sehen, dass das das beschriebene System hauptsächlich dazu konzipiert ist, Internetzugang zu gewähren (z.B. für die Kunden einer Tankstelle). Verfügt ein Angreifer also bereits über eine schnelle Internet-Anbindung, so ist es nicht sinnvoll für ihn, dieses System zu überlisten.

Location-Key-Verfahren haben den Vorteil, dass weder hochpräzise Uhren, aufwändige Berechnungen auf dem Endgerät oder manipulationssichere Hardware notwendig sind. Es wird aber ein unidirektionaler Kommunikationskanal zum Backend-System mit geringer Latenzzeit benötigt.

1.5. Request-Response-Verfahren

Das Grundprinzip der Request-Response-Verfahren ist es, das zwischen dem mobilen Endgerät und einem stationären Sender Signale ausgetauscht werden und aus der gemessenen Laufzeit eine Entfernung abgeleitet wird: auf die Anfrage (erste Nachricht) muss hierbei unmittelbar vom Empfänger (sog. Prover) eine Antwort geschickt werden, wobei die Anfrage eine Zufallsinformation (etwa Bitstring) hinreichender Größe enthält, die in die Antwort einfließen muss; so wird gewährleistet, dass eine Antwort nur dann gesendet werden kann, wenn auch die entsprechende Anfrage empfangen wurde. Die Laufzeitmessung wird immer von der anfragenden Partei (sog. Verifier) vorgenommen; der Prover weißt gegenüber dem Verifier also eine Höchstentfernung nach. Für den Fall der Eigenortung würde das mobile Endgerät (Verifier) die Anfrage an eine Bake (Prover) schicken; in der Literatur wird dieses Verfahren aber für Fremdortung beschrieben, das Anfragesignal wird also von einer stationären Basisstation als Verifier ausgestrahlt. Je nach Anwendungsfall kann es sinnvoll sein, dass der Prover durch ein Signal den Verifier dazu auffordert, mit der Request-Response-Sequenz zu beginnen.

Ein solches Verfahren wird von Sastry et al. [SWa03] mit dem sog. Echo-Prokoll beschrieben: der Prover (mobiles Endgerät) schickt hierbei die Koordinaten seines angeblichen Aufenthaltsortes sowie die Rechenzeit, die zur Beantwortung des aus einer Zufallsbitfolge („Nonce“) bestehenden Requests benötigt wird, an einen Verifier. Die Besonderheit des Echo-Protokolls ist es, dass die Antwort des Provers über Ultraschall und nicht normale Funkwellen gesendet werden muss. Der Verifier berechnet, wie lange es dauern darf, bis seine Anfrage reflektiert wird, wobei hierbei die unterschiedlichen Fortbewegungsgeschwindigkeiten von Ultraschall- und Funk berücksichtigt werden. Um den Fall einer zu hoch angegebenen Berechnungszeit des Provers zu berücksichtigen, wird die Entfernung, die Ultraschallsignale in dieser Zeitspanne zurücklegen können, als Messungenauigkeit angenommen. Da die Fortbewegungsgeschwindigkeit von Ultraschallsignalen um den Faktor 10^6 langsamer als die Lichtgeschwindigkeit ist, bewegen sich die so errechneten Messungenauigkeiten in akzeptablen Größenordnungen. In [WaFe03] wird ein weiterer Ansatz für sichere Ortung auf Basis eines Request-Response-Protokolls beschrieben; hierbei kommen aber nur Funkwellen zum Einsatz.

Der Vorteil der Request-Response-Verfahren besteht darin, dass keine hochpräzisen Uhren benötigt werden, auch sind keine rechenaufwändigen Public-Key-Verfahren notwendig. Es ist muss allerdings ein bidirektionaler Kommunikationskanal verfügbar sein, der eine geringe Latenz aufweist.

1.6. Digitale Signaturen

Durch den Einsatz digitaler Signaturen (ohne manipulationssichere Hardware) lassen sich zwei Arten von externen Angriffen verhindern: zum einen können die vom Ortungsnetzwerk ausgestrahlten Signale gegen Manipulationen gesichert werden, der externe Angreifer kann die Ortungssignale auf ihrem Weg zum mobilen Endgerät nicht verändern, so dass solche Manipulationen erkannt werden. Leitet das mobile Endgerät eine Eigenortung an ein Backend-System weiter, können Manipulationen dieser Nachrichten anhand der Signatur erkannt werden.

Bei Verwendung von Signaturen ist es zur Verhinderung von Replay-Angriffen immer noch notwendig, dass die signierte Nachricht einen Zeitstempel beinhaltet. Der Empfänger des Signals kann durch Vergleich mit seiner eigenen Uhr dann erkennen, ob das Signal umgeleitet oder als Aufzeichnung wiedergegeben wurde. Dies erfordert aber eine hinreichend genaue Zeitquelle, um Laufzeitabweichungen in der Größenordnung von Millise-

kunden erkennen zu können. Da auf mobilen Endgeräten i.d.R. nur einfache Quarz- (und keine Atom-)Uhren zur Verfügung stehen, wird zumindest gelegentlich ein Rückkanal zum Zeitabgleich benötigt.

3. Absicherung für reale Systeme

Die bisher vorgestellten Ansätze sind — wenn überhaupt — nur als Forschungsprototypen realisiert. Doch auch schon in den 1970ern konzipierte „Global Positioning System“ (GPS) beinhaltet Schutzmaßnahmen gegen Spoofing: Die Satelliten strahlen auf zwei Frequenzen L1 und L2 jeweils die gleichen Navigationsnachrichten aus. Das L2-Band ist für militärische Nutzung ausgelegt, weshalb die von ihm transportierten Daten (sog. P-Code) mit einer geheimen Spreizcodesequenz aufmoduliert werden. Dieses Verfahren kann als symmetrische Verschlüsselung aufgefasst werden, ohne seine Kenntnisse ist es für einen externen Angreifer nicht möglich, selbst die entsprechenden Signale zu erzeugen. Allerdings muss der symmetrische Schlüssel auch auf allen Endgeräten gespeichert sein, so dass wenn nur eines davon kompromittiert wird der symmetrische Schlüssel auf allen Endgeräten ausgetauscht werden muss, um weiterhin eine Absicherung gegen Spoofing-Angriffe darzustellen. Die Verwendung des P-Codes bei GPS sollte aber vielmehr gewährleisten, dass eine höhere Ortungsgenauigkeit nur einem bestimmten Nutzerkreis (nämlich der US-Armee und ihren Verbündeten) zur Verfügung steht. Mit sog. Codeless-Receivern kann mittlerweile aber auch ohne Kenntnis der geheimen Spreizcode-Sequenz unter Ausnutzung des L2-Bandes eine höhere Ortungsgenauigkeit erreicht werden.

Das von der EU geplante Satelliten-Navigationssystem „Galileo“ soll für bestimmte Dienste Signale ausstrahlen, die mit einem Public/Private-Key-Verfahren signierte Navigationsnachrichten enthalten. Die einzelnen mobilen Endgeräte verfügen also nur über den öffentlichen Schlüssel, um die empfangen Nachrichten verifizieren zu können.

4. Fazit

Es wurde ein Überblick über verschiedene Konzepte gegeben, mit denen Manipulationsversuche von Ortungsverfahren verhindert werden sollen. Hierbei wurde eine Klassifikation nach bestimmten Grundprinzipien vorgenommen. Spezielle Ansätze, um Ortung in AdHoc-Netzwerken (z.B. Sensornetzwerken) zu schützen, wurden aber nicht vorgestellt.

Literatur

[CBGo06] Cho, Y.; Bao, L.; Goodrich, M.: LAAC: A Location-Aware Access Control Protocol. Third Annual International Conference on Mobile and Ubiquitous Systems: Networking & Services, 2006, pp. 1-7.

[DeMa96] Denning, D.; MacDoran, P.: Location-Based Authentication: Grounding Cyberspace for Better Security. Computer Fraud & Security, Elsevier, February 1996, pp. 12-16.

[GaWo98] Gabber, E.; Wool, A.: How to Prove Where You Are: Tracking the Location of Customer Equipment. Proceedings of the 5th ACM Conference on Computer and Communications Security, 1998, pp. 142-147.

[Mund05] Mundt, T.: Location Dependent Digital Rights Management. Proceedings of the 10th IEEE Symposium on Computers and Communications (ISCC 2005), 2005.

[SWa03] Sastry, N.; Shankar, U.; Wagner, D.: Secure Verification of Location Claims. Proceedings of the Conference on Wireless Security (WiSe ’03), 2003, pp. 1-10.

[WaFe03] Waters, B.R.; Felten, E.W.: Secure, Private Proofs of Location. Technical Report TR-665-03, Department of Computer Science, Princeton University, 2003.

[WaJo03] Warner, J.; Johnson, R.: GPS Spoofing Countermeasures. Los Alamos Research Paper LAUR-03-6163, 2003.

Combining Web 2.0 and NGN: Mobile geo-blogging as Service Enabler for Next Generation Networks

Niklas Blum

Fraunhofer FOKUS
Kaiserin-Augsta-Allee 31
10589 Berlin
niklas.blum@fokus.fraunhofer.de

Lajos Lange

lajos.lange@fokus.fraunhofer.de

Thomas Magedanz

thomas.magedanz@fokus.fraunhofer.de

Abstract

Modern telecommunication networks and classical roles of operators are subject to fundamental change. As the value of communications networks decreased rapidly, many network operators are currently seeking for new sources to generate revenue by exposing network capabilities to 3rd party service providers.
At the same time we can observe that applications on the World Wide Web (WWW) are becoming more mature in terms of the definition of APIs that are offered towards other services. The combinations of those web-based services are commonly referred to as Web 2.0 mash-ups. Web 2.0 is the economic community buzzword that encompasses several technologies supporting the creation, collaboration and cross-linking of Internet societies. Really Simple Syndication (RSS) is one of these technologies and is commonly used to publish changing content.
This report describes our approach to prototype a mobile geo-blogging services combing technology from the Web 2.0 and real-time signaling from the telecommunications domain. Telecommunications Service Delivery Platform (SDP) and IP Multimedia Subsystem (IMS) technologies as standardized session control entities are highly potent and are set to form the future infrastructure of telecommunications service providers and operators.

1. Introduction

Telecommunications is at crossroads, the convergence of fixed and mobile telecommunications, cable networks, as well as the Internet leads into a global all-IP based Next Generation Network (NGN). Through this ongoing process of the convergence of access networks and the existence of new players in the telecommunications market, traditional operators and carriers are seeking for new business models to increase their revenue. The reuse of an extensible set of existing service components to rapidly create new market driven applications is a key aspect of telecommunications platforms since many years and gains a new momentum with the definition of dedicated application enablers for NGNs. One real-life example is British Telecom's BT Web21C SDK [1] solution that defines an API to expose telecommunications specific core network functionalities to 3rd party service developers using Web Services.
In this context we have prototyped a geo-blogging enabler for telecommunications service providers that combines technologies from the Web 2.0 domain as syndication with real-time signaling from the IP-based NGN to offer 3rd party service integration towards the WWW domain.

2. Related Standards and Technology

The following subsections describe emerging standards as the IMS, IMS enablers, OSE, and related technologies to the term Web 2.0 like Ajax and the mash-up service architectures.

1.1. NGN / 3GPP IMS

The 3GPP IP Multimedia Subsystem (IMS) provides the interfaces for interaction and underlying communication control infrastructure. The IP Multimedia Subsystem [2] is defined from 3GPP Release 5 specifications on as overlay architecture on top of the 3GPP Packet Switched (PS) Core Network for the provision of real time multi-media services.
Due to the fact that the IMS overlay architecture is widely abstracted from their interfaces, the IMS can be used for any mobile access network technology as well as for fixed line access technology as currently promoted by the European Telecommunications Standards Institute's (ETSI) Telecoms & Internet converged Services & Protocols for Advanced Networks (TISPAN) [3] within the Next Generation Network reference architecture definition.
The central session control protocols are the Session Initiation Protocol (SIP) [4] and Diameter [5]. The SIP Application Server (AS) is the service relevant part in the IMS. How multimedia applications are programmed is out of scope of the standardization committees [6]. But the SIP AS needs to support well defined signalling and administration interfaces (3GPP ISC and Sh-interfaces) to connect to the standardized network architecture. The following figure 1 depicts the simplified IMS architecture.

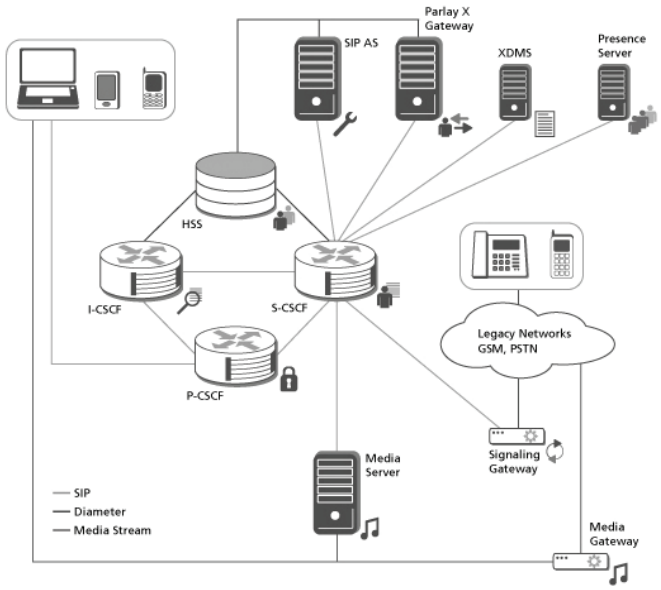


Figure 1: Basic IMS Architecture

The particular techniques and methodologies that are required to gain the advantages of these key functionalities are not completely new, but the IMS provides the first major integration and the interaction of all key functionalities.

1.2. Telecommunications Service Enabler

Similar to Service-Independent Building Blocks (SIBs) which form part of the conceptual model for Intelligent Networks, the Open Mobile Alliance (OMA) defined during the last years service enablers for the IP Multimedia Subsystem. The ideas was initially born during the specification of a Push-to-Talk over Cellular (PoC) [7] service, a walkie-talkie like communication service between several mobile peers based on the Internet Protocol (IP) using the SIP, Real-time Transfer Protocol (RTP) and Real-time Transfer Control Protocol (RTCP). PoC uses Presence, Group Management and Instant Messaging as enablers to provide information to the users as well as to the PoC service. This led alongside the standardization of PoC to the definition of Presence SIMPLE [8] for Presence and Instant Messaging and XML Documents Management (XDM) [9] for group and list management.
PoC as a public available service never received real acceptance besides the U.S. market, but the concept of abstract application enablers is by now widely used.
Service developers for next generation network based applications, especially those offered by 3rd party service providers will want to make use of the advanced multimedia communication functionalities offered by IMS-based applications. But core communication functionality like voice- and video call control as well as legacy messaging and location will reside at the operator's domain for security reasons and a well-defined integration of the service platforms into the operator's charging and provisioning functionality. Most application developers will also not have the capability and resources to economically develop such complex communication features into their services. The OMA currently standardizes the OMA Service Environment (OSE) [10] as an abstract enabler layer that serves as an access gateway for 3rd parties and operator services. The OSE has been introduced to enable operators with the functionality to provide their communication and application capabilities to users without the need for the application developer to implement such functionality into their applications themselves. An OSE provides therefore abstract programming interfaces for 3rd party service developers. There is one such API that is standardized by the Parlay Group based on Web Services, namely the Parlay X API [11]. Parlay X defines a set of powerful yet simple, highly abstracted, building blocks of telecom capabilities that developers and the IT community can both quickly comprehend and use to generate new applications. Each building block will be abstracted from the set of telecom capabilities exposed by the Parlay X APIs. The capabilities offered by a building block may be homogeneous (e.g. call control only) or heterogeneous (e.g. mobility and presence). A building block will usually not be application-specific. In order to use them from within the Web 2.0 domain a gateway has been developed, which allows access to the Parlay X API via JavaScript.

1.3. The WWW / Web 2.0

The WWW is by nature community-driven, not only with regard to content, but also from a technical point of view. Simple protocols like HTTP, description languages as HTML and CSS, and architecture paradigms (e.g. Representational State Transfer - REST) made the Web successful and its simplicity is the decisive factor for the developer community's acceptance of extensions to the Web technology stack. Web 2.0 is less a question of novel technologies in the Web technology stack, but rather a question of how existing technologies are applied to create services tailored to user communities.
In this respect, client-side active scripting and the inherent capability of HTML to integrate content from different sources play a major role. Active scripts are shipped along with the web content to control content presentation and interactivity. The object based programming language ECMAScript [12], better known as JavaScript, is today's mostly used scripting language for Web pages. In addition to operations on the associated document, all noteworthy Web browsers allow active scripts to self-reliantly utilize the HTTP client interface in a pared-down configuration. This feature of active scripts to access their origin server for the exchange of any message is referred to as Asynchronous Java Script (Ajax) [13]. Although the Ajax API introduces with the XMLHttpRequest [14] just one new language construct the amount of available developer tools based on Ajax show its current importance.
The varieties of client-server interaction, given to active scripts through Ajax include Remote Procedure Call (RPC) and Publish-Subscribe. The representatives for RPC over HTTP in favor of the developer community are XML-RPC [15] and JSON-RPC [16]. The major difference between both can be found in the representation of request and response, i.e. marshalling of method calls and objects. While JSON-RPC utilizes a light-weight, non-standard syntax, XML-RPC is based on W3C's XML. However, RPC frameworks for Ajax usually require

a respective counterpart on the server-side. In practice, tool support for the selected backend platform (e.g. .NET, J2EE, PHP) is often the decisive criterion for the selection of a RPC framework.

3. Geo-blogging Service Enabler

3.1 Design & Implementation

The realization of the geo-blogging service has been transferred to Parlay X Web Services enabling IMS functionality. An overview about this concrete gateway is depicted in the following figure. The implementation is Java- and JavaScript based and the communication between client and gateway has been realized using JSON-RPC.

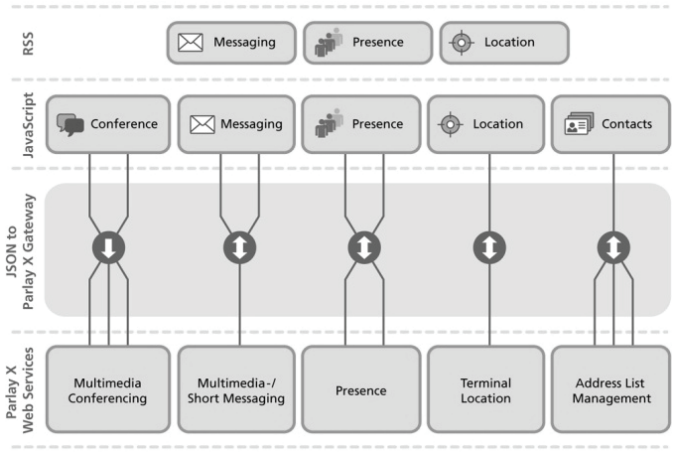


Figure 2: Network abstraction gateway

We make use of the Terminal Location Interface [18] of the Parlay X API to make location information accessbile towards 3rd parties, location information itself is treated as presence information [19] inside the telecommunications infrastructure and respectively stored on a presence server.
To realize the the geo-blogging service, we have extended the OMA XDMS enabler for the storage of RSS feeds that are enriched with geo information. Figure 3 below depicts the architecture that enforces the dynamic distribution of RSS Feeds. The RSS XML Document Management Server (XDMS) and the RSS Checker are illustrated on top of this architecture. These components are controlled by the RSS IMS Enabler and attached to the Open IMS core network. The controlling server offers a web interface that enables the IMS User Agents (UA) to administer their RSS Feeds.

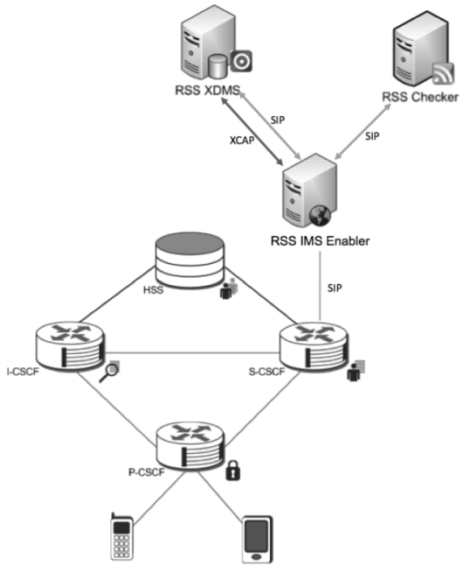


Figure 3: IMS-based RSS distribution architecture

Among other things, the RSS XDMS is in charge of supporting community specific RSS feeds, called RSS blogs. Every registered IMS client is capable of creating and administering its own RSS blog. The interaction between IMS User Agents (UAs) and XDMS occurs via the RSS IMS Enabler. The user controls the XDMS through a HTML web interface and the intentions of the user are forwarded to the XDMS using the XML Configuration Access Protocol (XCAP) [20]. XCAP enforces the creation, deletion and modification of XML documents as well as XML search operations on the XDMS. IMS users can also subscribe to the XDMS via the RSS IMS Enabler.
The remaining application server in figure 3 is the RSS Checker that is responsible for managing the general third party RSS feeds from the World Wide Web domain, such as news feeds, Google Mail or Calendar and various more RSS-based information sources. A user who subscribes to a certain RSS feed will be notified whenever there is a change to the XML file. Hence, the RSS Checker periodically invokes the accordant RSS feed for updates. In the event of content changes the RSS Checker will send SIP Notify requests to the previously subscribed RSS IMS Enabler. The incoming notification is parsed, formatted and delivered as an SIP Message request to the IMS parties that have subscribed to the RSS XML changes. These application servers support the integration of Web 2.0 RSS Feeds and Blogs into the IMS world. Figure 4 depicts the signaling flow chart in detailed.

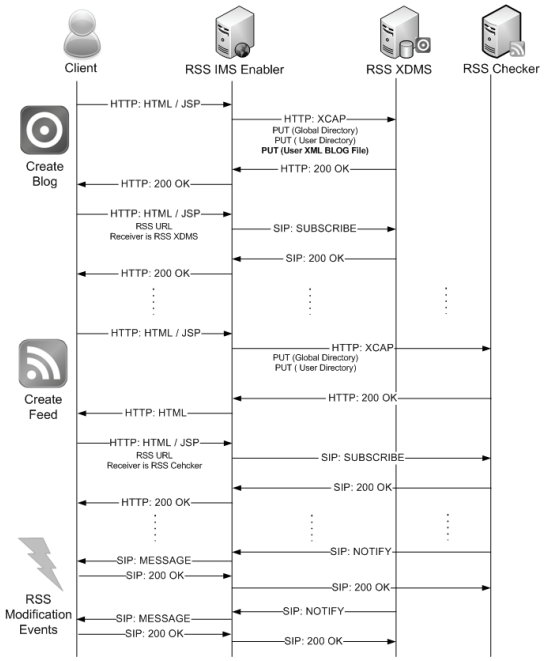


Figure 4: IMS-based RSS signalling diagram

Alongside the notification approach to RSS feed changes, we included furthermore location information support, since the integration of geographical data could facilitate the creation of location-based services. In order to publish location information, the RSS </gid> tag has been used to include geo data.

```
<gidisPermaLink="true">
   52.514497;13.350105
</gid>
```

Within the following section, the implementation of the Facebook “IMS Community Mash-up” application is introduced as a proof-of-concept.

3.2 Proof-of-Concept / Demonstrator

The realization of our mash-up is based on the community platform Facebook [21]. Facebook provides an API that allows developer to easily plug their applications into the portal getting access to Facebook user data, for example, retrieving information about all buddies of a user. The access on community-related information allows an easy mash-up of telecommunications features with user data. In order to demonstrate a meaningful way to use information from within the IMS, a map has been integrated into the application to present presence

and location information in the Web as well as the geo-blogs. For this integration the Google Maps API has been used, but could have been exchanged by any other map API (e.g. Yahoo Maps). Additionally, user information from Facebook has been merged into the map to follow the Web2.0 idea of information mash-up. Telecommunications specific-features have been realized in the following modules that each encapsulates a single functionality:

- **Multimedia Conference**: Voice conferencing through IMS. Participants can be contacted within the Public Switched Telephone Network (PSTN) as well as all-IP Next Generation Networks (NGN).
- **Presence**: Rich presence information, such as online/offline status and mood, of a certain user is published.
- **Terminal Location**: Request location information of a certain user from the IMS. It is necessary that the user provides location information from the IMS client, for example, with the help of a GPS receiver.
- **Short Messaging**: Sending simple text messages (SMS).
- **Multimedia Messaging**: Sending and receiving multimedia messages (MMS), consisting of plain text as well as application data. These messages are sent instantly and are also used for Instant Messaging (IM). A feedback channel has been also realized for this module to receive incoming messages.
- **Address List Management**: Transfer existing user- and group-information from the Web- to the Telco-domain and vice versa.

4. Conclusion & Outlook

In this report, we described our implementation of a geo-blogging service for telecommunications services and a proof-of-concept integration as a Facebook application. Our focus was on the creation of a very high level API that meets the needs and programming paradigms of developers from the WWW community. The term Web 2.0 encompasses a lot of different technologies including, among others, RSS feeds and blogs. The depicted integration of SIP and HTTP to realise event based push approach enhances the way of being dynamically updated to modifications through the system instead of statically polling a server for new information. Due to the presented concept we accomplished the integration of WWW RSS feeds into the NGN domain. In addition this solution improved the Web 2.0 by bonding user-generated content of the IMS community that is enriched with mobile information (location, presence, etc.) from the telco domain.
Future work will be done on security issues related to exposing core network capabilities to the WWW. Furthermore the integration of an OMA compliant Policy Enforcer as part of OSE will be implemented to provide flexible mechanisms for Service Level Agreements (SLAs) between an operator and the service provider using the gateway.

5. References

[1] BT. Web21C SDK. http://web21c.bt.com/.
[2] 3GPP. TS 23.228. IP Multimedia Subsystem (IMS). Stage 2 v.7.10.0. 2007.
[3] ETSI. http://www.etsi.org/tispan/.
[4] H. Schulzrinne, et al.. IETF RFC 3261. SIP: Session Initiation Protocol. 2002.
[5] P. Calhoun. IETF RFC3588. Diameter Base Protocol. 2003.
[6] Maes S., “Pragmatic Approaches to True Convergence with or without IMS”, Proc. of Network Operations and Management Symposium Workshops, NOMS pp. 44-51, 2008. Workshops 2008, ISBN 978-1-4244-2067-4
[7] Open Mobile Alliance (OMA). *Enabler Release Definition for Push-to-talk over Cellular. Candidate Version 2.0 – 11 Dec 2007.* 2007.
[8] Open Mobile Alliance (OMA). *Presence SIMPLE Architecture Document. Approved Version 1.0.1 – 28 Nov 2006.* 2006.
[9] Open Mobile Alliance (OMA). *XML Document Management Architecture. Candidate Version 2.0 – 24 Jul 2007.* 2007.
[10] Open Mobile Alliance (OMA). *OMA Service Environment. Approved Version 1.0.4 – 01 Feb 2007.* 2007.
[11] Parlay X Web Services API, http://www.parlay.org/en/specifications/pxws.asp
[12] ECMAScript Language Specification, 3’rd Edition, 1999, http://www.ecma-international.org/ publications/standards/Ecma-262.htm
[13] J. J. Garrett. Ajax: A new Approach to Web Applications. 2005.
[14] The XMLHttpRequest Object, W3C Working Draft, October, 2007 http://www.w3.org/TR/XMLHttpRequest/
[15] D. Winer; XML-RPC Specification, June, 1999, http://www.xmlrpc.com/spec
[16] JSON-RPC Specification 1.1, Working Draft, August, 2006, http://json-rpc.org/wd/JSON-RPC-1-1-WD-20060807.html
[17] Blum, N., Linner, D., Krüssel, S., Magedanz, T. and Steglich, S.; "Definition of a Web 2.0 Gateway for 3rd Party Service Access to Next Generation Networks", 2008, in IFIP International Federation for Information Processing, Volume 284; Wireless and Mobile Networking; Zoubir Mammeri; (Boston: Springer), pp. 247-258, ISBN 978-0-387-84838-9
[18] ETSI, “Open Service Access (OSA); Parlay X Web Services; Part 9: Terminal Location (Parlay X 2)”, ETSI ES 202 391-9 v0.0.6 (2007-06)
[19] J. Peterson, “IETF RFC 4119;: A Presence-based GEOPRIV Location Object Format, December 2005
[20] J. Rosenberg, ‘‘RFC 4825 – The Extensible Markup Language (XML) Configuration Access Protocol (XCAP)”, May 2007.
[21] Facebook, http://www.facebook.com

Kontextbasierte Adressierung und Routing in mobilen Ad-hoc-Netzwerken

Robert Eigner

eigner@in.tum.de

Christoph Mair

christoph.mair@gmail.com

Technische Universität München
Boltzmannstraße 3
85748 Garching

Abstract

In dieser Arbeit wird eine grundsätzlich andere Art der Adressierung und des Routing für mobile Ad-hoc Netze auf der Basis von Kontextinformationen wie z. B. wie Luftdruck, Helligkeit, Windrichtung und -stärke sowie oder GPS-Position untersucht und anhand eines Beispielszenarios prototypisch implementiert. Der häufigste Anwendungsfall kontextbasierter Adressierung ist die Gruppenkommunikation: Ein Teilnehmer versendet eine Nachricht an eine nicht festgelegte Anzahl von Empfängern, gibt aber an, in welchem Kontext die Nachricht für einen potentiellen Empfänger sinnvoll sein könnte. Im Gegensatz zu infrastrukturbasierten Netzen legt also nun nicht mehr der Sender mittels einer Adresse fest, wer der Empfänger seiner Nachricht sein soll, sondern jeder Empfänger entscheidet anhand seines lokalen Kontexts selbst, ob sie für ihn bestimmt ist und ob sie erneut ausgesendet werden soll. Die Modellierung des dafür nötigen Anwendungswissens erfolgt als Ontologien in OWL. Als Beispielszenario wird eine Windböenwarnung auf Autobahnen verwendet, die über ein vehicular ad-hoc network (VANET) an alle Fahrzeuge verteilt werden soll, die sich auf der gleichen Route befinden, die den Ort enthält, an dem die Windböe detektiert wurde. Die Leistungsfähigkeit des Ansatzes wird durch eine Simulation gezeigt, die mit Hilfe des JiST/SWANS Simulators für mobile Ad-hoc Netze durchgeführt wurde. Die Ergebnisse zeigen, dass die Anzahl der Nachrichten, die nötig sind, um Fahrzeuge in einer bestimmten Umgebung vor der Windgefahr zu warnen auf ungefähr die Hälfte reduziert werden kann – im Vergleich zu einem simplen Fluten des Netzes.

1. Einführung und Motivation

Die stetig zunehmende Zahl mobiler miteinander vernetzter Geräte, wie z. B. Mobiltelefone, PDAs und in letzter Zeit auch Fahrzeuge (VANETs) stellt die Anbieter von drahtlosen Kommunikationsnetzen vor ständig neue Herausforderungen, u. a. auch hinsichtlich der Verteilung und Verwaltung der Adressen der Endgeräte. Die Zuteilung kann entweder manuell oder automatisch erfolgen, wobei Letzteres in großen Netzen zur Beibehaltung einer vernünftigen Relation zwischen Nutzen und Aufwand unbedingt nötig ist. Was in Netzwerken mit einer zentralen Koordinierungsstelle noch gut funktioniert, stellt die Betreiber von drahtlosen Ad-hoc-Netzen, welche grundsätzlich eine hohe Fluktuationsrate der Teilnehmer aufweisen und oft über keine feste Infrastruktur verfügen, vor neue Probleme.
In diesem Artikel soll eine grundsätzlich andere Art der Adressierung untersucht und prototypisch implementiert werden. Die so genannte kontextbasierte Adressierung kombiniert eine Menge von im Umfeld des Teilnehmers verfügbaren Informationen zu einer Adresse. Dies können zum Beispiel Umwelteinflüsse wie Luftdruck, Helligkeit, Windrichtung und -stärke sowie Parameter des Endgerätes, GPS-Position oder Ladezustand der Batterie sein.
Kapitel 2 entwirft ein beispielhaftes Szenario, in dem die in Kapitel 3 entworfenen Kontext-Ontologien und Konzepte eingesetzt werden können. Abschnitt 4 geht auf das entwickelte Routing-Verfahren ein, das mit Hilfe des in Kapitel 5 beschriebenen Prototypen simuliert wurde. Die Ergebnisse der Simulation werden in Kapitel 6 dargestellt. Kapitel 0 schließt den Artikel mit eine kurzen Fazit ab.

2. Szenario

Zum Testen der kontextbasierten Adressierung wurde folgendes Beispielszenario aus dem Bereich der Fahrzeug-zu-Fahrzeug-Kommunikation gewählt:
Ein Fahrzeug ist an einem windigen Tag in hügeligem Gelände auf der Autobahn unterwegs. Auf Brücken übt der kräftige und in Böen auftretende Seitenwind je nach Fahrzeugtyp eine unterschiedlich große Kraft quer zur Fahrtrichtung des Fahrzeugs aus, was bei hohen Geschwindigkeiten leicht zu einem Abdriften aus der Fahrspur führen kann. Zum Erkennen einer Windbö werden dreierlei Messdaten benötigt:

- *Geschwindigkeit*: Die Geschwindigkeit ist in allen modernen Fahrzeugen über Bussysteme verfügbar.
- *Querbeschleunigung*: Die vom Wind verursachte, quer zur Fahrtrichtung wirkende Beschleunigung. Diese Messdaten werden für das ESP benötigt und sind demnach auch verfügbar.
- *Lenkwinkel*: Der aktuelle Einschlag der Vorderräder bestimmt neben der Geschwindigkeit den Radius einer Kurve maßgeblich mit. Mit diesem Messwert kann eine kurveninduzierte von einer von äußeren Einflussfaktoren erzeugten Querbeschleunigung unterschieden werden.

Eine Warnung wird durch ein Fahrzeug ausgelöst, welches eine plötzliche Querbeschleunigung ohne nennenswerten Lenkeinschlag registriert. Aus der gemessenen Querbeschleunigung lassen sich fahrzeugspezifisch Rückschlüsse auf die ungefähre Windstärke ziehen. Diese wird zusammen mit den Koordinaten des Auftretens an weitere Fahrzeuge, welche sich auf derselben Straße oder in deren Nähe befinden, gesendet.
Adressaten der Nachricht sind alle Fahrzeuge, die in absehbarer Zeit in eine ebensolche Gefahrensituation geraten könnten. Die Gefahrenstelle kann durch Koordinaten und einen eindeutigen Bezeichner für die Straße, wie sie Navigationssysteme verwenden, identifiziert werden. Um möglichst viele Fahrer zu warnen, soll die Nachricht für beide Fahrtrichtungen sowie auch andere in der Nähe verlaufende Strecken gültig sein. An diesem Beispiel ist gut zu erkennen, dass die Anwendung es erfordert, Nachrichten eher an Fahrzeuge mit bestimmten Eigenschaften zu richten als an Netzwerkknoten mit einer a priori nicht bekannten Adresse. Da die Adresse aus den Kontextinformationen gewonnen wird, bleibt diese nicht konstant, sondern ändert sich sobald sich das Umfeld des Teilnehmers ändert. Diese Eigenheit muss bei der Entwicklung von Routing-Protokollen berücksichtigt werden, da eine bidirektionale eins-zu-eins Kommunikationsbeziehung unter den genannten Bedingungen nur schwer möglich ist. In diesem Szenario versendet also ein Teilnehmer eine Nachricht an eine nicht festgelegte Anzahl von Empfängern. Diese entscheiden dann, ob die Nachricht für sie bestimmt ist und ob sie weiterverarbeitet oder verworfen wird.

3. Kontextmodell

Alle Kontextinformationen werden mit OWL [1] als Ontologien modelliert. Das hat den Vorteil, dass eine umfassende semantische Beschreibung der Fahrzeuge mit deren Eigenschaften sowie der Umweltsituation relativ problemlos möglich ist. Aus der so modellierten Faktenbasis lässt sich durch einen Reasoner weiteres Wissen ableiten, um neue Ereignisse klassifizieren und darauf reagieren zu können. Dadurch ist es möglich, mit relativ

wenig Code auch komplexe Sachverhalte zu untersuchen und diese an neue Erfordernisse oder weitere Anwendungen wie z. B. in [2] anzupassen. Für das gewählte Szenario werden zwei Ontologien modelliert: Eine mit den Informationen zur Warnung und eine zweite, die den Adressatenkreis bestimmt.

3.1. Warnungs-Ontologie

Um eine Windbö zu erkennen, werden die in der Szenariobeschreibung angegebenen Messwerte ermittelt. Eine gemessene Querbeschleunigung wird nur dann als Windbö klassifiziert, wenn diese einerseits groß genug ist, um Unebenheiten der Fahrbahn als Ursache ausschließen zu können und die Vorderräder zum Zeitpunkt der Messung keine nennenswerten Winkeländerungen verzeichneten, die als Lenkbewegung interpretiert werden könnten. Die aktuelle Geschwindigkeit wird erfasst, da starker Seitenwind vor allem bei hohen Geschwindigkeiten gefährlich ist. Wird eine entsprechende Konstellation von Messwerten festgestellt, werden diese in eine neue Ontologie eingetragen und sowohl im Fahrzeug ausgewertet als auch an andere Fahrzeuge versandt. Als `Ereignis` werden nur Beobachtungen akzeptiert, welche die Objekteigenschaft `hat_Messung` besitzen. Diese ist ein Attribut mit dem Definitionsbereich `Ereignis` und dem Wertebereich `Messung`, weshalb eine Beobachtung mit diesem Attribut vom Reasoner als `Ereignis` klassifiziert werden kann. Die einzelnen Messwerte werden als Subklassen von `hat_Messwert` als Datentyp-Eigenschaften `hat_geschwindigkeit`, `hat_querbeschleunigung` und `hat_lenkwinkel` modelliert. Diese Definition von `hat_Messwert` dient dazu, einen oder mehrere Messwerte an eine Instanz der Klasse `Messung` zu binden. Aufbauend auf dieser Grundlage kann nun eine `Warnung` als Subklasse eines `Ereignisses` modelliert werden.
Die Definition der Klasse `Seitenwind` (als Subklasse von `Ereignis`) ist in Abbildung 1 in OWL Functional Syntax notiert. Diese ist deutlich kürzer und semantisch äquivalent zur Darstellung mittels RDF/XML: Durch das Schlüsselwort `EquivalentClasses` mit den Parametern `Seitenwind` als Indiz für die Warnung und einer Schnittmenge von Klassen, definiert durch `ObjectIntersectionOf`, wird eine `Messung` auf eine Warnung abgebildet. Es wird jeweils die Existenz der einzelnen Attribute `hat_geschwindigkeit`, `hat_lenkwinkel` und `hat_querbeschleunigung` gefordert, für deren Belegung nur eine begrenzte Menge an Werten zulässig ist. Dies sind Werte vom Typ `double` mit entsprechend nach oben oder unten beschränktem Wertebereich. Zum Schluss wird noch gefordert, dass `Seitenwind` eine Subklasse von `Ereignis` und mit der hier nicht näher spezifizierten Klasse `Glatteis` nicht äquivalent ist. Diese Struktur erlaubt es dem Reasoner, einer Instanz, welche über die drei Messwert-Attribute mit der geforderten Belegung verfügt, den Typ `Seitenwind` zuzuordnen.

```
EquivalentClasses(
   Seitenwind
   ObjectIntersectionOf(
      ObjectSomeValuesFrom(
         hat_Messung DataAllValuesFrom(
         hat_geschwindigkeit DatatypeRestriction(
            xsd:double minInclusive "8"^^xsd:int)))
      ObjectSomeValuesFrom(
         hat_Messung DataAllValuesFrom(
         hat_lenkwinkel DatatypeRestriction(
            xsd:double maxInclusive "5"^^xsd:int)))
      ObjectSomeValuesFrom(
         hat_Messung DataAllValuesFrom(
         hat_querbeschleunigung DatatypeRestriction(
            xsd:double minInclusive "10"^^xsd:int)))
   )
)
SubClassOf(Seitenwind Ereignis)
DisjointClasses(Seitenwind Glatteis)
```

Abbildung 1: Die Klasse `Seitenwind`

3.2. Adress-Ontologie

Damit die als Broadcast versendeten Pakete nicht von jeder Anwendungssoftware ausgewertet werden müssen – die übertragenen Ontologien können durchaus umfangreicher sein als die für dieses Szenario modellierten – ist es sinnvoll, eine „Adressierung“ einzuführen, da die Adressauswertung i. A. einfacher ist als die Bearbeitung der Nutzdatenontologie. Die Zieladresse wird ebenfalls als Ontologie modelliert, um Kontextinformationen einzubeziehen. Für das hier modellierte Szenario ist eine Adressierung anhand der zukünftigen Route des Verkehrsteilnehmers denkbar. Eine Warnung soll an all jene Fahrzeuge adressiert werden, welche sich auf dem Weg zur Gefahrenstelle befinden. Wenn die Verkehrsteilnehmer durch ihre zukünftige Route identifiziert werden, dann ist diese Adresse weder zwingend eindeutig noch konstant, da sich die Fahrzeuge entlang ihrer Route weiterbewegen, bereits besuchte Teile aus der Route löschen und neue Wege hinzufügen. Ein Streckenabschnittbasiertes Adressierungsverfahren lässt sich relativ einfach mit OWL implementieren. Die zukünftige Route eines Fahrzeugs kann in dessen Adressontologie z. B. als Liste von Streckenabschnittsnummern dargestellt werden.

Adressaten einer Meldung sind trivialerweise erst einmal alle Fahrzeuge, die sich in unmittelbarer Umgebung der Gefahrenstelle befinden. Diese doch sehr unscharfe Adressierung kann durch weitere Einschränkungen noch verbessert werden:

- Die Warnung ist nur für die Fahrzeuge interessant, welche die Gefahrenstelle auch wirklich passieren wollen. Das Navigationsgerät kann dies mit Hilfe der geplanten Route sehr einfach entscheiden.
- Die Wahrscheinlichkeit, dass die vom Navigationsgerät geplante Route auch tatsächlich befahren wird, ist abhängig von der Qualität der Routenplanung und dem Fahrer, der die Anweisungen nicht befolgen oder auch falsch interpretieren kann. Deshalb ist eine Warnung auch für diejenigen Fahrzeuge interessant, die laut der aktuellen Routenplanung die Gefahrenstelle zwar nicht passieren, sich aber darauf zu bewegen beziehungsweise in deren Nähe fahren. Ein falscher Abbiegevorgang kann schnell zu einer Neuberechnung der Route unter Einbeziehung des gefährlichen Streckenabschnitts führen. Deshalb sollen auch Fahrzeuge, die in Richtung des Gefahrenbereichs unterwegs sind die Warnung erhalten.

4. Routing

Die kontextbasierte Adressierung kann weder Eindeutigkeit noch Konstanz der Adressen garantieren. Diese Eigenheiten führen dazu, dass bidirektionaler Unicast nicht zuverlässig funktioniert. Damit scheiden Protokolle, welche Routen suchen und speichern wie z.B. AODV oder OSPF als mögliche Kandidaten aus. Da eine selektive Adressierung von Teilnehmern zwar möglich, aber einzelne Stationen nicht zwingend unterscheidbar sind, bleibt das so genannte Selektive Fluten, eine Art Broadcast-Routing als mögliche Lösung: Nachrichten werden an alle Stationen in der Umgebung geschickt – dies wird durch die Funkschnittstelle implizit erledigt – und der Empfänger entscheidet, ob er das Paket weiterleiten (nochmals aussenden) soll oder nicht. Dabei wird eine Nachricht anstatt an die vom Sender spezifizierte Zielgruppe von den einzelnen Stationen im Netzwerk zu einer von ihnen als gewünscht vermuteten Zielgruppe weitergeleitet.
Die Entscheidung, ob eine erneute Übertragung notwendig ist, kann ein Teilnehmer anhand verschiedener Kriterien und mittels des in den Ontologien modellierten Zusatzwissens selbst treffen. Die Bedingungen zum Weitersenden von Nachrichten können wieder mit OWL implementiert werden. Die Positionsberechnung kann entweder über das bereits bekannte Matching von Klassen oder aber durch SWRL-Regeln [3] implementiert werden. Der regelbasierte Ansatz ist dabei um einiges flexibler und einfacher realisierbar, wird aber nicht von allen Reasonern unterstützt.

Selektives Fluten

Der entwickelte Prototyp implementiert die Routing-Entscheidung mit Java-Logik. Um die Nachrichtenflut im Netzwerk einzudämmen, darf eine Nachricht höchstens einmal pro Fahrzeug versendet werden. Diese Schutzmaßnahme wird durch das Speichern eines URI für jedes empfangene Paket, z. B. den Hashwert der Nachricht und einen Zeitstempel, welcher den Messzeitpunkt festhält, realisiert. Dieses Verfahren stellt sicher, dass Duplikate erkannt werden. Der Speicher ist als begrenzter Ringbuffer organisiert, sodass nach einer gewissen Anzahl an empfangenen Paketen die ältesten wieder überschrieben werden. Die Größe des Ringbuffers muss an das Verkehrsaufkommen der Funkschnittstelle angepasst werden und ist ein Kompromiss zwischen sicherer Vermeidung von Duplikaten und Speicherverbrauch. Die Routing-Implementierung ist auf Ebene drei des ISO/OSI Schichtenmodells angeordnet. Alle empfangenen Pakete durchlaufen neben der Adressauswertung der Netzwerkschicht auch die Routing-Implementierung. Diese speichert bisher noch unbekannte Nachrichten und plant für einen zukünftigen Zeitpunkt deren erneutes Aussenden. Die Wartezeit ist notwendig, weil die Sendeversuche von mehreren Fahrzeugen, die die Nachricht gleichzeitig zum ersten Mal empfangen, sofort zu einer Kollision auf der Funkstrecke führen würden. Eine während der Wartezeit empfangene Nachricht durchläuft denselben Prozess wie alle anderen Nachrichten. Wird sie jedoch als Dublette einer bereits zur Weiterverbreitung geplanten Nachricht identifiziert, muss diese aus der Warteschlange genommen werden, um das Netz nicht durch eine erneute, oftmals nutzlose Wiederholung zu belasten.

5. Implementierung

Im folgenden Abschnitt werden einige Details zur Abarbeitungsreihenfolge einer empfangenen Nachricht skizziert; Abbildung 2 verdeutlicht den Weg einer Information oberhalb von ISO/OSI-Schicht zwei.
Die Netzwerkkomponente auf Schicht drei erhält von Schicht zwei ein kürzlich empfangenes Datenpaket. Nachdem eine Kopie davon an die Routing-Komponente weitergeleitet wurde, muss geklärt werden, an wen die empfangene Nachricht adressiert war. Dazu wird zuerst die URI der Adressontologie mit einer Liste von bereits bekannten URIs verglichen. Um den Speicherverbrauch zu begrenzen, werden nur die jeweils 1000 letz-

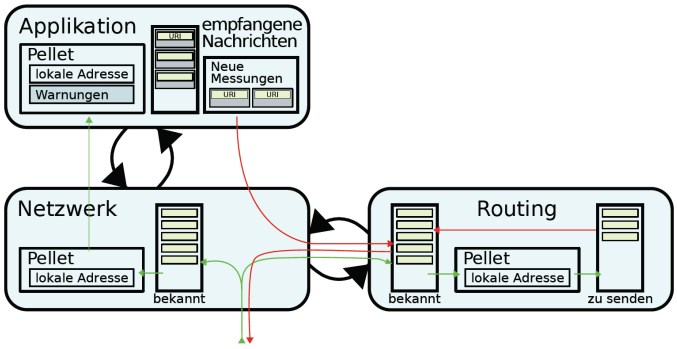


Abbildung 2: Architektur innerhalb eines Netzwerkknotens

ten URIs in einer FIFO-Warteschlange gespeichert. Bei sehr hohem Nachrichtenaufkommen muss dieser Wert eventuell angepasst werden. Weil eine Nachricht über die URI eindeutig identifizierbar ist, können empfangene Duplikate schnell aussortiert werden.
Wurde der erste Test bestanden und handelt es sich um eine noch unbekannte Nachricht, wird eine Instanz des Reasoners damit beauftragt, herauszufinden, ob sich der in der Adressontologie gespeicherte Straßenabschnitt, auf dem die Messung durchgeführt wurde, auch in der Liste der eigenen zukünftigen Route wiederfindet. In diesem Fall ist die Nachricht möglicherweise eine Warnung, die dem Fahrer mitgeteilt werden muss. Deshalb wird sie an die Applikation weitergeleitet. Wurde die Nachricht als uninteressant klassifiziert, wird sie von der Netzwerkkomponente verworfen. Die Routing-Komponente bearbeitet ihre Kopie der Nachricht völlig unabhängig vom restlichen Kontrollfluß. Die Überprüfung auf empfangene Duplikate findet auch hier nach dem gleichen Schema statt, aber es wird zusätzlich zur URI auch eine zufällige Wartezeit im Bereich von null bis fünf Sekunden gespeichert, nach welcher eine erneute Aussendung der Nachricht stattfinden soll. An dieser Stelle sind auch weitere intelligentere Mechanismen zur erneuten Aussendung einer Nachricht denkbar, die auf anderen Kontextdaten (z. B. Position oder Netzauslastung) basieren.
Eine selbst erzeugte Nachricht durchläuft beim Senden ebenfalls die Netzwerk- und die Routing-Komponente. Die Netzwerkschicht erhält von der Applikation eine Nachricht, welche umgehend an das Routing-Modul weitergeleitet wird, wo die URI notiert wird, um später empfangene Duplikate zu erkennen. Anschließend übernimmt die Netzwerkkomponente die Nachricht und beauftragt Schicht zwei mit deren Versand.

6. Evaluation

Zur Untersuchung des Netzwerkverhaltens wurde der Simulator SWANS++ [4] verwendet. Dabei wird das STRAW-OD Mobilitätsmodell zur Simulation der Fahrzeugbewegung verwendet, weil es eine für das Szenario wichtige Eigenschaft mitbringt: Die Bewegung ist nicht zufällig, sondern folgt einer zuvor festgelegten Route. Diese wird dann als Kontextinformation zur Adressierung von Fahrzeugen verwendet. Das simulierte Straßennetz basiert auf realistischen digitalen Kartendaten der Stadt Boston, der zur Simulation herangezogene Straßenausschnitt ist ca. 1000m x 750m groß.
Auf Schicht zwei des ISO/OSI–Referenzmodells wird der Medienzugriff simuliert. Die MAC–Schicht bringt neben den Kanalzugriffsverfahren auch eine Adressierung der Teilnehmer über die MAC-Adressen mit. Grundsätzlich wird diese Adressierung nicht benötigt und könnte durch die Kontextmodelle substituiert werden. Die Möglichkeit der eindeutigen Adressierung von Teilnehmern im unmittelbaren Umfeld kann jedoch für spätere Erweiterungen der kontextbasierten Adressierung genutzt werden.
Zum Umgang mit Ontologien wurde OWL-API [5] verwendet, als Reasoner kam Pellet[1] zum Einsatz. Alle Komponenten lassen sich gut mit den in Java geschriebenen Anwendungen integrieren.
Zur Erprobung der entwickelten Verfahren wurden mit dem Prototypen mehrere Simulationsläufe mit unterschiedlichen Parametern durchgeführt. Die Fahrzeugdichte wurde in Zweierschritten auf bis zu 48 Fahrzeuge pro Simulationsdurchlauf erhöht. Im simulierten Szenario wurde als zusätzliche Bedingung zur Weiterleitung einer Nachricht eine Gültigkeitsbeschränkung eingeführt: Fahrzeuge sendeten eine Nachricht nur dann neu aus, wenn sie sich in einem bestimmten Umkreis um das berichtete Ereignis befanden. Jede dieser Konfigurationen wurde mit 10 verschiedenen Gültigkeitsradien (zwischen 0m und 450m, jeweils in 50m-Schritten) und ohne jegliche Gültigkeitseinschränkung der Nachrichten simuliert. Um die Einflüsse verschiedener Signalausbreitungsmodelle auf die Performanz der ISO/OSI-Schicht 3 zu untersuchen, wurden die Simulationsläufe jeweils mit dem Shadowing-Pathloss-Modell und mit dem Two-Ray-Pathloss-Modell durchgeführt.
Zur Klärung der Frage, ob sich der erhöhte Implementierungsaufwand für selektive Flooding-Verfahren im Vergleich zu simplen Flooding-Algorithmen lohnt, werden in Abbildung 3 die Anzahl der verschickten Nachrichten beider konkurrierenden Verfahren gegenübergestellt. Das Fluten des gesamten Netzes erfordert wie erwartet einen weitaus höheren Aufwand als das durch die Gültigkeitsbeschränkung limitierte selektive Fluten – der Graph steigt ungefähr doppelt so schnell. Die Frage, inwieweit die semantisch aufwendige Implementierung

[1] `http://pellet.owldl.com`

gerechtfertigt wäre, wird in Abbildung 4 beantwortet. Einer Zunahme der gewarnten Fahrzeuge um etwa 10% steht eine Steigerung des Nachrichtenaufkommens um bis zu 100% gegenüber. Der Graph zeigt die Ergebnisse bei einer durchschnittlichen Gültigkeitsbeschränkung von 225m.
Diese Simulationsdaten bestätigen, dass sich selbst die Implementierung von einfachen selektiven Algorithmen lohnt, weil das Netz dadurch stark entlastet und zusätzliche Übertragungskapazität für andere Nachrichten frei wird.

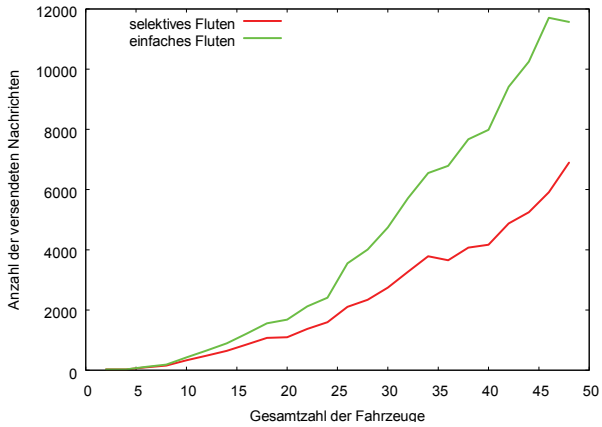


Abbildung 3: Vergleich der Netzbelastung zwischen einfachem und selektivem Fluten. Während die Zahl der erzeugten Nachrichten gleich bleibt, übersteigt die Zahl der insgesamt durch Wiederholung versandten Nachrichten beim einfachen Fluten die selektive Variante bis um den Faktor 2

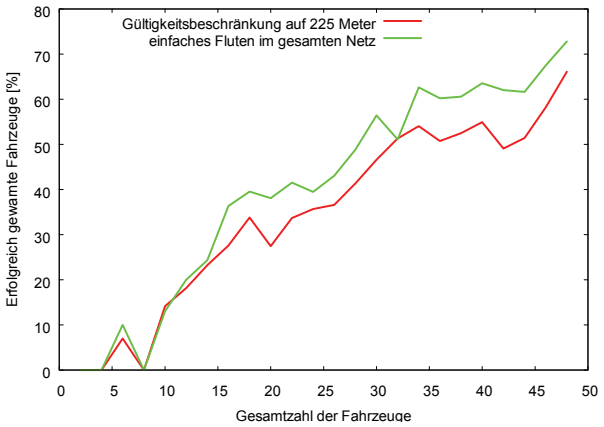


Abbildung 4: Vergleich des Warnerfolgs zwischen einfachem und selektivem Fluten. Bei ersterem werden zwar mehr Fahrzeuge gewarnt, dieser Erfolg steht aber mit einem Gewinn von maximal 15% in keinem Verhältnis zum zu leistenden Mehraufwand, welcher laut Abbildung 3 um bis zu 100% zunimmt.

7. Fazit

Die Simulationsergebnisse aus Kapitel 6 zeigen, dass die kontextsensitive Adressierung für Warnsysteme in Fahrzeug-zu-Fahrzeug-Netzwerken erfolgreich eingesetzt werden kann. Das selektive Fluten des Netzwerks erreicht selbst mit einfachen Techniken gute Ergebnisse bei der Zustellung der Warnungen. Im Vergleich zu einfachem Fluten wird das Netz, bei ähnlich hohem Warnerfolg, weit weniger stark belastet. Um die Universalität der Einsetzbarkeit von kontextbasierten Adressierungs- und Routing-Verfahren darzustellen, reicht das behandelte Szenario leider nicht aus. Die Untersuchung weiterer Szenarien, welche verschiedenste Anforderungen an Modelle und Algorithmen stellen ist demnach zur weiteren Erforschung der Möglichkeiten der kontextbasierten Adressierung notwendig. Die grundsätzliche Einsetzbarkeit von Kontextmodellen, zur Adressierung von Knoten in einem mobilen Ad-hoc-Netzwerk sowie zur Ableitung von Routing-Metriken, konnte mit dem Prototypen erfolgreich gezeigt werden.
Die kontextbasierte Adressierung lässt sich auch für Bereiche außerhalb des Fahrzeug-zu-Fahrzeug-Umfelds einsetzen. Der Routing-Algorithmus sollte jedoch an die jeweiligen Erfordernisse angepasst werden, um die besten Ergebnisse zu erzielen. Eine allgemein gehaltene Implementierung ist vorstellbar, wobei die Modellierung der Adressontologien entsprechend komplizierter wird.

8. Literatur

[1] Patel-Schneider, P. F., P. Hayes und I. Horrocks: *OWL Web Ontology Language – Semantics and Abstract Syntax*. W3C Recommendation, W3C, 2004. `http://www.w3.org/TR/owl-semantics/`.
[2] Eigner, R. und G. Lutz: *Collision Avoidance in VANETs - An Application for ontological context models*. In: Proc. 5th IEEE Workshop on Context Modeling and Reasoning (CoMoRea 2008), Hong Kong, PR China, 2008
[3] Horrocks, I., P. F. Patel-Schneider, H. Boley, et al.: *SWRL: A Semantic Web Rule Language Combining OWL and RuleML*. W3C Member Submission, W3C, 2004. `http://www.w3.org/Submission/SWRL/`.
[4] Barr, R.: SWANS – Scalable Wireless Ad hoc Network Simulator Users Guide. `http://jist.ece.cornell.edu/docs/040319-swans-user.pdf`, 2004.
[5] Horridge, M., S. Bechhofer und O. Noppens: *Igniting the OWL 1.1 Touch Paper: The OWL API*. In: Proc. 3rd International Workshop on OWL Experiences and Directions (OWLED 2007), 2007

Tracking von Fahrerlosen Transportfahrzeugen mittels drahtloser Sensornetzwerke und Erweitertem Kalman-Filter

Dipl.-Inf. (FH) Sarah Spieker

Joseph-von-Fraunhofer-Str. 2-4
44227 Dortmund

Prof. Dr. Christoph Röhrig

Emil-Figge-Str. 42
44227 Dortmund

Marcel Müller

Emil-Figge-Str. 42
44227 Dortmund

Abstract

Im Bereich der Lagerverwaltung ist die Optimierung der Lager- und Fördertechnik von großem Interesse. Lagerverwaltungssysteme (LVS) werden eingesetzt, um alle wichtigen Informationen in einer zentralen Instanz zu verwalten. In herkömmlichen LVS werden diese Daten in einer Datenbank hinterlegt, sodass Veränderungen im Lager manuell aktualisiert werden müssen. Da diese Art der Lagerorganisation mit hohem Aufwand verbunden ist, ist eine weitgehende Automatisierung der Informationsübermittlung wünschenswert.
Drahtlose Sensornetzwerke optimieren die Lagerverwaltung durch eine Dezentralisierung der Datenspeicherung und bieten die Grundlage zur Positionserfassung. Besonders interessant ist der Einsatz bei der Kursverfolgung (Tracking) beweglicher Objekte. Über die Interaktion der Sensorknoten kann kontinuierlich die aktuelle Position eines Fahrzeugs erfasst werden.
Dieser Artikel befasst sich mit der Kursverfolgung eines Fahrerlosen Transportfahrzeugs (FTF) in einem Lager. Dazu wird ein konkretes Sensornetzwerk untersucht. Die Lokalisation erfolgt über das verwendete nanoLOC System sowie über das Erweiterten Kalman-Filter mittels Trilateration.

1. Einleitung

Lagerverwaltungssysteme (LVS) werden eingesetzt, um wichtige Informationen in einer zentralen Instanz zu verwalten. Von Interesse sind unter anderem die Position und der Zustand eines Ladungsträgers, wie beispielsweise einer Palette. In herkömmlichen LVS werden diese Daten zentral in einer Datenbank hinterlegt. Veränderungen im Lager müssen regelmäßig in der Datenbank aktualisiert werden. Da diese Art der Lagerorganisation aufwändig und unsicher gegen Ausfälle, ist eine weitgehende Automatisierung der Informationsübermittlung sowie eine dezentrale Verteilung der Daten auf die Ladungsträger selbst erstrebenswert. Dieses Ziel kann durch den Einsatz von drahtlosen Sensornetzwerken, wie beispielsweise mit dem nanoLOC System des Herstellers NANOTRON Technologies, erreicht werden. Über die Interaktion der Sensorknoten kann kontinuierlich die jeweils aktuelle Position von Gütern oder Transportfahrzeugen ermittelt werden, was viele Arbeitsabläufe in einem Lager vereinfacht. Die Ladungsträger lokalisieren sich selber und übermitteln ihre Position an eine zentrale Verwaltungsinstanz. (vgl. [1, S.46f.])

1.1. Themenbezogenen Arbeiten und Forschungsstand

In [2] wurden drahtlose Sensornetzwerke bereits für stationäre Lokalisierungsapplikationen eingesetzt. Dabei wurde die Messgenauigkeit des nanoLOC Systems anhand der Positionsbestimmung von Ladungsträgern in einem Hochregallager (HRL) untersucht. Da ungefähr eine Positionsgenauigkeit von $0{,}5\ \mathrm{m}$ erzielt wurde, bietet das System Potentiale für weitere Untersuchungen im Bereich der Lagerverwaltung und der Transportlogistik, was im in dieser Arbeit näher betrachtet wird.

2. Das nanoLOC System

NANOTRON Technologies entwickelt robuste und energieeffiziente *Real-Time Location Systems* (RTLS) und schafft die Voraussetzung für *Location-Based-Services* (LBS) sowie für *Asset-Tracking-Applications* im zwei- und dreidimensionalen Raum. Das dabei verwendete Ranging-Verfahren *Symmetrical Double-Sided Two Way Ranging* (SDS-TWR) gestattet eine funkbasierte Abstandsmessung anhand der Signallaufzeiten und bietet die Grundlage für eine metergenaue Positionsbestimmung eines mobilen Objektes. Die drahtlose Kommunikation sowie das Ranging-Verfahren sind in einem einzigen Chip, dem Transceiver nanoLOC TRX integriert. Dieses Hochfrequenz-Funkmodul arbeitet in dem weltweit verfügbaren ISM-Band von $2{,}4\ \mathrm{GHz}$. Die drahtlose Kommunikation basiert auf der von NANOTRON patentierten Chirp-Modulationstechnik *Chirp Spread Spectrum* (CSS) nach dem IEEE-Standard 802.15.4a. Die Datenrate liegt zwischen $125\ \mathrm{kbit/s}$ und $2\ \mathrm{Mbit/s}$. (vgl. [3])

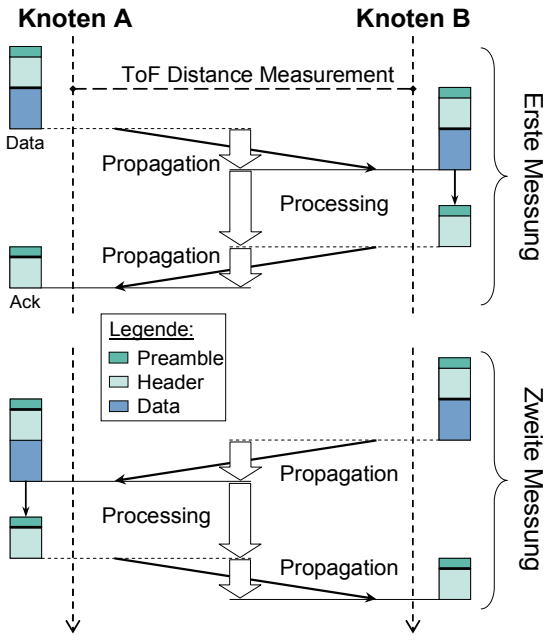


Abbildung 1: SDS-TWR [4]

Der Vorteil von SDS-TWR liegt darin, dass keine zeitliche Synchronisierung zwischen den Stationen erforderlich ist und dadurch bekannte Probleme zeitbasierter Lokalisierungsverfahren behoben werden. Der genaue Ablauf wird in der Abbildung 1 dargestellt. Bei SDS-TWR wird die Signalübertragungszeit in zwei Richtungen gemessen (*Two-Way Ranging*). Dabei entstehen zwei Zeitspannen: Die Signallaufzeit (*Signal Propagation Delay*) die benötigt wird, um ein Datenpaket von Knoten A zu B zu übermitteln und eine Bestätigung zurückzusenden, wird von Knoten A gemessen. Die Zeit um das eintreffenden Datenpaket zu verarbeiten, die Bestätigung zu generieren und um das Versenden vorzubereiten, wird als Verzögerungszeit (*Processing Delay*) bezeichnet und von Knoten B gemessen. Die Differenz der beiden Zeitangaben (Signallaufzeit - Verzögerungszeit) beschreibt somit die zweifache Signalübertragungszeit. Zudem wird eine doppelseitige Messung (*Symmetrical Double-Sided*) durchgeführt, um Fehler wie Uhrenabweichungen (*Clock Drift*) zu eliminieren. Dazu wird das gleiche Verfahren in umgekehrter Laufrichtung durchgeführt, also von Knoten B zu A und zurück zu B. (vgl. [4])

1.2. Lokalisation

Zusammen mit dem Mikrocontroller ATmega128L und dem Hochfrequenz-Funkmodul nanoLOC TRX formt das System eine einsatzfähige Hardware zur drahtlosen Distanzmessung und Lokalisation. Das Sensornetz besteht aus vier fest positionierten Ankerknoten, die als Referenzpunkte innerhalb eines zweidimensionalen Koordinatensystems fungieren, sowie aus einem zu lokalisierenden Tag. In der Abbildung 2 beschreibt $\mathrm{P}(p_x, p_y)$ die Tagposition, $\mathrm{P}(a_{x,i}, a_{y,i})$ die Ankerpositionen und r_i die Distanzen zwischen den Ankern und dem Tag.

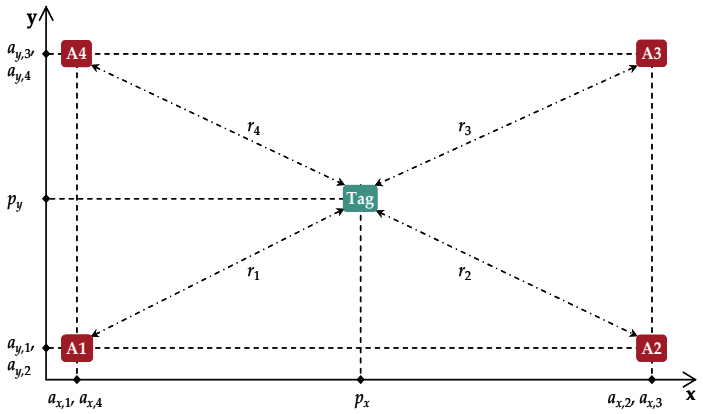


Abbildung 2: Positionsbestimmung mittels Trilateration [2]

Für die Positionsbestimmung stellt NANOTRON zudem eine Auswertungssoftware sowie eine Basisstation, die die Schnittstelle zwischen dem Sensornetzwerk und dem Auswertungscomputer darstellt, zur Verfügung. Der Tag sendet die gemessenen Abstandswerte an die Basisstation, woraufhin die Software über einen nicht veröffentlichten Algorithmus die Position des Tags berechnet. (vgl. [2, S. 1834f.])

3. Positionsbestimmung von Transportfahrzeugen unter Verwendung des Erweiterten Kalman-Filters

Bei der Betrachtung eines dynamischen Systems ist der interne Zustand, wie zum Beispiel die Position eines Fahrerlosen Transportfahrzeugs (FTF), nur über Messungen erfassbar. Jedoch sind Messwerte oftmals ungenau oder mit einem Rauschen überlagert. Das Messrauschen kann jedoch mit dem Kalman-Filter reduziert werden. Das Kalman-Filter beschreibt ein rekursives Verfahren, um den Zustand eines Systems aufgrund von Messungen zu schätzen, mit dem Ziel, den mittleren quadratischen Fehler zu minimieren. Die Grundform des Filters, das Diskrete Kalman-Filter (DKF), schätzt den internen Systemzustands in dem Fall, dass der gesamte Prozess durch lineare Gleichungen dargestellt wird. In dem Fall, dass Prozesszustände geschätzt werden sollen, die durch ein zeitdiskretes, nichtlineares System beschrieben werden, kommt das Erweiterte Kalman-Filter (EKF) zum Einsatz. Die Algorithmen der beiden Filter-Varianten sind in [5] beschrieben.

1.3. EKF Entwurf

Mit dem EKF können die sich verändernden Positionen sowie die Geschwindigkeit des Tags auf Basis der gemessenen Abstände zu den vier Ankern ermittelt werden. Basierend auf der Trilaterationsgleichung gilt folgende Berechnung für die Ankerdistanzen mit $i \in \{1, 2, 3, 4\}$:

$$r_i = \sqrt{(p_x - a_{x,i})^2 + (p_y - a_{y,i})^2} \qquad (1)$$

Um die unbekannte Position des Tags zu ermitteln, werden die vier Gleichungen aus (2) nach p_x und p_y umgestellt und in die folgende Matrixdarstellung transformiert:

$$\begin{pmatrix} 2\cdot a_{x,1} - 2\cdot a_{x,2} & 2\cdot a_{y,1} - 2\cdot a_{y,2} \\ 2\cdot a_{x,1} - 2\cdot a_{x,3} & 2\cdot a_{y,1} - 2\cdot a_{y,3} \\ 2\cdot a_{x,1} - 2\cdot a_{x,4} & 2\cdot a_{y,1} - 2\cdot a_{y,4} \end{pmatrix} \cdot \begin{pmatrix} p_x \\ p_y \end{pmatrix} = \begin{pmatrix} r_2^2 - r_1^2 + a_{x,1}^2 - a_{x,2}^2 + a_{y,1}^2 - a_{y,2}^2 \\ r_3^2 - r_1^2 + a_{x,1}^2 - a_{x,3}^2 + a_{y,1}^2 - a_{y,3}^2 \\ r_4^2 - r_1^2 + a_{x,1}^2 - a_{x,4}^2 + a_{y,1}^2 - a_{y,4}^2 \end{pmatrix} \qquad (2)$$

Beim EKF wirken nichtlineare Funktionen auf das System ein, durch die der interne Systemzustand $\boldsymbol{x}_k$ und die Beobachtungen am System $\boldsymbol{y}_k$ beschreiben werden:

$$\begin{aligned} \tilde{\boldsymbol{x}}_{k+1} &= \boldsymbol{f}\left(\hat{\boldsymbol{x}}_k, \boldsymbol{u}_k, \boldsymbol{w}_k\right) \\ \tilde{\boldsymbol{y}}_k &= \boldsymbol{h}\left(\tilde{\boldsymbol{x}}_k, \boldsymbol{v}_k\right) \end{aligned} \tag{3}$$

Der Zustandsvektor $\boldsymbol{x}_k$ repräsentiert die zu schätzende Position des Tags und der Ausgangsvektor $\boldsymbol{y}_k$ die gemessenen Abstände zu den vier Ankern. Für die Schätzung einer Bewegung erfordert der interne Systemzustand $\boldsymbol{x}_k$ neben den Positionen p_x und p_y zusätzlich die jeweilige Geschwindigkeit v und Beschleunigung a:

$$\begin{aligned} \boldsymbol{x}_k &= \begin{pmatrix} p_x & v_x & a_x & p_y & v_y & a_y \end{pmatrix}^{\mathrm{T}} \\ \boldsymbol{y}_k &= \begin{pmatrix} r_1 & r_2 & r_3 & r_4 \end{pmatrix}^{\mathrm{T}} \end{aligned} \tag{4}$$

Die Störgrößen w_k und v_k repräsentieren das Prozess- und Messrauschen mit gegebenen Kovarianzmatrizen $\boldsymbol{Q}_k$ und $\boldsymbol{R}_k$. Die Messfunktion $\boldsymbol{h}$ fungiert als Bindeglied zwischen $\boldsymbol{x}_k$ und $\boldsymbol{y}_k$. Die Prozessfunktion $\boldsymbol{f}$ setzt den vorherigen Systemzustand zum Zeitpunkt k mit dem zum Zeitpunkt $k+1$ in Beziehung. Da die Berechnung der Geschwindigkeit und der Beschleunigung lineare Gleichungen darstellen, existiert keine Prozessfunktion $\boldsymbol{f}$ und das Bewegungsmodell wird über die Transitionsmatrix $\boldsymbol{A}_k$ dargestellt:

$$\boldsymbol{A}_k = \begin{pmatrix} 1 & \Delta t & \Delta t^2/2 & 0 & 0 & 0 \\ 0 & 1 & \Delta t & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \Delta t & \Delta t^2/2 \\ 0 & 0 & 0 & 0 & 1 & \Delta t \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{5}$$

Zur Bestimmung der ungefähren Startposition $\mathrm{P}_0(x_0, y_0)$ wird der anfängliche Systemzustand $\hat{\boldsymbol{x}}_0$ über die Gleichung (2) ermittelt. Für die anschließende Schätzung der Tagpositionen, werden die Funktionswerte der nichtlinearen Funktion $\boldsymbol{h}$ an die realen Tagpositionen $\mathrm{P}(x_k, y_k)$ approximiert. Die Funktion $\boldsymbol{h}$ beinhaltet dabei die Trilaterationsgleichungen aus (1) und berechnet den approximierten Systemausgang $\tilde{\boldsymbol{y}}_k$, mit dem die aktuelle Schätzung $\tilde{\boldsymbol{x}}_k$ korrigiert wird. Die zugehörige Jacobi-Matrix $\boldsymbol{C}_k$ beschreibt die partiellen Ableitungen der Funktion $\tilde{\boldsymbol{y}}_k = \boldsymbol{h}(\tilde{\boldsymbol{x}}_k)$ nach p_x und p_y. Da nur die Positionsberechnung über Differentialgleichungen beschrieben wird, sind die Matrixelemente innerhalb der Matrix $\boldsymbol{C}_k$ für die Geschwindigkeit und die Beschleunigung Null:

$$\boldsymbol{C}_k = \begin{pmatrix} \partial r_1/\partial p_x & 0 & 0 & \partial r_1/\partial p_y & 0 & 0 \\ \partial r_2/\partial p_x & 0 & 0 & \partial r_2/\partial p_y & 0 & 0 \\ \partial r_3/\partial p_x & 0 & 0 & \partial r_3/\partial p_y & 0 & 0 \\ \partial r_4/\partial p_x & 0 & 0 & \partial r_4/\partial p_y & 0 & 0 \end{pmatrix} \quad \text{mit} \quad \begin{aligned} \frac{\partial r_i}{\partial p_x} &= \frac{p_x - a_{x,i}}{\sqrt{(p_x - a_{x,1})^2 + (p_y - a_{y,1})^2}} \\ \frac{\partial r_i}{\partial p_y} &= \frac{p_y - a_{y,i}}{\sqrt{(p_x - a_{x,1})^2 + (p_y - a_{y,1})^2}} \end{aligned} \tag{6}$$

Mit den gemessenen Ankerabständen als Eingabeparameter in das Filter können die sich kontinuierlich ändernden Tagpositionen erfasst werden und in Verbindung mit den Geschwindigkeitsangaben als Trajektorie dargestellt werden. Der gesamte EKF-Algorithmus ist in der Abbildung 3 dargestellt. Für jede Position wird im zuerst ein Systemzustand vorhergesagt, der anschließend anhand der bekannten Fehlerkovarianzen sowie der Messwerte korrigiert wird.

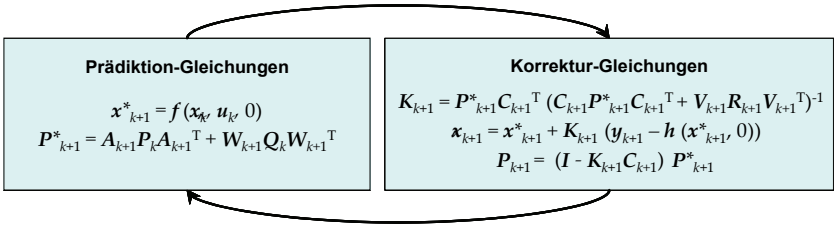


Abbildung 3: EKF-Gleichungen

4. Tracking von Fahrerlosen Transportfahrzeugen

Für die Positionserfassung wird das nanoLOC Sensornetzwerk entlang der Fahrbahn des FTFs installiert. Dabei ist zu beachten, dass die Fahrspur komplett von den Ankern eingekreist ist. Um eine robuste Kommunikation zu ermöglichen, besteht Sichtkontakt zwischen allen Hardwarekomponenten (Line-of-sight) und die Antennen sind in dieselbe Richtung ausgerichtet. Anhand von zwei Messszenarien wird untersucht, wie präzise das nanoLOC System die sich kontinuierlich ändernden Positionen erfasst. Das FTF wird gestartet und beschleunigt mit $a = 260\ \mathrm{mm/s^2}$ bis es seine Maximalgeschwindigkeit von $v = 1130\ \mathrm{mm/s}$ erreicht hat.

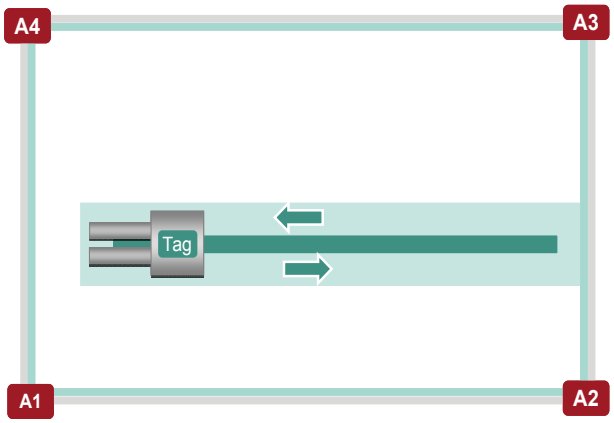


Abbildung 4: Versuchsumgebung gradlinige Strecke

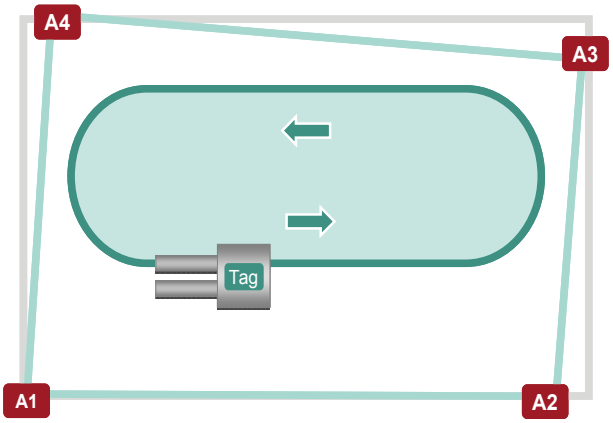


Abbildung 5: Versuchsumgebung Kreisbahn

In dem ersten Versuchsszenarium bewegt sich das Fahrzeug auf einer geraden Linie, wobei es die Hinfahrt im Vorwärtsgang und die Rückfahrt im Rückwärtsgang absolviert. In dem zweiten Versuchszenarium fährt das FTF ein U auf einer ovalen Kreisbahn ab. Die Abbildungen 4 und 5 stellen die beiden Versuchsumgebungen grafisch dar.

4.1. Experimentelle Ergebnisse

Die Liniendiagramme in der Abbildung 6 visualisieren die Positionen des FTFs über die zurückgelegte gradlinige Strecke und die Diagramme in der Abbildung 7 stellen den Verlauf der U-förmig gefahrenen Strecke auf der Kreisbahn dar.

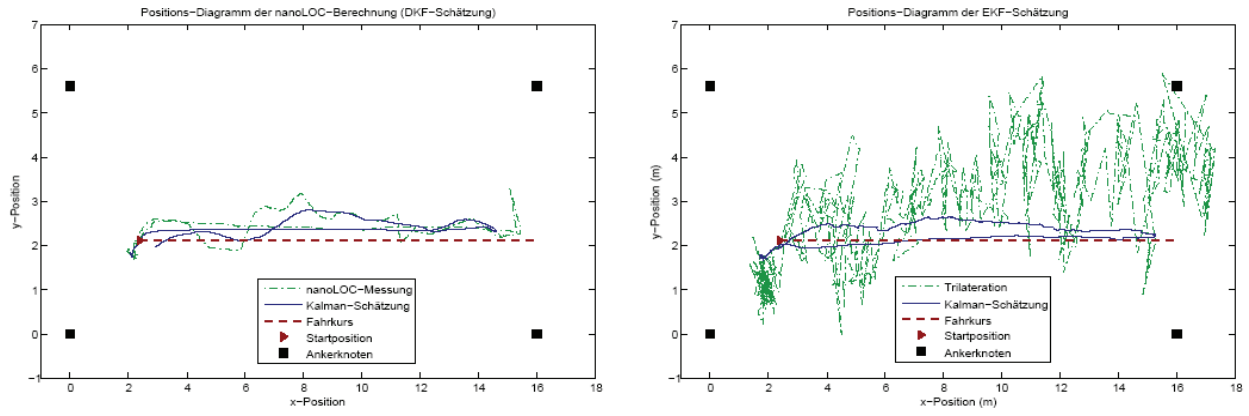


Abbildung 6: Positionsbestimmung gradlinige Strecke (nanoLOC l. / EKF r.)

Das jeweils linke Diagramm stellt die DKF-Schätzung der nanoLOC-Berechnung dar und das rechte Diagramm die EKF-Schätzung mit der Trilaterationsberechnung. Es ist zu erkennen, dass die Bewegung sowohl durch den nanoLOC-Algorithmus als auch durch die Trilateration erfasst wird. Jedoch sind die Ergebnisse stark verrauscht, sodass es sinnvoll ist, diese über das Kaman-Filter zu glätten und dadurch den realen Werten anzunähern. Das Kalman-Filter reduziert das Rauschen deutlich. Dabei ist zu erkennen, dass die EKF-Schätzung die gefahrene Strecke genauer wiedergibt als die DKF-gefilterte nanoLOC-Berechnung.

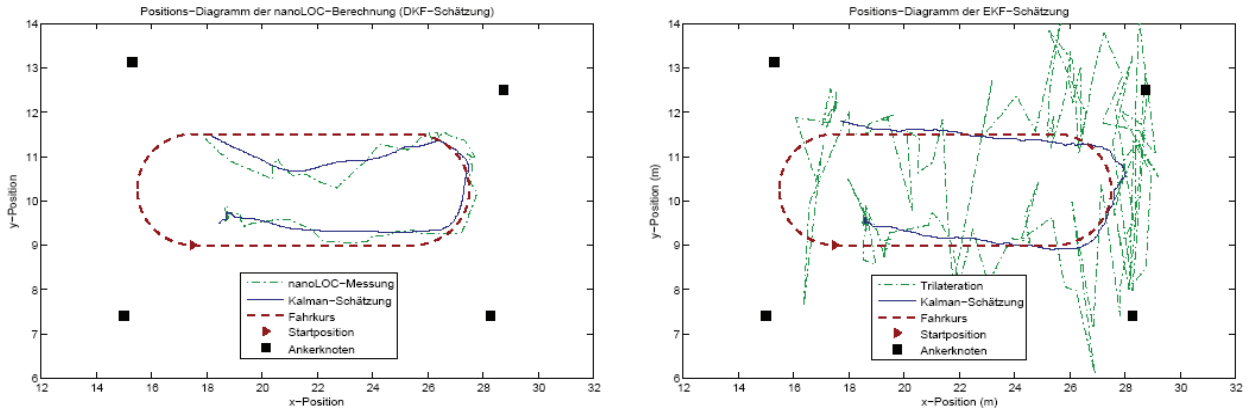


Abbildung 7: Positionsbestimmung Bahnkurve (nanoLOC l. / DKF r.)

5. Fazit und Ausblick

Abschließend ist festzuhalten, dass das nanoLOC System geeignet für die Aufzeichnung eines Bewegungsmodells ist. Eine alternative Berechnungsmöglichkeit zum nanoLOC-Algorithmus stellt das Erweiterte Kalman-Filter auf Basis der Trilaterationsberechnung dar. Im Durchschnitt wird eine Positionsgenauigkeit von ungefähr $0{,}5\ \mathrm{m}$ erzielt, was aufgrund der Fahrzeuggröße ausreichend genau ist. Durch Parameter-Tuning der Prozessrausch-Kovarianz Q und der Messrausch-Kovarianz R kann das Schätzergebnis zusätzlich noch optimiert werden. Auch bei der Positions- und Geschwindigkeitsbestimmung durch das Erweiterte Kalman-Filter kann über die Anpassung von Q und R das EKF-Schätzergebnis noch mehr an die realen Werte angenähert werden.

Bei den Algorithmen zur Positionsbestimmung handelt es sich bislang um ein Test- und Analysewerkzeug. Bei einem Bewegungsmodell soll mit dem Kalman-Filter neben den sich kontinuierlich ändernden Positionen auch die zugehörige Geschwindigkeit und Beschleunigung geschätzt werden, was bislang noch nicht umgesetzt und untersucht wurde. Zudem ist eine Erweiterung auf 3D-Anwendungen für eine flexible Lokalisierung von Transportfahrzeugen erstrebenswert. Das nanoLOC System stellt die Voraussetzungen für 3D-RTLS bereit und der EKF-Algorithmus ist ebenfalls problemlos auf eine weitere Raumdimension erweiterbar.

Quellenverzeichnis

[1] Schier, Arkadius: Drahtlose Sensornetzwerke in der Logistik - Einsatz drahtloser Sensornetzwerke zur dezentralen Lagerhaltung; 1. Auflage, Verlag Dr. Müller, 2008

[2] Röhrig, Christof and Spieker, Sarah: Localization of Pallets in Warehouses using Wireless Sensor Networks; In: Proceedings of the 16th Mediterranean Conference on Control and Automation; Corsica, France; June 2008

[3] NANOTRON Technologies GmbH (Hrsg.): nanoLOC System and TRX Transceiver; http://www.nanotron.com/EN/PR_nanoLOC_Transceiver.php; Abruf: 16. September 2008

[4] NANOTRON Technologies GmbH (Hrsg.): SDS-TWR (Symmetrical Double-Sided Two Way Ranging); http://www.nanotron.com/EN/CO_techn-sdstwr.php; Abruf: 16. September 2008

[5] Spieker, Sarah: Lokalisation in der Lagerverwaltung – Nutzenpotentiale drahtloser Sensornetzwerke zur Positionsbestimmung sowie Genauigkeitsoptimierung mit dem Kalman-Filter; 1. Auflage, Verlag Dr. Müller, 2008

OGAS - Open Geographic Applications Standard: An Open User-centric Description Language for Exchangeable Location-based Services

Johannes Martens

johannes.martens@ifi.lmu.de

Ulrich Bareth

ulrich.bareth@ifi.lmu.de

Georg Treu

georg.treu@ifi.lmu.de

Ludwig-Maximilians-Universität München
Institut für Informatik
Lehrstuhl für Mobile und Verteilte Systeme
Oettingenstr. 67, 80538 München

Abstract

Location-based Services are gaining more and more users and are available on a broad set of mobile devices. Due to their complexity such services are currently only provided by large organizations that use their own proprietary specifications thus leaving the full potential untapped. As seen during the development of the World Wide Web a tremendous growth can be achieved by enabling end-users to create services on their own.
In this paper we present a platform independent description language comprising a tool box of standard functions allowing end-users to easily design and exchange innovative location-based services. Besides its simplicity it also enables larger organizations to describe more sophisticated location-based services and run them on different platforms. The language supports event based triggers, like entering a zone, as well as temporal constraints and flow control. The created code is human readable thus its executed actions are completely transparent to the user. It is interpreted by a generic parser on the mobile device managing the positioning process and performing actions like notifying the user or calling a 3rd party application.

1. TODAY'S LBS

Location-Based Services (LBS) have not been as successful as estimated in the past yet. Besides other factors, implementing an LBS requires a lot of know-how, effort and probably even an own server infrastructure, thus in the end a lot of time and money. Alternatively one of the big LBS-Providers could be used which is gainless for small projects or simple services like a personal tour recommendation. A lot of tools have been developed to ease the implementation of LBS by taking care of the positioning process or providing generic functionalities, like geocoding, etc. Nevertheless, the mobile application itself still has to be implemented and ported to different handsets.

To overcome these obstacles one could think about a generic way to create a LBS by simply describing its behavior like writing a web site in HTML. In our opinion, the simplicity of HTML is one of the main reasons for the success of the World Wide Web. Just like for building a new web site by writing a document in HTML, uploading it onto a web server and displaying it in a web browser, creating an LBS means writing a short service description in an XML dialect, publishing it on a web server for being accessible by anyone and interpreting it by a client application on the mobile device of the user who subscribed for that service.

Therefore we propose an XML-based language similar to HTML, to provide an easy way for users and service-providers to implement LBSs. This paper presents first results of the current work in progress towards an open user-centric description language for exchangeable location-based services.

In the next section we present the state of the art of existing LBS solutions and potential scenarios for our approach. The concept behind OGAS is described in section 3. A proof of concept via a prototype implementation is given in section 4. Afterwards we summarize the results and describe future work.

2. OUR VISION OF LBS

In this chapter we present existing solutions in the area of LBS. Later on, we give a detailed description of the problem domain and picture our vision of location-based services in the future by introducing two scenarios. Finally we discuss the existing solutions to retrieve requirements for our proposed concept which will be described in the following chapter.

1.1. Related work

Nowadays Location-Based Services are realized by several LBS-Providers who mostly implement their own proprietary protocols and applications. We identified four different fields of LBS solutions, namely "Enterprise LBS solutions", "Lightweight web-based LBS", "self contained commercial LBS" and finally "Scientific platforms and Standards". The main focus of our investigations will be on offered LBS functions like reactive, proactive or community features, the used protocols, interfaces and data formats and advanced concepts like temporal constraints for LBS. Another critical aspect is ease of implementation.

Enterprise LBS solutions

If a company offers a product for location-based services to third parties, we speak of an "Enterprise LBS solution". Usually those companies sell their products or customized solutions like location-aware logistics-applications to other companies.

A typical example is **DeCarta's DDS** (Drill Down Server)[1]. The DDS provides several services like geocoding, routing or proximity lookup of nearby fixed objects that can be accessed by a proprietary xml-based query language, the so called "DDS Query Language" or directly by accessing the "DDS Web Services". Responses from those queries will be returned as proprietary RTF (Rich Text Format). Unfortunately DeCarta only provides reactive and no proactive or community services and also doesn't allow constraints for LBS.

Autodesk's platform **LocationLogic**[2] provides location-based services like geocoding, routing, search for POIs and location-based alarms. Its services are compliant to standards like OpenLS[3] and Autodesk is also member of the WAP-Forum, the OMA (formerly known as LIF) and the OGC. Though LocationLogic tries to support rapid application development by the concept of application-based building blocks that consist of prefabricated code-blocks, a service provider still needs quite some know-how in the area of Java and XML. Furthermore its services lack proactive LBSs and constraints.

The AS (Application Server) of Oracle's 10g databases[4] and **Oracle Application Server** Wireless (Oracle 9iASW) also provide the development of mobile LBS. The applications will be written in Java and Oracle's MobileXML dialect. Oracle's ASW offers reactive services as well as Geofencing and will implement proactive Push-Services, location-based and distance-based alarms and community features similar to the buddy-tracker application in the future.

Lightweight web-based LBS

A relatively new solution in the field of location-based services is **Yahoo's Fire Eagle**[5], a web-based service platform where users can specify access policies for their location data in a profile for every single third party application that wants to use their data. Thus Fire Eagle is more a position broker that responds, based on specific policies, to other applications that request the users location information. Unfortunately it is not able to recognize spatial events like proximity detection or Geofencing services which doesn't make it a real LBS-provider.

Self-contained commercial LBS

This field introduces LBS providers who do not make use of existing platforms but implement most of the LBS functions by themselves to build their own proprietary LBS application.

WHERE, a product of **uLocate**[6] is an "LBS widget platform", as they call it, which enables developers to create their own WHERE-based LBSs. Though their API was "written for quick development" it still requires a lot of geo and programming experience. Furthermore it only includes some reactive local search LBSs but no proactive LBSs or temporal constraints.

The company **Qiro**[7] spun off in 2006 from Deutsche Telekom Laboratories and offers mobile location-based community- and information-services. It has been developed for a lot of mobile devices mainly for series 60 and windows mobile platforms. The application only realizes reactive LBSs and does not provide any interfaces for 3rd party developers.

Due to a massive marketing campaign, **Nokia Maps**[8] (formerly known as smart2go) might be one of the best-known LBS solutions today. The platform allows for mapping, routing, POI-search and supports turn-by-turn navigation. Also users are allowed and encouraged to upload their own user-generated content onto the platform. However the platform doesn't have an interface for 3rd party developers, which also makes it a close community for owners of Nokia devices.

Scientific platforms and Standards

Usually developed in the environment of a university, the scientific platforms aren't primarily interested in commercial success of their solutions like the following examples show.

Nexus[9], a research project of the Universität Stuttgart in Germany describes a virtual contextual representation of the real world, the augmented world model, with different context domains like spatial or other context but solely enables developers to retrieve location information which is insufficient for creating LBSs in our approach.

The Ludwig-Maximilians-Universität München in Munich, Germany developed **TraX**[10] (track and exchange), a location provider, that effectively notifies users about spatial events they have subscribed for like geofencing or proximity detection of dynamically moving objects. Furthermore it also offers a plethora of basic services like periodic, distance-based or hybrid position update events. Thus it can be seen as an ideal toolkit providing an LBS-enabling platform.

The research projects **L-ToPSS**[11] of the University of Toronto in Canada, **myMytileneCity Guide**[12] of the University of Aegaen in Greece and **Flame 2008**[13] developed at Fraunhofer Institut Software- und Systemtechnik (ISST) and the Institute of Computing Technology of the Chinesische Akademie der Wissenschaften won't be discussed in detail here.

1.2. Discussion

Most of the aforementioned platforms have no open interfaces for 3rd party developers to create their own LBSs and if they do, the standard like OpenLS usually is not up to date anymore because it only provides reactive LBSs which makes the offered interfaces or protocols useless for the creation of individual and modern LBSs like proactive notifications on entering a geofence or on approaching a friend in a mobile community. Also none of the platforms is able to define constraints that can inhibit service execution if they aren't fulfilled.
Due to high complexity, limited resources and a long time-to-market, using "Enterprise LBS solutions" for creating new LBSs is not feasible for small companies or individuals. The "Lightweight web-based LBS" providers like Fire Eagle only support storing and retrieving location information but can neither give any reactive nor proactive LBS features thus being useless for the creation of individual LBSs. "Self-contained commercial LBS"

solutions are profit oriented companies that exclude third parties from their paying customers. The “Scientific platforms and standards” also don’t provide reactive or proactive LBS, but TraX implements mechanisms for notifying registered users about simple and complex spatial events which can be seen as basis for LBS providers implementing proactive services.

1.3. Scenarios

In the following we introduce the two scenarios which show an easy way to create and use Geofencing for location-based advertising and proactive detection of nearby friends for a Buddy Alert service.

Location-based Advertising

In the first scenario, the owner of a small coffee company wants to use new marketing strategies by offering location-based advertising on the mobile devices of his potential customers passing by her shop to increase turnover during lunch time. She wants to offer discount on a lunch menu by sending electronic vouchers to her customers. She simply writes a little XML document that specifies her LBS advertising service and places it on the website of her company as download. Now her customers can easily download the service and execute them on their LBS-enabled devices which, from now on, show an electronic voucher when they enter the vicinity of the coffee shop during lunchtime and the service is activated. That can be seen as a benefit for both involved parties.

Buddy Alert

Next generation mobile LBSs will also be able to detect if two or more moving objects approach each other up to a predefined distance. In this scenario, a student who wants to learn for her exam is looking for fellow students who want to group up with her and learn together for their exams. So she specifies a tiny LBS in a simple XML description language which notifies her if one of her fellow students approaches within 100 meters, so she can ask her directly if they want to learn together for minimizing effort and maximizing fun. After she met three buddies she can learn with, she deactivates the service and is glad that she took the 2 minutes that it took to specify it.

3. OGAS CONCEPT

This paper proposes OGAS; a user-centric XML-based description language for location-based services. Our main focus lies on the simplicity of the description language to also enable small companies and individuals that aren’t very familiar with programming languages to easily create services on their own. That’s why we chose XML as basis for our language because it can be created with a simple text editor and is human-readable which also also increases transparency for users compared to binary formats. Furthermore XML dialects are generic and extenable. Another focus of our language is that despite its simplicity it is powerful enough to create sophisticated business applications. For that reason we divide the components of our language into a basic serviceset for standard users and an extended serviceset for professional use. Documents described by our proposed language must be interchangeable directly between users and executable on every OGAS-enabled mobile device.
A so called toolbox of LBS features (consisting of the basic serviceset and extended serviceset) can be mapped to basic LBS modules of the TraX platform which is used by our description language as location provider. The TraX platform generates spatial events that will trigger actions predefined by the service on the mobile device if certain constraints are met.

A standard service description consists of a triplet containing **actions**, **triggers** and **constraints**. Actions can be the popup of a browser displaying a predefined website, vibrating or ringing of the mobile device or starting another application. Triggers are spatial events generated by the TraX platform like entering a geofence or approaching another person. Limitations like opening hours of a store would be described as temporal constraints but also other constraints can be conceivable. The example below illustrates the empty body of an OGAS document:

```
<?xml version=“1.0“ encoding=“UTF-8“ ?>
<OGAS>
  <header author=„…“ name=„…“><description> ...</description></header>
<service action=„…“ trigger=„…“ constraints=„…“></service>
  <action name=„…“ type=„…“><parameters>…</parameters></action>
  <trigger name=„…“>…</trigger>
  <constraints name=„…“ type=„…“><parameters>…</parameters></constraints>
</OGAS>
```

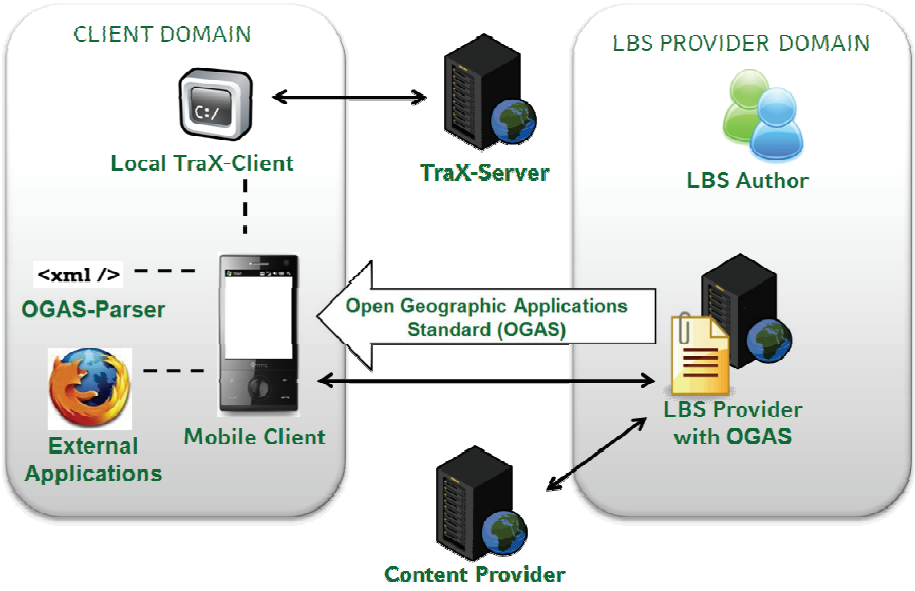


Figure 1: OGAS Architecture

As shown in Figure 1, our architecture consists of the following components: In the LBS provider domain an LBS provider making the OGAS document available for downloading, an optional content provider for external contents. In the client domain there is the OGAS enabled mobile client, which communicates with the OGAS-Parser, external applications and the local TraX client that sends monitoring jobs to the TraX server and receives spatial events.
Services have to be created, published, subscribed (or downloaded), activated and then executed on the mobile device which means that the OGAS application on the mobile device parses the service description document and translates it into TraX jobs that monitor spatial events and triggers services on the mobile device. If a service is triggered, corresponding constraints will be checked and the according actions then being executed.

A typical process from service creation until execution will contain the following steps:

Creation: The author uses our XML dialect to describe his service
Provisioning: The service can be uploaded on a conventional webserver for download by other users
Subscription: A user loads the service into the lbs repository of his OGAS-enabled device
Activation: Services have to be activated explicitly and can be deactivated anytime by user interaction
Execution: The OGAS document will be parsed on the device and translated into TraX jobs that throw spatial events or triggers. If a service is triggered and the constraints are met, the corresponding action, e.g. calling an external application, will be performed.

4. PROTOTYPE REALIZATION

Based on the aforementioned concept we developed a prototype implementation of a LBS providing location and time based reminders for appointments. The service consists of a mobile client application presenting events and notification to the user on her mobile phone as well as a back-end system providing a web frontend for entering location trigger and time constraints.

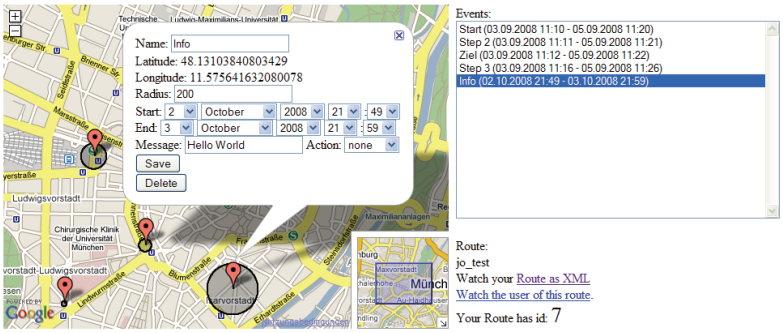


Figure 2: Web interface of prototype

The mobile client is a J2ME-MIDlet including a data aggregation, a location and a presentation module. The data aggregation module downloads the selected xml file from the server, reads the triggers and constraints

out and sends the location triggers to the location module. This module is provided by the TraX framework acting as the underlying location platform. After retrieving spatial or temporal events the presentation module notifies the user via vibration, alarm sound or a pop-up about the specific event.
The back-end system was developed on top of Ruby on Rails. The user can configure triggers on a Google Maps enhanced webpage (see picture 1), which are then saved as a “service” in the proposed XML format.
The created services can be shared among other users and downloaded by the mobile client. In the following the listing of an example XML file containing one route is shown:

```
<?xml version=“1.0“ encoding=“UTF-8“ ?>
<OGAS>
  <header user_hash="963" name=„ Start"><description></description></header>
<service created-at="Wed Sep 03 11:10:24 +0000 2008" updated-at="Wed Sep 05 11:20:20 +0000 2008" action=„notify_vibrate“ trig-
ger=„test_geofence“ constraints=„test_constraints“></service>
  <action name=„notify_vibrate“ type=„vibrate“><message>Starting the tour</message></action>
  <trigger name=„test_geofence“ type=”geofence”>
    <latitude>48.1300072512693</latitude>
    <longitude>11.557788848876953</longitude>
    <radius>20</radius>
  </trigger>
  <constraints name=„test_constraints“ type=„temporal“>
    <start-time timestamp="1220440200000">Wed Sep 03 11:10:00 +0000 2008</start-time>
    <end-time timestamp="1220613600000">Fri Sep 05 11:20:00 +0000 2008</end-time>
  </constraints>
</OGAS>
```

The document consists of several services, which are described by an action to be executed and a message shown to the user, a trigger containing location events and temporal constraints.
This prototype implementation serves as proof of concept for the described approach towards an open and generic description language for LBS.

5. CONCLUSION AND FUTURE WORK

This paper presented a new approach towards an open user-centric description language for exchangeable location-based services. We showed, that existing LBS solutions do not target user created LBS. Therefore we introduced the Open Geographic Applications Standard and provided a proof of concept via a prototype implementation.
Our future work will focus on the definition of a flow control for actions and triggers and professional functions in an extended service set enabling more complex LBS.

Bibliography

[1] DeCarta, DDS (Drill Down Server), http://www.decarta.com/products/dds/index.html, accessed on oct 8, 2008.
[2] Autodesk, Location Logic, 2008, http://www.usa.autodesk.com, accessed on oct 8, 2008.
[3] Open Geospatial Consortium. OpenGIS Location Services (OpenLS): Core Services, Version 1.1, Mai 2005, http://portal.opengeospatial.org/files/?artifact id=8836, accessed on oct 8, 2008.
[4] Oracle, http://www.oracle.com, accessed on oct 8, 2008.
[5] Yahoo, Fire Eagle, http://fireeagle.yahoo.net/, accessed on oct 8, 2008.
[6] uLocate, WHERE, http://www.ulocate.com/where.php, accessed on oct 8, 2008.
[7] Qiro, http://www.myqiro.de/, accessed on oct 8, 2008.
[8] Nokia Maps, http://europe.nokia.com/maps, accessed on oct 8, 2008.
[9] D. Nicklas, Mitschang, B., Institute of Parallel and Distributed Systems (IPVS), Universität Stuttgart, On building location aware applications using an open platform based on the NEXUS Augmented World Model, Software and Systems Modeling, Volume 3, Nr. 4, pages 303-313, dec 2004.
[10] A. Küpper, Treu, G., Linnhoff-Popien, C., Ludwig-Maximilians-Universität (LMU) München, Trax: A device-centric middleware framework for location-based services, IEEE Communications Magazine, Advances in Service Platform Technologies for Next Generation Mobile Systems, sep 2006.
[11] I. Burcea, Jacobsen, H.-A., Department of Electrical Engineering and Department of Computer Science, University of Toronto, L-ToPSS - Push-Oriented Location-Based Services, Lecture Notes in Computer Science, Volume 2819, pages 131-142, 2003.
[12] M. Kenteris, Gavalas, D., Economou, D., Cultural Heritage Management Laboratory, Department of Cultural Technology and Communication, University of the Aegean, Lesvos, Greece, An innovative mobile electronic tourist guide application, Personal and Ubiquitous Computing, feb 2007.
[13] P. Gupta, Februar Hoffmann, M., Holtkamp, B., Möhr,W., Peters, J., Ritscher, M., Voisard, A., Fraunhofer Institute IGD, IPSI, ISST, SIT, Mobile kontextabhängige Multimediadienste, Informatik-Spektrum, Volume 27, Nr. 1, pages 35-43, 2004.

Ortsbezogene Verwaltung von Informationen für Fahrzeug-zu-Fahrzeug-Anwendungen

Vivian Prinz

Technische Universität München
- Institut für Informatik

Boltzmannstr. 3, 85748 Garching b. München

Tel: +49 89 289-18650
Fax: +49 89 289-18657
E-Mail: prinzv@in.tum.de

Wolfgang Wörndl

Technische Universität München
- Institut für Informatik

Boltzmannstr. 3, 85748 Garching b. München

Tel: +49 89 289-18686
Fax: +49 89 289-18657
E-Mail: woerndl@in.tum.de

Abstract

Aktive Sicherheitsanwendungen wie Frühwarnsysteme werden als wichtigstes künftiges Einsatzgebiet der Fahrzeug-zu-Fahrzeug Kommunikation gesehen. Beispielsweise können sich Fahrzeuge frühzeitig vor Unfällen oder Glatteis warnen. Informationen wie Glatteiswarnungen müssen hierfür durch das Fahrzeugnetz verwaltet werden. Sie müssen ortsbezogen publizierbar und anschließend für die Dauer ihrer Gültigkeit in der Region verfügbar sein. In diesem Artikel stellen wir einen Lösungsansatz für die ortsbezogene verteilte Verwaltung von Informationen vor, um deren regionale Verfügbarkeit zu gewährleisten. Dabei kommt eine strukturierter Peer-to-Peer (P2P)-Algorithmus zum Einsatz. P2P-Algorithmen wurden bisher vorwiegend Internetanwendungen zugrunde gelegt. Um netzspezifische Unterschiede zu adaptieren, wird das Verkehrsnetz in Segmente gegliedert, die separate, interagierende P2P-Netze bilden. Informationen werden von den Fahrzeugen der Segmente verteilt verwaltet und gegebenenfalls unter benachbarten Segmenten weitergereicht. Anwendungen können durch die vorgeschlagene Lösung Informationen ohne Kenntnis von der Fahrzeugnetzsegmentierung ortsbezogen und gültigkeitsbeschränkt publizieren. Anschließend sind die Informationen abfrag-, modifizier- und löschbar.

1. Einleitung

Die Kommunikation zwischen Fahrzeugen, bei der andere Kommunikationsendpunkte nicht ausgeschlossen werden, wird vereinheitlichend Car-to-X (C2X)-Kommunikation genannt. C2X-Kommunikation bildet zum einen die Grundlage für Anwendungen aus dem Bereich der *aktiven Sicherheit*. Dabei kommunizieren Fahrzeuge miteinander, um Situationen zu erkennen, die nicht von lokalen Sensoren, dem Fahrer selbst oder einem einzelnen Auto erfassbar sind. Beispielsweise können sich Fahrzeuge frühzeitig vor Unfällen, Glatteis oder Stauenden warnen. Folglich wird C2X-Kommunikation als eine wichtige künftige Möglichkeit angesehen, Verkehrsunfälle und deren Folgen zu reduzieren. Einen zusätzlichen Mehrwert für Autofahrer sollen komfortable Zusatzanwendungen bieten, die so genannten *Deployment-Anwendungen*. Durch sie können Fahrern zum Beispiel aktuelle Informationen zu freien Parkplätzen oder zu Tankstellen und deren Benzinpreisen in der aktuellen Umgebung angezeigt werden. Die Informationen können von Parkhäusern und Tankstellen wie auch von anderen Fahrzeugen in das *Vehicular Ad-Hoc Network (VANET)* eingespeist werden.

Zur Realisierung derartiger Anwendungen müssen fortgeschrittene Verfahren zur ortsbezogenen Verwaltung von Informationen bereitgestellt werden. Dieser Artikel thematisiert diese Verfahren und stellt einen eigenen Ansatz auf Basis von Peer-to-Peer-Mechanismen vor. Dabei wird in Abschnitt 2 zunächst genauer auf Anforderungen an C2X-Kommunikationssysteme eingegangen. Abschnitt 3 gibt einen Überblick über den aktuellen Stand in verwandten Forschungsfeldern und beschreibt bestehende Herausforderungen. In Abschnitt 4 wird ein Lösungsansatz zur Realisierung dieser Herausforderungen erläutert. Abschnitt 5 gibt eine Zusammenfassung und einen kurzen Ausblick auf das weitere Vorgehen.

2. Anforderungen an eine Kommunikationsplattform für C2X-Anwendugen

In den beiden einführend genannten Bereichen existiert eine Vielzahl möglicher Anwendungen. Um diese Anwendungen breitgefächert zu unterstützen, gilt es, eine Kommunikationsplattform mit einer wohl definierten Schnittstelle bereitzustellen, auf die beliebige Anwendungen aufgesetzt werden können. Die Kommunikationsplattform muss dabei folgenden Anforderungen genügen:

- **Dauerhafte Verfügbarkeit:** Informationen müssen über einen bestimmten Zeitraum verfügbar sein. Es genügt nicht, eine Glatteiswarnung einmalig an alle folgenden Fahrzeuge zu versenden. Fahrzeuge, die zu einem späteren Zeitpunkt in die Gefahrenzone einfahren, müssen ebenfalls gewarnt werden.
- **Ressourceneffizienz:** Da über die Luftschnittstelle kommuniziert wird, führen viele Nachrichten zu hohen Kollisionswahrscheinlichkeiten und langen Zugriffszeiten. Dies sollte insbesondere in Hinblick auf aktive Sicherheitsanwendungen vermieden werden.
- **Abfragbarkeit:** Ressourceneffizienz bedingt, dass einzelne Informationen nicht einfach kontinuierlich versendet werden sollten, um sie dauerhaft bereitzustellen. Daten wie Parkplatzinformationen sind nur für wenige Fahrzeuge von Interesse. Sie sollten von diesen bei Bedarf abgefragt werden können.
- **Modifizierbarkeit:** Parkhäuser müssen bei Belegung oder Freiwerden eines Parkplatzes die Informationen zu ihren Kapazitäten ändern können.
- **Löschbarkeit/Verfallen:** Fahrer sollten nicht auf Glatteis hingewiesen werden, wenn sich die Straßenverhältnisse bereits wieder normalisiert haben.
- **Ortsbezogenheit:** Es muss möglich sein, Informationen ortsbezogen zu publizieren. Beispielsweise sollte eine Warnung über schlechte Straßenverhältnisse für alle Fahrzeuge in der betreffenden Straße zugänglich sein. In weit entfernten Regionen ist die Warnung hingegen irrelevant.
- **Dezentralität:** VANETs sind von Natur aus dezentrale, selbstorganisierende Netze. Entsprechend sollte ein C2X-Kommunikationssystem nicht auf zentrale Vermittler- beziehungsweise Kontrolleinheiten angewiesen sein. Zentrale Einheiten müssten im gesamten Verkehrsnetz verfügbar sein, was zu hohen Mobilfunkkosten führt. Zudem können sie aufgrund der Vielzahl von Fahrzeugen zu einem Engpass werden. Ein Ausfall führt zu einem Gesamtausfall des Systems.
- **Ausfallsicherheit:** Nachdem keine zentrale Infrastruktur zum Einsatz kommen soll, werden Informationen von den beteiligten Fahrzeugen selbst verwaltet. Informationen dürfen dabei nicht verloren gehen, wenn ein Fahrzeug unerwartet per Funk nicht mehr erreichbar ist, weil es beispielsweise einen Tunnel passiert.

Im Folgenden wird das verteilte Speichern von Informationen, durch das sie anderen Fahrzeugen zur Verfügung gestellt werden, Publizieren genannt. Ihre verteilte und ortsbezogene Verwaltung durch die beteiligten Fahrzeuge, die deren Abfragbarkeit, Modifizierbarkeit und Löschbarkeit gestattet, wird als ortsbezogenes Informationsmanagement bezeichnet.

3. Ortsbezogenes Informationsmanagement in VANETs

In diesem Abschnitt werden Arbeiten aus verwandten Forschungsfeldern beleuchtet. Es ist von Interesse, wie bestehende Ansätze vorgehen, um die geschilderten Anforderungen zu erfüllen.

3.1. C2X-Kommunikationssysteme

Im Bereich der C2X-Forschung gibt es viele Arbeiten zu Medienzugriff, Routing und Gruppenadressierung. Medienzugriffsverfahren steuern den Zugang zu Kanälen der Luftschnittstelle unter verschiedenen Fahrzeugen. Routing-Mechanismen setzen andere Fahrzeuge als Vermittler ein, um Fahrzeuge außerhalb der Funkreichweite eines Senders zu erreichen (Multi-hop-Routing). Schließlich gibt es Protokolle, über die ganze Fahrzeuggruppen adressierbar sind. Zum einen sind dies Protokolle, durch die klassische Multicast-Mechanismen in VANETs umsetzbar sind. Des Weiteren existieren Geocast-Protokolle, die das Versenden einer Nachricht an alle Fahrzeuge innerhalb einer geografischen Region erlauben. Abgesehen davon wird derzeit der *Dedicated Short Range Communication-Standard (DSRC)* als übergreifender Standard in Amerika, Europa und Japan entwickelt. Im Fokus steht dabei die Interoperabilität der Transponder unterschiedlicher Hersteller, die im Fahrzeug und am Straßenrand eingesetzt werden. Ein wichtiger Punkt ist, dass der DSRC ein Derivat der WLAN-Technologie IEEE 802.11a als Basisfunksystem festschreibt – den sich in der Entstehung befindenden Standard 802.11p. Die Randbedingungen für diesen Standard sehen unter anderem Fahrzeuggeschwindigkeiten von bis zu 200 km/h vor. Des Weiteren soll direkte Kommunikation bis zu einer Entfernung von einem Kilometer möglich sein (vgl. u.a. [MeTo 08]).

Die genannten Arbeiten ermöglichen, dass Nachrichten über die Luftschnittstelle an einzelne Fahrzeuge und Fahrzeuggruppen gesendet werden. Ortsbezogenes Informationsmanagement wird nicht behandelt. Diesbezüglich existieren einige Ansätze, die sich mit dem intelligenten Wiederholen von Broadcast-Nachrichten befassen, um dauerhafte Verfügbarkeit zu gewährleisten (vgl. u.a. [WER+ 03]). Fahrzeuge eines Gebietes wiederholen das Versenden von Informationen nach bestimmten Regeln, so dass auch neu einfahrende Fahrzeuge informiert werden. Problematisch ist, dass durch kontinuierliche Rebroadcasts eine Vielzahl redundanter Nachrichten entsteht. Ein alternativer Ansatz ist der so genannte *Stored Geocast*. Dabei werden Nachrichten nicht von mehreren Knoten wiederholt versendet, sondern ein Knoten nimmt Stored Geocast-Nachrichten entgegen und liefert sie aus. Beispielsweise wird in [ImNa 96] angenommen, dass das Ad-hoc-Netz in feste Zellen aufgeteilt ist, denen je ein so genannter *GeoNode* zugewiesen ist. Der GeoNode speichert die Nachrichten dauerhaft und liefert sie aus. In [MFE 03] wird eine infrastrukturunabhängige Lösung vorgeschlagen. Nach einem initialen Geocast wird ein verteilter Wahlalgorithmus gestartet, durch den der Knoten bestimmt wird, der die Geocast-Nachricht weiterhin verwaltet und ausliefert. Verlässt er die Region, stößt er eine Neuwahl an. Auch in Geocast-Ansätzen muss weiterhin jede Information kontinuierlich durch den zugehörigen Geocast Server versendet werden. Alternativ kann jedes Fahrzeug periodisch seinen Standpunkt proklamieren. Dadurch weiß ein Geocast Server, wenn ein Fahrzeug in seine Zielregion einfährt. Er kann die Nachricht bei Bedarf ausliefern. Kontinuierliches Versenden von Informationen ist jedoch in jedem der Fälle unumgänglich.

Es lässt sich festhalten, dass dauerhafte regionale Verfügbarkeit in bestehenden Ansätzen dadurch erreicht wird, dass Informationen in regelmäßigen Abständen an alle Fahrzeuge einer Region ausgeliefert werden. Es besteht das Problem der Ressourcenineffizienz. Zudem wurde ortsbezogenes Informationsmanagement im Sinne von Modifizierbarkeit, Löschbarkeit oder Abfragbarkeit bisher nicht thematisiert.

3.2. Peer-to-Peer-Algorithmen

P2P-Algorithmen adressieren unter anderem ausfallsicheres, verteiltes Informationsmanagement. Dabei wurden sie bisher vorwiegend Internetanwendungen zugrunde gelegt. Bekannte Anwendungsbereiche sind verteilte Dateisysteme oder Kommunikationsanwendungen wie Skype. Peers sind in der Regel PCs der Internetnutzer. Die Anwendungen setzen unterschiedliche P2P-Algorithmen ein, die das zugrunde liegende Netz verwalten. Es entstehen dezentrale, selbstorganisierende P2P-Netze.

Unterschieden werden strukturierte und unstrukturierte P2P-Netze. Ein unstrukturiertes P2P-Netz besteht aus Peers, die das Netz nach wenigen, losen Regeln betreten und verlassen. In strukturierten P2P-Netzen ist die Topologie hingegen streng kontrolliert. Sowohl Daten als auch Informationen über andere Peers werden auf bestimmen Peers abgelegt, um spätere Anfragen effizienter zu gestalten. Dabei kommen verteilte Hashtabel-

len (engl. *Distributed Hash Tables (DHTs)*) zum Einsatz. Eine DHT ist eine Datenstruktur, durch die der Speicherort einer gesuchten beziehungsweise zu speichernden Information ermittelt wird. Informationen haben eindeutige Bezeichner. Sie werden, wie Peer-Adressen, mittels Hash-Funktionen auf Identifikatoren abgebildet. Informationen sind stets auf demjenigen Peer zu speichern beziehungsweise zu suchen, dessen Identifikator dem der Information am ähnlichsten ist. Ist eine Gültigkeitsdauer angegeben, löscht dieser zuständige Peer eine Information nach Ablauf. Peers halten kleine Routingtabellen mit den Identifikatoren und Adressen Ihrer Nachbarpeers. Anfragen werden anhand der Routingtabellen an den Peer mit dem gesuchten Identifikator weitergeleitet. Die zugehörigen P2P-Algorithmen unterscheiden sich in ihren Datenobjektschemata, ihren Schlüsselräumen und ihren Routing-Strategien. Auch werden unterschiedliche Replikationsmechanismen zur Gewährleistung von Ausfallsicherheit eingesetzt.

VANETs und P2P-Netze weisen ähnliche Eigenschaften auf. In beiden Fällen handelt es sich um dezentrale, selbstorganisierende Netze. Ferner sind kommende und gehende Fahrzeuge beziehungsweise Peers a priori nicht bekannt – sie sind anonym und verhalten sich autonom. Auch sind sie potenziell sehr groß – Internet und Verkehrsnetz. Schließlich sind P2P-Netze und VANETs starken topologischen Schwankungen ausgesetzt. PCs werden nach belieben ein- und ausgeschaltet. Fahrzeuge werden angelassen und abgestellt oder passieren Tunnel. Entsprechend berücksichtigen P2P-Algorithmen Eigenschaften, die auch VANETs eigen sind.

Es existieren Arbeiten, die einen Zusammenhang zwischen mobilen Ad-hoc-Netzen (MANETs) und P2P herstellen (vgl. u.a. [MKR 04]). Die Ansätze befassen sich allerdings nicht mit Fahrzeugnetzen. VANETs sind große Netze, in denen beispielsweise hohe Geschwindigkeiten zu berücksichtigen sind oder ein Bezug zur Straßentopologie herstellbar sein muss. Im Bereich der C2X-Forschung betrachten das SmartWeb-Projekt [Wahl 04] sowie Huang et al. [HMM+ 02] P2P-Kommunikation. Sie verstehen unter P2P-Kommunikation allerdings lediglich eine direkte Ende-zu-Ende-Verbindung zwischen zwei Fahrzeugen. Im Zuge des SmartWeb-Projekts wurde eine direkte WLAN-basierte Verbindung zwischen zwei realen Fahrzeugen umgesetzt, über die starkes Abbremsen propagiert werden kann. Huang et al. nutzen die Ende-zu-Ende-Verbindung zweier aufeinander zufahrender Fahrzeuge, um Informationen zu Fahrtrichtung und Geschwindigkeit auszutauschen. Daraus wird berechnet, ob eine Kollision möglich ist.

Es lässt sich festhalten, dass P2P-Algorithmen eine Möglichkeit für verteiltes Informationsmanagement darstellen, die die C2X-Forschung bisher noch nicht in Betracht gezogen hat.

4. Einsatz von P2P-Algorithmen in C2X-Netzen

In diesem Abschnitt werden zunächst Herausforderungen besprochen, die sich in Zusammenhang mit dem Einsatz von P2P-Algorithmen in C2X-Netzen ergeben. Anschließend werden konzeptionelle Lösungsmöglichkeiten dargelegt. Abschließend erfolgt eine Diskussion der vorgestellten Lösung.

4.1. Herausforderungen

Bisher wurden P2P-Algorithmen vorwiegend Internetanwendungen zugrunde gelegt. Im Internet unterscheidet sich die Verbindungsqualität zwischen unterschiedlichen Rechnern geringfügig. In Verkehrsnetzen besteht insbesondere die Problematik, dass nicht jedes Fahrzeug mit jedem anderen Fahrzeug gleichermaßen verlässlich kommunizieren kann, da Hindernisse Funkverbindungen stören. Zum anderen können Fahrzeuge nicht gleichermaßen schnell kommunizieren, da Routing-Protokolle für VANETs aufgrund hoher Geschwindigkeiten viele Nachrichten zur Aktualisierung von Routingtabellen implizieren. Dies führt zu längeren Zugriffszeiten. Erfolgt der Verbindungsaufbau auf Anfrage, entstehen Verzögerungen. Infolge der hohen Knotengeschwindigkeiten sind Verbindungen über mehrere Fahrzeuge zudem nicht sehr stabil. Abgesehen davon ermöglichen existierende P2P-Algorithmen nicht, Informationen mit Bezug auf bestimmte Verkehrsnetzregionen zu verwalten.

4.2. Ortsbezogenes Informationsmanagement auf Basis überlappender P2P-Verkehrsnetzsegmente

Eine Lösungsmöglichkeit zur Adaption dieser unterschiedlichen Gegebenheiten besteht darin, das Verkehrsnetz nicht als ein großes P2P-Netz zu betrachten, sondern in viele kleinere Netze aufzuteilen. Es wird in Segmente gegliedert, die je ein strukturiertes P2P-Netz bilden. Die Segmentgrößen werden nach den technischen Möglichkeiten für den Kommunikationsradius der Fahrzeuge gewählt. Beispielsweise werden in urbanen Gegenden Straßenabschnitte von bis zu 1000 Metern selektiert, wenn man von dem maximalen DSRC-

Kommunikationsradius ausgeht (vgl. Abschnitt 3.1). Der Adressraum eines P2P-Netzes wird auf diese Segmente beschränkt, wobei sich benachbarte Segmente logisch an ihren Rändern überlappen (Abb. 1).

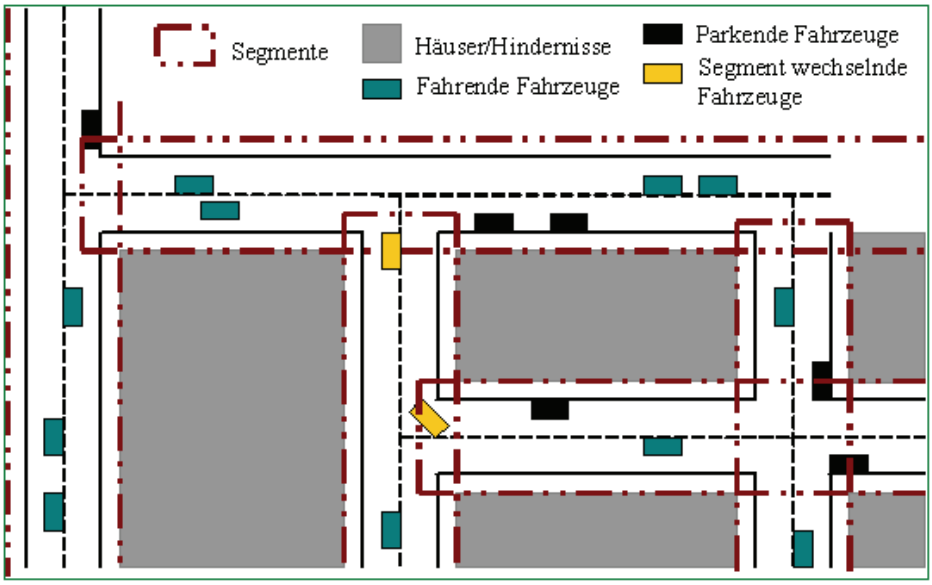


Abbildung 1: Überlappende P2P-Verkehrsnetzsegmente

Innerhalb eines solchen Segments können Fahrzeuge über DSRC direkt untereinander kommunizieren. Somit können sie in adäquater Geschwindigkeit alle Informationen austauschen, die zur Aufrechterhaltung der P2P-Topologie benötigt werden. Die Adressen der Fahrzeuge werden also, wie in Abschnitt 3.2 beschrieben, auf Identifikatoren abgebildet. Informationen werden stets auf demjenigen Fahrzeug abgelegt, das dem Identifikator der Information am nächsten ist. Somit ist zu jeder Zeit bekannt, wo eine Information abzulegen ist und wo sie zu suchen ist, wenn sie modifiziert, gelöscht oder abgefragt werden soll.

Fahrzeuge kennen ihre eigene Position, ihre Fahrtrichtung und ihre Geschwindigkeit. Ferner wissen sie durch digitales Kartenmaterial, wann sie Segmentgrenzen überschreiten. Die Informationen werden durch Navigationssysteme bereitgestellt. Nähert sich ein Fahrzeug einem neuen Segment, tritt es über Beacon-Nachrichten mit dem erstbesten Fahrzeug des zugehörigen P2P/C2X-Netzes in Verbindung. Es erhält so alle benötigten Informationen, um sich in die neue Topologie integrieren zu können. Ferner meldet es sich bei dem alten Netz ab, um Overhead durch Ausfall-Recovery zu vermeiden.

Nun kann es Informationen geben, die in einem Segment publiziert werden, aber auch über die Grenzen des Segments hinaus relevant sind. Um diese Informationen zwischen benachbarten P2P/C2X-Netzen weiterzuleiten, sucht ein Fahrzeug, dass eine Segmentgrenze überschreitet, zunächst entsprechende Informationen in seinem alten Segment. Anschließend publiziert es die Informationen, deren Reichweite für mehr als ein Segment angegeben wurde, in dem neuen Segment. Schließlich kennzeichnet es die Informationen durch Modifikation für dieses Segment als „weitergeleitet“. Redundantes Weiterleiten wird damit vermieden. Um segmentübergreifende Informationen spezifizieren zu können, wird ein entsprechender Informationstyp definiert. Somit sind sie für alle Fahrzeuge such- und modifizierbar. Eine Ausnahme bilden hier segmentübergreifende Warnmeldungen. Sind Warnmeldungen noch im benachbarten Segment von Bedeutung, werden sie gebroadcastet. Es kann nicht gewartet werden, bis das nächste Fahrzeug das Segment verlässt. Sind sie auch noch im übernächsten Segment gültig, müssen die Warnungen geflutet werden. Da dies für DSRC über einem Kilometer Reichweite entspricht, kann davon ausgegangen werden, dass das Szenario relativ selten eintritt.

Das beschriebene Vorgehen bezieht sich auf die Funktionalität einer C2X-Kommunikationsplattform. Die Fahrzeugnetzsegmentierung und alle übrigen Mechanismen werden auf dieser Ebene ein- und umgesetzt. Über die Schnittstelle der Schicht werden C2X-Anwendungen Methoden zum publizieren, auffinden, modifizieren und löschen bereitgestellt. Bei Publikation einer Information kann an der Schnittstelle ihre Reichweite, beispielsweise in Metern, angegeben werden. Die Plattform selektiert relevante Segmente gemäß dieser Angabe und publiziert die Information entsprechend. C2X-Anwendungen können so ohne Kenntnis von der Segmentierung des Fahrzeugnetzes entwickelt werden und Informationen ortsbezogen publizieren.

4.3. Diskussion des Lösungsansatzes

Von Vorteil ist bei diesem Lösungsansatz, dass alle Fahrzeuge direkt untereinander kommunizieren können. Es ist kein teures Multi-hop-Routing notwendig. Folglich können die Stabilisierungsroutinen unterschiedlicher P2P-Algorithmen effizient durchgeführt werden. Zudem ist der anwendungsspezifische Informationsaustausch unter den Fahrzeugen weniger teuer. Schließlich werden Informationen nur in relevanten Segmenten verwaltet.

Es besteht jedoch das generelle C2X-Problem, dass sich innerhalb des Kommunikationsradius, also innerhalb eines Segments, zu jeder Zeit mindestens ein Knoten befinden muss. Andernfalls kann kein Nachrichtenaustausch mehr stattfinden. Geht man davon aus, dass fest installierte Funkeinheiten von Parkhäusern und Tankstellen, mobile und gegebenenfalls auch parkende Fahrzeuge Teil des Netzes sind, so kann zumindest in urbanen Gegenden angenommen werden, dass stets ein Knoten innerhalb eines Kilometers verfügbar ist. In P2P-Netzen reicht ein Knoten aus, um sicherzustellen, dass die Informationen des Netzes nicht verloren gehen. Voraussetzung ist dabei, dass der Knoten in der Lage ist, alle Informationen zu speichern. Da von textuellen Inhalten ausgegangen wird, benötigen Informationen wenig Speicherplatz. Folglich ist dieses Kriterium im Falle von Fahrzeugen unkritisch. Im Vergleich zu anderen Systemen kann das Ausreichen einer Netzwerkdichte von einem Knoten pro Kilometer auch als Vorteil erkannt werden (vgl. u.a. [MML 04]).

Ein klarer Nachteil ist hingegen, dass strukturierte P2P-Netze Aufwand durch Nachrichten, die zur Aktualisierung der Topologieinformationen benötigt werden, implizieren. Diesem Nachrichtenaufwand stehen bei allen verwandten Arbeiten, die dauerhafte Verfügbarkeit von ortsbezogenen Informationen umsetzen, kontinuierliche Rebroadcasts gegenüber. Die Möglichkeiten des anschließenden Bearbeitens stellt eine Neuerung dar.

5. Zusammenfassung und Ausblick

Um aktive Sicherheits- und Deployment-Anwendungen breitgefächert zu unterstützen, müssen Informationen wie Glatteiswarnungen durch ein Fahrzeugnetz verteilt verwaltet werden. Sie müssen ortsbezogen publizierbar sein und anschließend für die Dauer ihrer Gültigkeit zur weiteren Bearbeitung bereitstehen. Anhand einer Zusammenfassung verwandter Arbeiten wurde gezeigt, dass die genannten Funktionalitäten Herausforderungen der C2X-Forschung darstellen. Vor diesem Hintergrund wurde ein Lösungsansatz vorgestellt, der einen strukturierten P2P-Algorithmus für die ortsbezogene Verwaltung von Informationen einsetzt. Um den Einsatz zu ermöglichen, wird das Verkehrsnetz in Segmente gegliedert, die separate P2P-Netze bilden. Informationen werden von den Fahrzeugen der Segmente verteilt verwaltet und falls nötig unter benachbarten Segmenten weitergereicht. Anwendungen können durch die vorgeschlagene Lösung Informationen ohne Kenntnis von der Fahrzeugnetzsegmentierung ortsbezogen und gültigkeitsbeschränkt publizieren. Anschließend sind die Informationen abfrag-, modifizier- und löschbar. Unser nächster Schritt ist die Implementierung des Lösungsansatzes, um dessen Eignung evaluieren zu können. Dabei soll unter Verwendung eines geeigneten C2X-Simulators insbesondere der zusätzliche Nachrichtenaufwand, der zur Aufrechterhaltung der P2P-Topologie in Kauf genommen werden muss, analysiert werden.

Referenzen

[HMM+ 02] HUANG, Q., R. MILLER, P. MCNEILLE, D. DIMEO und G. C. ROMAN: Development of a peer-to-peer collision warning system. Technischer Bericht, Washington University, Ford Research Lab, 2002.

[ImNa 96] IMIELINSKI, T. und J. NAVAS: GPS-based addressing and routing. Technischer Bericht, The Internet Society, 1996.

[MeTo 08] MEROTH, A. und B. TOLG: Infotainmentsysteme im Kraftfahrzeug. Vieweg-Verlag, 2008.

[MFE 03] MAIHÖFER, C., W. FRANZ und R. EBERHARDT: Stored Geocast. Kommunikation in Verteilten Systemen. Springer, 2003.

[MKR 04] MILLER, B. N., J. A. KONSTAN und J. RIEDL: PocketLens: Toward a personal recommender system. ACM Transactions on Information and System Security, 22(3):437–476, 2004.

[MML 04] MATHEUS, K., R. MORICH und A. LÜBKE: Economic background of car-to-car communication. Proceedings of the 2. Braunschweiger Symposium Informationssysteme für mobile Anwendungen. ACM, 2004.

[Wahl 04] WAHLSTER, W.: SmartWeb: Mobile applications of the semantic web. Advances in Artificial Intelligence, 3238(27):50–51, 2004.

[WER+ 03] WISCHOFF, L., A. EBNER, H. ROHLING, M. LOTT und R. HALFMANN: SOTIS - a selforganizing traffic information system. Proceedings of the 57th Vehicular Technology Conference. IEEE, 2003.

LBS-Applikationen bei temporärer Netzentkopplung mittels Hoarding-Prozesse

Gefördert durch das BMWi innerhalb des Projektes Sm@rtLogistics, Next Generation Media

Prof. Dr.-Ing. habil. Werner Bärwald

Technische Universität Dresden
Fakultät Verkehrswissenschaften „Friedrich List“
Lehrstuhl Verkehrsnachrichtensysteme
01069 Dresden

werner.baerwald@tu-dresden.de

Abstract

Wird ein Hoarding-Bereich als wahrscheinlicher geografischer Aufenthaltsbereich eines mobilen Nutzers definiert und werden gleichzeitig die gespeicherten historischen Aufenthaltsdaten (Tracking Points) für eine Extrapolation des Bewegungsvorganges herangezogen und einem geografischen Informationssystem überlagert, so können Bewegungsvorgänge prognostiziert und den so bestimmbaren künftigen Aufenthaltsorten zuordnete Informationen auch in einem zeitweise entkoppelten Betrieb bereitgestellt werden. Man spricht von Movement Awareness für Location Based Services. Grundlage dafür ist eine entkoppelte Form der Bewegungsverfolgung, wobei mutmaßliche Standorte (Location) aus dem Abgleich der extrapolierten Standortkoordinaten und den in GIS ausgewiesenen Point of Interests mit hinreichender Genauigkeit als Basis für eine ortsbezogene Content-Bereitstellung bestimmbar sind.

1. Problemstellung

Als Hoarding bezeichnet man eine Vorabübertragung von Informationen mit dem Ziel, einen entkoppelten Betrieb zu ermöglichen. Unter entkoppelten Betrieb wird die Betriebsart definiert, in der ein mobiles Endgerät ohne Netzanbindung ist. In dem Fall ist eine netzgestützte Lokalisierung des Endgerätes nicht möglich. Somit können dem Nutzer auch keine auf den augenblicklichen Standort bezogenen Informationen angeboten werden. Die zeithaltende Beseitigung eines zu erwartenden Informationsdefizits an einem künftigen Standort kann durch vorausschauendes Handeln in dreifacher Weise verwirklicht werden:

- Bereitstellung von ortsbezogenem Content, bevor der betreffende Standort erreicht wird,
- Bereitstellung von ortsbezogenem Content in dem Zeitpunkt, zu dem ein betreffender Standort erreicht wird (allgemein als Prefetching-Prozess bezeichnet).
- Bereitstellung von ortsbezogenem Content zu dem Zeitpunkt, zu dem ein bestimmter Standort erreicht ist, aber technisch keine direkte Lokalisierung des Nutzers mehr möglich ist (Hoarding-Prozess).

Location Based Services sind eine Menge von Kommunikationsdiensten, bei denen der Standort des Nutzers einen steuernden Einfluss auf das Erbringen und vor allem auf die Art und Weise der Erbringung des Dienstes besitzt. Standortbezogene Dienste dienen der Aufwertung von Information durch Steigerung ihres Standortbezugs.

2. Standortabhängige Informationen

Location Sensitive Information sind Informationen, die vom Standort eines Nutzers abhängig sind. Das betrifft vorrangig Mobilfunkdienste. Dabei werden in Abhängigkeit vom fixierten Einzugsbereich Anwendungen verwirklicht. Location Sensitive Information ist in einem engen Zusammenhang mit Bewegungsvorgängen zu sehen. Mit Hoarding in der Ausprägung eines Information-Hoarding steht eine Methode zur Nutzung bereit für das vorab zur Verfügung stellen von standortbezogenen Informationen. Die Standortbestimmung ist im vorliegenden Zusammenhang ein Paradoxon. Augenblickliche Standorte mit zugeordneten Positionsdaten sind kurzlebig. Eigentlich gibt es sie in praxi nicht, da der Nutzer mobil ist und sich unmittelbar nach der Standortbestimmung bereits an einer anderen Stelle befindet. Die so vorhandene Ungenauigkeit ist abhängig von der Bewegungsgeschwindigkeit. Ein mit üblichen Verfahren ermittelter Standort /1/ definiert einen begrenzten geographischen Ereignisraum. Dazu ist eine Umkreis-Semantik definierbar, die die Suche nach brauchbaren Informationen, Zeitmuster und Mobilitätsgrade ermöglicht. Diese Semantik ergibt eine funktionale Reichweite in einem definierten Umfeld des ermittelten Standortes. Die Umkreis-Semantik überdeckt den Hoarding-Bereich und wird zu einem logischen Bereich, der aber nicht mit dem in Zellen-Cluster eingeteilten Versorgungsgebiet eines Mobilfunknetzes gleichzusetzen ist.

Wird ein Hoarding-Bereich als wahrscheinlicher geographischer Aufenthaltsbereich eines mobilen Nutzers definiert und werden gleichzeitig die gespeicherten historischen Aufenthaltsdaten (Tracking Points) für eine Extrapolation des Bewegungsvorganges herangezogen und einem geographischen Informationssystem überlagert, so können Bewegungsvorgänge prognostiziert und den so bestimmbaren künftigen Aufenthaltsorten zuordnete Informationen auch in einem zeitweise entkoppelten Betrieb bereitgestellt werden. Man spricht von Movement Awareness für Location Based Services. Grundlage dafür ist eine entkoppelte Form der Bewegungsverfolgung, wobei mutmaßliche Standorte (Location) aus dem Abgleich der extrapolierten Standortkoordinaten und den in GIS ausgewiesenen Point of Interests mit hinreichender Genauigkeit als Basis für eine ortsbezogene Content-Bereitstellung bestimmbar sind. Bei der Bereitstellung von Content gilt es zu unterscheiden zwischen:

- ortsbezogener Selektion,
- ortsbezogener Präsentation und
- ortsbezogener Aktion.

Diese Ortsbezogenheiten sind verknüpft mit kontextbezogenen Nutzerprozessen und Anwendungen. Dafür lassen sich die Kontextparameter wie Zeit, Bewegungsrichtung, Bewegungsgeschwindigkeit und bestimmte Benutzerpräferenzen fixieren. Als Personalisierung bezeichnet man die möglichst genaue Anpassung eines Dienstes an die Bedürfnisse und/oder das Verhalten und die erfassten Gewohnheiten eines Nutzers. Gewohnheiten eines Nutzers können aus bestimmten Bewegungsprofilen abgehoben werden. Damit sind aktuelle Mobilitätsaktivitäten mit aus der Vergangenheit verfügbaren Daten abgleichbar.

3. Standort und Bewegungsprofile

Für mobilfunkgestützte Hoarding-Prozesse ist einen Grundsatz beachten: Nicht die Person, sondern ein Endgerät wird erfasst. Dieser Prozess kann vereinheitlicht werden durch:

- Punktuelle Erfassung wiederholter logischer Ortskomponenten,
- EDV-mäßiger Abgleich der erfassten Koordinaten anhand elektronischer Straßen- oder Umkreis- Semantik (GIS),
- Darstellung einer Wegsemantik mit Abschnitten, wo die Anwesenheit nur gemutmaßt werden kann (Erfassungslücke).

Kombinationen mit bekannten Präferenzen führen zu neuen Kundenprofilen. Hoarding bedeutet hier, eine mit hinreichender Genauigkeit mögliche Vorhersage eines zu erwartenden Standortes und eines diesem Standort zuzuordnenden Benutzerverhaltens zu treffen. Das ist mit einer Simulation der Standortbestimmung eines nicht mehr ortbaren Terminals vergleichbar. Darüber hinaus werden auch alle Verfahren als Hoarding-Verfahren bezeichnet, die bestimmte Daten anhand von statistischen Methoden im Voraus bereitstellen, um im Fall einer späteren fehlenden Netzabdeckung dennoch die Anwendung korrekt ausführen zu können. Der vereinfachte Hoarding-Prozess ist im Bild 1 zu sehen.

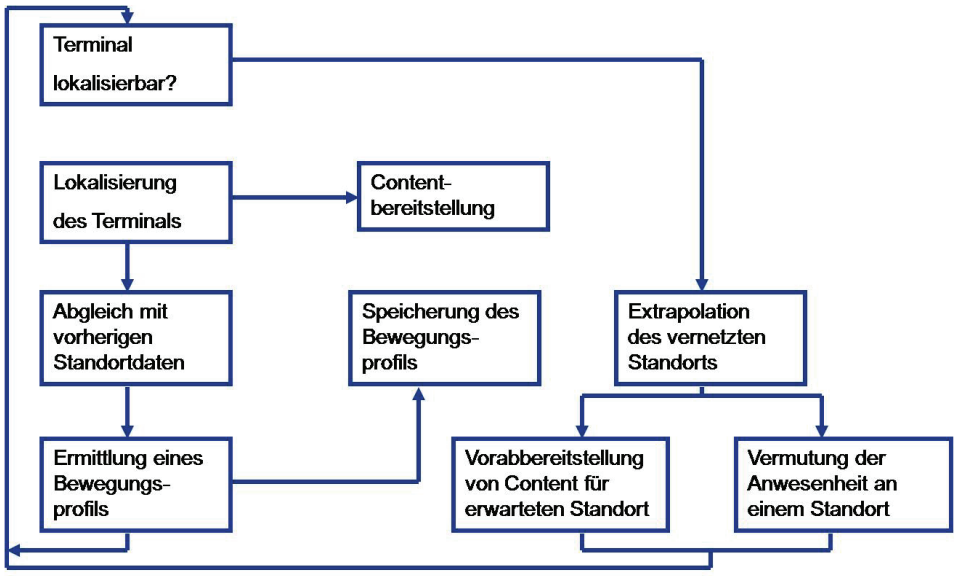


Bild 1: Grundprinzip des LBS-gestützten Hoarding-Prozesses

Hoarding-Prozesse können zwei Ausprägungen haben, die sich in der Art der Selektion unterscheiden:

- Im ersten Fall setzt man voraus, dass der Nutzer weiß, welche Daten (Objekte) er künftig benötigen wird. Das ist in der praktischen Anwendung damit verbunden, dass der Nutzer grundsätzlich Kenntnis über den verfügbaren Content besitzt, was in der Regel nicht gegeben ist.
- Der zweite Fall beinhaltet eine automatische Selektion des zu einem Standort verfügbaren Content. Selektionskriterien können sein der Standort, das Bewegungsprofil, das Nutzerprofil, wobei Präferenzen als ein Verhaltensmuster aus der Historie so abgehoben werden, dass ein möglicher semantischer Zusammenhang ableitbar ist.

Automatisch ablaufende Hoarding-Verfahren unterstützen somit die Bereitstellung von ortsabhängigen Informationen, wobei als Entscheidungskriterium für die gezielte Bereitstellung Bewegungsmuster und Präferenzen sowie der Ortsbezug dieser Informationen sind. Für Hoarding wird ein Verfahren zur Positionsbestimmung mobiler Objekte oder Personen benötigt, das immer dann wirksam werden kann, wenn andere Ortungssysteme nicht mehr nutzbar sind. Das ist an zwei Voraussetzungen gebunden:

- Die Kenntnis der Ausgangsposition oder bekannter Referenzpunkte sind für die Anwendung des Verfahrens zwingend (Dead Reckoning).

- Durch Bestimmung der Bewegungsrichtung, der Streckenmessung und (wenn nötig und möglich) der Beschleunigung bei Geschwindigkeitsänderungen kann der augenblicklich erreichte Standort ermittelt werden.

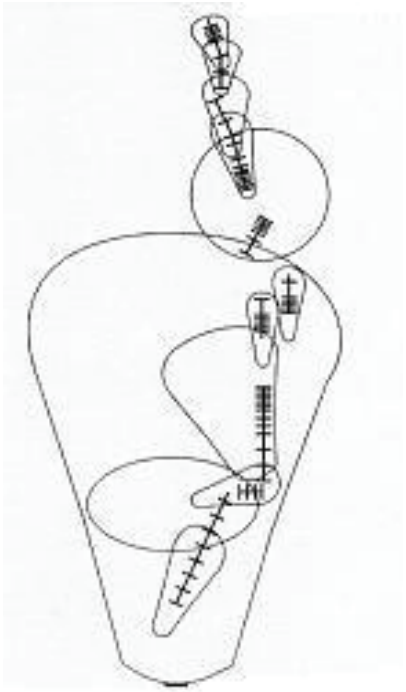


Bild 2: Wegpunkte als Ergebnis einer Standortbestimmung

4. Prophetie eines künftigen Standortes

Eine mehrfach durchgeführte Ortung führt zu Wegpunkten als Ergebnis einer Standortbestimmung. Ein Beispiel dafür wird im Bild 2 deutlich. Wird Content Awareness als bewegungsbedingte Verknüpfung für eine Standortbestimmung mittels extrapolierter Mobilitätsdaten angestrebt, so entsteht folgende Problemstellung:

- Mobile Nutzer und/oder Endgerät hinreichend genau lokalisieren.
- Position einem definierten Bereich (Aufenthaltsbereich) zuordnen.
- Position mit standortbezogenen Datenbeständen verknüpfen.

Die Kenntnis des Nutzerverhaltens (Nutzerprofile, Präferenzen) erlaubt dem Service Provider, ein „maßgeschneidertes“ Diensteangebot zu erstellen. Dieses Diensteangebot (Anwendungen) kann

- auf die Person/Gruppe bezogen,
- auf das Verhalten der Person/Gruppe bezogen,
- auf die Situation bezogen,
- auf zeitliche Kriterien bezogen,
- auf definierte Standort/Lokalitäten bezogen oder
- auf das Verhalten einer Person in Gruppe bezogen

sein.

Typisch ist, dass nicht mit Sicherheit vorhergesagt werden kann, welche Informationen in Zukunft benötigt werden. Auch ist nicht vorher bekannt, wann der Nutzer einen Bereich mit Funkabdeckung verlässt und wann er in diesen oder einen anderen Bereich wieder eintreten wird.

LBS ermöglichen eine Auswertung regelmäßig wiederkehrender Bewegungsvorgänge oder regelmäßig oder zu definierten Zeitpunkten aufgesuchter Lokalitäten. Für das Diensteprofil werden Informationen (Content) mit Orten und Objekten logisch verknüpft. Die Informationen können sich auf ortsfeste und/oder auf mobile Objekte beziehen. Das verlangt eine Erfassung der Daten mit hinreichender Genauigkeit und eine Auswertung in einer EDV-Basis.

5. Anwendungsbeispiele

Standortbezogene Dienste führen zu einem Mehrwert, indem sie Lokalisierungsdaten mit geographischen Zusatzinformationen (Standortinformationen) verknüpfen und in kundenspezifische Anwendungen integrieren. Damit lassen sich eine Vielzahl nutzerabhängiger Anwendungen entwickeln. Anwendungen für Hoarding-Prozesse kann man in Indoor- und in Outdoor-Lösungen erkennen. In beiden Fällen liegen folgende Fragestellungen zugrunde:

- Wann wird welcher Standort erreicht?
- Welche Bedingungen sind an diesem Standort vorzufinden?
- Wie kann das damit verbundene Informationsdefizit beseitigt werden?

Aus Bild 2 wird auch deutlich, dass mit Tracking-Prozessen das Bewegungsprofil von Personen erfassbar ist. Wird dieses Bewegungsprofil einem GIS-System überlagert, dann lassen sich künftig zu erreichende Standorte abschätzen. Mit der Prophetie eines künftigen Standortes können standortbezogene Informationen verknüpft werden. So lassen sich zwei Fragestellungen beantworten:

- Was ist zu erwarten wenn … ?
- Was ist zu veranlassen, weil … ?

Diese Fragestellungen können auch in der anderen Richtung betrachtet werden. Dann gilt:

- Wo komme ich hin?
- Wann komme ich dorthin?

Auf mobile Kommunikationsdienste abgestützte Touristenführer können in einem Hoarding-Bereich alle interessanten Objekte vorhalten. Aufgrund von Präferenzen eines bestimmten Nutzers kann darüber eine gezielte Auswahl in der Content-Bereitstellung getroffen werden. So lassen sich alle sonstigen und wahrscheinlich nicht im Interessenfeld des Nutzers liegende Points of Interest ausblenden. Ebenso können aus der Content-Bereitstellung alle Points of Interest eliminiert werden, die im Augenblick des Besuchs des Hoarding-Bereichs z. B. durch Schließzeiten von Kulturstätten nicht verfügbar sind. Der Nutzer muss über die Hoarding-Bereiche und über die Bereitstellungsmodi von Content nicht in Kenntnis gesetzt werden. Es handelt sich um ein unterstützendes System, das im Hintergrund arbeitet.

Andere Anwendungen für derartige Hoarding-Prozess-gestützte Verfahren lassen sich im Verkehrsmanagement abheben. Das Ziel, eine individuelle, situationsbezogene Bereitstellung von Informationen über einen noch nicht erreichten Standort zu verwirklichen, führt stets zu einem gleichen Prozess, der die Ortung und das Abheben eines Bewegungsprofiles, die Überlagerung dieses Bewegungsprofils mit einer geografischen Karte und dem Ableiten daraus, wann der künftige Standort errecht wird. Das ist interessant, um Informationen z. B. über touristische Attraktivitäten bereitzustellen, um am betreffenden Ort eine Fahrtunterbrechung zu veranlassen. Gleichermaßen können so Informationen über den Straßenzustand gegeben werden, um vorbeugend ein bestimmtes Fahrverhalten zu erreichen (Aquaplaning, Glatteis). Informationen über bestehende oder zu erwartende Verkehrssituationen sind bereitstellbar, um diese in eine individuelle, situationsabhängige Routenempfehlung umzusetzen. Dabei sind entsprechende kommunikationstechnische Komponenten notwendig.

Anwendungen in der Verkehrlenkung und –leitung sind die Grundlage für regionales und/oder überregionales Management großer Verkehrsströme. Nach Ortung der mobilen Objekte und einer Verdichtung der Aussagen für bestimmte Trassen kann aus dem kollektiven Bewegungsprofil die künftige Belastung eines Knotenpunktes oder eines Trassenabschnittes und damit die Situation an einem bestimmten Ort abgeleitet werden. Anwendung für diese Prozesse findet man in der kollektiven Verkehrslenkung zur Entlastung kritischer Standorte durch alternative Routenempfehlungen, bevor die Überlast erreicht wird. Hier ist die kollektive Verkehrslenkung durch Vorgabe von maximalen und minimalen Fahrtgeschwindigkeiten ebenso einzuordnen wie die kollektive Information, um durch Verkehrsvermeidungen oder durch zeitliche Verlagerung temporäre Überlast zu vermeiden. Das Herangehen ist mit den in Telekommunikationsnetzen praktizierten Strategien vergleichbar.

Die klassische Anwendung für Hoarding-Prozesse findet man in dynamischen Navigationssystemen. Im Abgleich mit dem erfassbaren Bewegungsprofil ist der voraussichtliche Informationsbedarf zu vermuten. Es müssen nicht alle Informationen vorgehalten werden. Eine Konzentration auf die der zu erwartenden weiteren Wegpunkten entsprechenden Information ist machbar. Da der künftige Aufenthalt an einem bestimmten Punkt

abschätzbar ist, werden die Informationen vorab selektiert, die in die Umkreis-Semantik eingeordnet werden können. Diese Informationen werden im Bedarfsfall präsentiert.

Die Umkehrung dieses Verfahrens liegt bei der Ortung in Notfällen vor. Der letzte Standort ist ebenso wie die Bewegungsrichtung bekannt. Die Bewegung einer geographischen Karte überlagert, lässt eine Schlussfolgerung zu, wo ein bestimmtes Terminal sich befinden müsste, wenn die Bewegungsrichtung und die -geschwingkeit annähernd konstant geblieben sind. Für Bewegungsvorgänge im vor allem kommunikationstechnisch unterversorgten Raum (Abschattungen) oder in Gebieten mit unzureichender funktechnischer Abdeckung (Tunnels, Inhouse-Bereiche) wird es schwierig oder gar unmöglich, den augenblicklichen Standort eines mobilen Objektes zu bestimmen. Die Historie der Standortdaten ist jedoch im System bekannt. Daraus lassen sich die Bewegungsprofile ermitteln.

Verkehrsinformationen im Straßenverkehrsnetz sind entweder auf einen augenblicklichen Standort oder auf eine geplante Fahrtroute bezogen. Mit einem auf Hoarding-Prozesse gestützten vorausschauenden Management kann die Fahrtroute analysiert und extrapoliert werden. Somit wird es möglich, den zu einer bestimmten Zeit erreichten Punkt abzuschätzen. Sollte sich auf dem Weg zu diesem Punkt ein Verkehrshindernis gleich welcher Art befinden, so kann der dann künftig von diesem Sachverhalt betroffene Verkehrsteilnehmer vorab individuell (personalisiert) gewarnt und auf eine alternative Route geführt werden.

Während die Bereitstellung von Content für einen bestimmten Standort bei einem mutmaßlichen Aufenthalt eines Nutzers zu einem bestimmten Zeitpunkt an einem bestimmten Ort i. d. R. Ausgangspunkt für Hoarding-Prozesse ist, finden sich im Verkehrs- und Mobilitätsmanagement weitere davon abweichende Anwendungen. Oft ist dem Nutzer keine Information auf sein persönliches Terminal zu übermitteln. Es genügt, die Anwesenheit eines Nutzers zu vermuten, um erstens kollektive Verkehrsleitsysteme zu steuern und zweitens ein zweckentsprechendes Verkehrsmanagement zu gewährleisten (Verkehrsverlagerung, Verkehrsvermeidung, Verkehrsverdrängung u. a.) Anwendungen für derartige auf Hoarding-Prozess-gestützte Verfahren finden sich im Verkehrswesen, wenn Telematik-Endgeräte in Bereichen ohne Funkversorgung weiter betrieben werden müssen. Das kann Tunneldurchfahrten ebenso betreffen wie Aufenthalte in Tiefgaragen oder in ungenügend versorgten bergigen Regionen sowie in ausgewiesenen Indoor-Bereichen wie Produktions- oder Messehallen.

6. Ausblick

Mit der Standortbestimmung sich bewegender Objekte, mit der Erzeugung von Bewegungsprofilen, mit der Überlagerung dieser Profile auf geografische Informationssysteme und mit der Auswertung erfasster Präferenzen der Nutzer sind eine Vielzahl individuell und kollektiv wirkender, nutzerspezifischer Anwendungen zu verwirklichen. Überaus wichtig ist dabei, dass neben den reinen technischen Möglichkeiten der Verwirklichung derartiger Lösungen auch die entsprechenden Datenschutzbestimmungen in jedem Fall zu beachten sind.

Literatur

/1/ Bärwald, W.; Baumann, S.; Keil, R.; Richter, K.: Positionsbestimmung und Identifikation für ein innovativeres Verkehrsmanagement. In: Strobl/Blaschke/Griesebner (Hrsg.): Angewandte Geoinformatik 2007 – Beiträge zum 19. AGIT-Symposium Salzburg, S. 41 – 47, Wichmann Verlag Heidelberg 2007

/2/ Riegelmayer, W. P.: IN - Eine verteilte Service-Plattform mobiler Prozeßarchitekturen für verkehrstelematische Anwendungen. TUDpress, Dresden 2006

/3/ Bärwald, W.: Von der Standortinformation zu einem personalisierten Diensteprofil. In Sächsische Fachtagung „Location Based Services“. Sächsisches Telekommunikationszentrum April 2006, Tagungsband S. 6 – 59

/4/ Bärwald, W.: Location Based Services. expert verlag Renningen (in Vorbereitung)

/5/ Ammoser, H.; Bärwald, W.; Baumann, S.: Location Based Services für ein innovatives Verkehrsmanagemt. 20. Verkehrswissenschaftliche Tage der TU Dresden, Tagungssektion 3, Dresden Sept. 2005

/6/ Bärwald, W.; Baumann, S.: LBS- und GIS-gestützte Hoarding-Prozesse. In: Strobl/Blaschke/Griesebner (Hrsg.): Angewandte Geoinformatik 2008 – Beiträge zum 20. AGIT-Symposium Salzburg, S. 254 – 259, Wichmann Verlag Heidelberg 2008

/7/ Bärwald, W.; Ammoser, H.; Stuhr, A.: Nutzung personengebunder Daten bei Location and Personalization Based Services zur Unterstützung touristischer Aktivitäten. In: Freyer, W.; Groß, S.(Hrsg.): Sicherheit in Touristik und Verkehr – Schutz vor Risiken und Krisen . FIT Forschungsinstitut für Tourismus, Dresden 2004

/8/ Bageor, J.: Das Handy kennt den Weg. http.//www.heise.de/ct/01/22/168/default.html, Stand August 2002

Kontext- und Ontologie-basiertes persönliches Informationsmanagement für mobile Endgeräte

Wolfgang Wörndl

Technische Universität München
Institut für Informatik
Boltzmannstr. 3
85748 Garching b. München

Tel: +49 89 289-18650
Fax: +49 89 289-18657
E-Mail: woerndl@in.tum.de

Abstract

In diesem Artikel wird die Realisierung eines „semantischen Desktops“ für mobile Endgeräte wie PDAs vorgestellt. Dabei kann der Benutzer eine persönliche Ontologie definieren und erweitern, und Ressourcen aus seinem persönlichen Informationsraum zuordnen. Neben einer Abfrage der Beziehungen können die semantischen Strukturen dann genutzt werden, um Empfehlungen für Dokumente, Kontakte und andere Items für den Benutzer zu generieren. Bei der Empfehlung kann man unterscheiden zwischen Empfehlungen im persönlichen Informationsraum und Empfehlungen anderer Items. In ersten Fall werden Items abgeleitet, die aktuell für den Benutzer wohl interessant sind. Der Ansatz verwendet eine konfigurierbare Evaluierungsfunktion, um die semantischen Beziehungen zwischen den Ressourcen zu gewichten und zu bewerten. Empfehlungen weiterer Items wie Points-of-Interests außerhalb des persönlichen Informationsraums erfolgen auch auf Basis der Ontologie und können zusammen mit ortsbezogenen Daten im persönlichen Informationsraum auf einer Karte angezeigt werden. Die entwickelten Konzepte wurden in einem Prototyp für Windows Mobile PDAs implementiert und getestet.

1. Einführung

Das Ziel von *persönlichem Informationsmanagement* (personal information management, PIM) ist es, die Aktivitäten von Menschen zur Organisation ihrer Datenhaltung im täglichen Leben zu unterstützen. PIM beinhaltet die Akquise, Verwaltung, Abfrage und den Austausch von Informationen wie Kontaktdaten, Aufgabenlisten oder Terminen [1]. Dies erscheint besonders wichtig in einem mobilen Szenario, zum Beispiel die Unterstützung eines Außendienstmitarbeiters beim Besuch seiner Kunden. Im mobilen Umfeld werden dazu oft PDAs (Personal Digital Assistants) und andere mobile Endgeräte eingesetzt. Allerdings ist die Organisation der persönlichen Daten auf mobilen Geräten im Vergleich zu Desktop Rechnern noch schwieriger. Dies liegt an den Einschränkungen von mobilen Geräten in Bezug auf Eingabe und Ausgabemöglichkeiten, Speicher- und Rechenleistung, sowie die Verfügbarkeit eines Datennetzes. Zum Beispiel ist es auf dem kleinen Display eines PDAs umständlich, eine Liste von Suchergebnissen zu sichten, um ein bestimmtes Dokument zu finden. Deshalb ist es gerade im mobilen Umfeld wichtig, den Zugriff auf Informationen zu personalisieren und an den aktuellen Kontext des Benutzers anzupassen. Unter dem Kontext wird im folgenden dazu insbesondere Ort und Zeit verstanden.

Ein möglicher Lösungsansatz als Grundlage für PIM ist der *Semantic Desktop* [2]. Die Grundidee dabei ist es, eine anwendungsübergreifende Datenverwaltung durch Semantic Web Technologien wie Ontologien zu unterstützen. Benutzer können dazu eine persönliche Kategorisierung anlegen, und Ressourcen ihres Informationsraumes zuordnen. Eine Ressource ist dabei eine einzelne Datei wie ein Dokument oder ein Eintrag im Kalender oder Adressbuch eines Benutzers. Die entstehende Ontologie kann dann genutzt werden, um das Abfragen und Wiederfinden von Ressourcen zu erleichtern. Ein Beispiel für eine Semantic Desktop Implementierung ist das *Gnowsis* System [2]. Gnowsis besteht aus zwei Teilen, nämlich dem Gnowsis Server, der die Verarbeitung und Speicherung der Daten sowie die Interaktion mit anderen Anwendungen übernimmt, und einem Teil für die graphische Benutzeroberfläche, die Java Swing oder Web-basierte Schnittstellen zur Verfügung stellt.

Allerdings sind alle bisher bekannten Semantic Desktop Lösungen wie Gnowsis auf einen Betrieb an einem Desktop Rechner oder Server abstimmt. Wie ausgeführt wäre jedoch eine Semantic Desktop Lösung für mobile Endgeräte interessant, die es aber bisher so nicht gibt. Ziel dieser Arbeit ist es daher, die Idee des Semantic Desktops auch für PDAs zu realisieren. Dabei sollte eine Anwendung entwickelt werden, die als eigenständige Anwendung abläuft – also kein mobiler Client einer Server-basierten Lösung –, Daten wie Telefonanrufe und SMS mit integriert und die Einschränkungen in der PDA-Benutzerschnittstelle berücksichtigt.

Die Sammlung der verwalteten Daten im persönlichen Informationsmanagement wird als „persönlicher Informationsraum“ bezeichnet, eine einzelne Ressource wird in diesem Artikel auch „Item“ genannt, da dies eine gängige Bezeichnung von Gegenständen in einem Empfehlungsprozess ist. Der Rest dieses Artikels ist folgendermaßen aufgebaut. Zunächst erklären wir unseren Ansatz zur Verwaltung einer persönlichen Ontologie in unserer mobilen Semantic Desktop Anwendung. Dann gehen wir in Kapitel 3 darauf ein, wie diese Ontologie genutzt werden kann, um für den Benutzer aktuell relevante Ressourcen in seinem persönlichen Informationsraum abzuleiten. Kapitel 4 behandelt die Empfehlung weiterer Items wie Points-of-Interests und deren Anzeige auf einer Karte. Der Beitrag schließt in Kapitel 5 mit einer knappen Zusammenfassung und einem Ausblick auf zukünftige Arbeiten.

2. Verwaltung einer persönlichen Ontologie

Zunächst erklären wir in diesem Kapitel die PIMO Ontologie, die als Grundlage für unseren Semantic Desktop Ansatz dient, und zeigen anschließend wie eine persönliche Ontologie in unserer mobilen Lösung verwaltet werden kann.

2.1. Hintergrund: PIMO

Der Semantic Desktop Ansatz baut auf Ontologien auf, um Beziehungen zwischen Ressourcen formalisieren zu können. Eine Ontologie ist eine explizite formale Kategorisierung von Konzepten. Für das bereits angesprochene Gnowsis Projekt wurde die *Personal Information Model Ontology (PIMO)* Ontologie entwickelt [3], auf die auch unser Ansatz aufbaut. PIMO ermöglicht es dem Benutzer, seine Sichtweise auf seinen persönlichen Informationsraum, der z.B. seine Dokumente, Nachrichten, Termine, Kontakte und Aufgaben umfasst, zu definieren. Dazu unterscheidet PIMO zwischen zwei Grundkonzepten. Das eine dient zum Aufbau einer logischen Struktur, wie z.B. das Anlegen von „Projekten“, „Organisationen“, oder Unterklassen von „Person“. Diese Klas-

sen sind dabei Unterklassen von „Thing“ und repräsentieren die konzeptuelle Ebene des persönlichen Informationsraums. Darüberhinaus gibt es Unterklassen von „RessourceManifestation“. Dies sind die konkreten Objekte im persönlichen Informationsraum, die z.B. in Klassen wie „File“, „CalendarEvent“ oder „Task“ strukturiert sind. Eine Instanz einer dieser Klassen ist dann die Repräsentation einer konkreten Ressource, z.B. einer Datei im Dateisystem. Benutzer können nun Instanzen anlegen oder importieren, und diese den logischen Konzepten zuordnen. Dazu können Relationen verwendet werden, wozu PIMO einige definiert, z.B. „InstanceOf“ oder „HasPart“.

2.2. SeMoDesk

Wir haben PIMO und die Semantic Desktop Idee in der Anwendung *SeMoDesk* für mobile Endgeräte umgesetzt [4]. Das Zielgerät ist ein Microsoft Windows Mobile 5 oder 6 PDA mit Touchscreen Interface. Die meisten dieser Geräte haben zudem eine Telefonfunktion, sowie viele auch einen GPS-Empfänger integriert. Wir haben SeMoDesk unter anderem mit einem HTC P3600 getestet. Die Implementierung erfolgte in C# mit der Microsoft Visual Studio IDE. Die aktuelle SeMoDesk Version läuft unter dem .NET Compact Framework 3.5.

Nach dem Start hat der Benutzer die Option, seine Ontologie zu verwalten und abzufragen, sich Ressourcen empfehlen zu lassen (siehe Kapitel 3), oder Items auf einer Karte anzeigen zu lassen (siehe Kapitel 4). Zur Verwaltung der Ontologie kann der Benutzer im Ontologie-Baum blättern, wie in Abbildung 1, links, gezeigt. Der Benutzer kann und soll dabei die Ontologie um eigene Konzepte (d.h. Subklassen der vorgegebenen Top-Level Ontologie) erweitern, um seinen Informationsraum logisch zu strukturieren. Ein Beispiel ist die Untergliederung von Personen, die der Benutzer verwalten möchte. Dann können Ressourcen, z.B. Kontaktdaten aus MS Outlook, der logischen Struktur zuordnet werden, indem entsprechende Relationen definiert werden.

Instanzen von Klassen können zwar manuell angelegt werden, dies ist aber normalerweise nicht erforderlich, denn SeMoDesk unterstützt den Benutzer bei dieser Aufgabe soweit es möglich ist. Zum Beispiel werden bei einem Telefonanruf oder einem Nachrichtenaustausch per SMS oder Email automatisch Relationen zur betreffenden Person angelegt, falls diese im Adressbuch des Benutzers gefunden wird (siehe auch das Beispiel in Abb. 1, Mitte). Daten wie Kalendereinträge, Aufgaben oder Kontaktdaten können außerdem aus Microsoft Pocket Outlook integriert werden. Für Ressourcen im Dateisystem, steht ein „File Sniffer“ zur Verfügung (Abb. 1, rechts). Dabei kann der Benutzer einen Pfad sowie ein Dateimuster angeben, neue Dateien werden dann automatisch bei Start von SeMoDesk als Ressourcen integriert. Der „Target Folder“ wie in Abbildung 1, rechts, gezeigt, gibt dabei an, an welche Stelle im Baum der Ressourcen die Dateien erscheinen sollen. In dem gezeigten Beispiel würden diese unter „My Pictures“ zur Verfügung stehen. Die erfassten Daten können nun komfortabel abgefragt werden, wie das Beispiel in Abbildung 1, Mitte, zeigt. Dies erfolgt Anwendungsübergreifend durch die definierte Ontologie und die – manuell und automatisch – angelegten Relationen. Es gibt auch eine Möglichkeit der chronologischen Darstellung zeitbezogener Daten, wobei alle Ressourcen eines Tages, Monats oder Jahres übersichtlich dargestellt werden können.

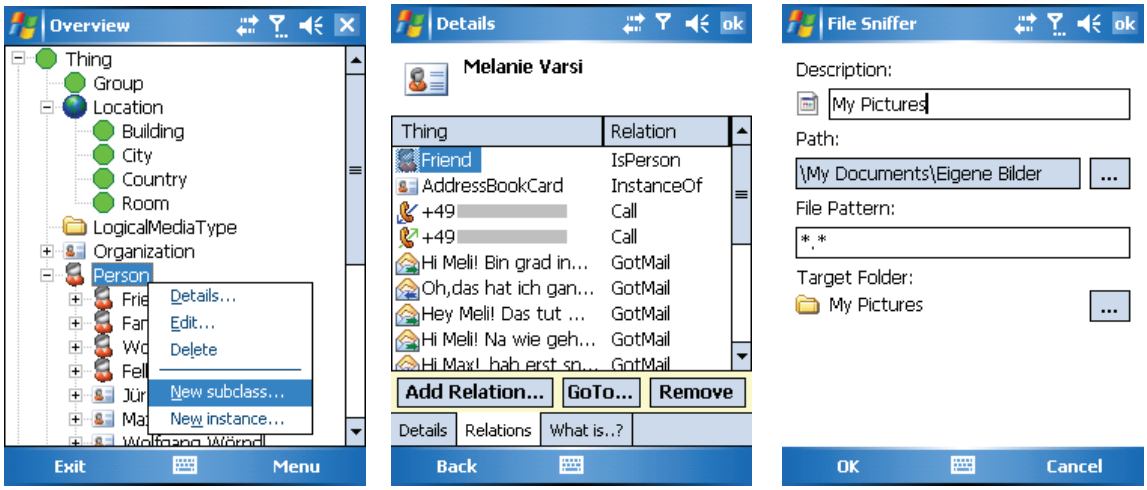


Abbildung 1: Verwaltung der Ontologie (links), Abfrage der Details zu einer Person (mitte), Konfiguration des File Sniffers (rechts)

3. Ableitung von Empfehlungen im persönlichen Informationsraum

Das Browsen und Abfragen von Konzepten und Ressourcen ist allerdings nicht ausreichend, da dabei immer nur direkte Beziehungen angezeigt werden. Wir haben daher eine intelligentere Empfehlungsfunktion entworfen und implementiert [5], die in diesem Kapitel erläutert wird.

3.1. Vorgehensweise und Benutzerschnittstelle

Um im aktuellen Kontext relevante Knoten im Graphen, der durch die persönliche Ontologie gebildet wird, ableiten zu können, ist es zunächst nötig, passende Startknoten für die Suche zu finden. Aus diesem Grunde besteht unser Algorithmus aus den folgenden beiden Schritten:

1. Dem Finden aktuell interessanter Ressourcen und Auswahl eines Startknotens
2. Dem Ableiten weiterer empfehlenswerter Ressourcen

Für den 1.Schritt bietet unser System verschiedene Möglichkeiten an. Der Benutzer kann zum einen manuell ihn aktuell interessierende Knoten auswählen (über Schaltflächen „Concepts“ und „Resources“, siehe Abb. 2, links). Zum anderen kann das System Items vorschlagen. Dies geschieht durch Auswahl der Schaltflächen „On Schedule“ bzw. „In Area“. Im ersten Falle ermittelt das System aktuelle Termine. Der relevante Zeitraum kann eingestellt werden und wird oben angezeigt (Abb. 2, links). Als Ergebnis dieses 1.Schritts wird dem Benutzer eine Liste von Ressourcen angezeigt, die schon interessante Informationen für den Benutzer enthalten kann, aber hauptsächlich als Liste möglicher Startknoten für den 2.Schritt dient. Der Benutzer kann jetzt eine Ressource auswählen und den eigentlichen Suchalgorithmus starten.

Abbildung 2: Anzeige aktuell relevanter Ressourcen (links), Ergebnis der Suche (Mitte), Anzeige des Pfades zu einem Ergebnis (rechts)

Abb. 2, Mitte, zeigt das Ergebnis einer durchgeführten Suche. In diesem Beispiel wurde nur eine empfehlenswerte Ressource mit einem „Match“ von 41% gefunden. „Match“ gibt an, wie relevant die Evaluierungsfunktion im Suchprozess eine Ressource einstuft, was im folgenden Abschnitt 3.2. näher erläutert wird. Nur Items mit einem Wert oberhalb eines konfigurierbaren Grenzwerts werden angezeigt. Die Items in der Ergebnisliste werden anhand des Evaluierungswertes sortiert, falls mehrere Items gefunden werden. In der unteren Hälfte von Abb. 2, Mitte, finden sich einige Symbole für Ressource-Typen, mit Hilfe derer der Benutzer den gewünschten Typ der Ergebnisse auswählen kann. Obwohl nur Ressourcen als Ergebnis angezeigt werden, erfolgt die Suche sowohl über Konzept- als auch Ressourcen-Knoten.

Zur Erklärung, warum eine Ressource gefunden wurde, kann der Benutzer den verwendeten Suchpfad anzeigen. In dem Beispiel in Abbildung 2, rechts, konnte eine Verwandtschaft des Ergebnisknotens mit dem Startknoten über einen gemeinsamen Ort abgeleitet werden. Eine Begründung von Empfehlungen ist generell eine wichtige Eigenschaft von Personalisierungssystemen.

3.2. Suchalgorithmus und Evaluierungsfunktion

Aufgrund der Einschränkungen in der Rechenleistung eines PDAs wurde kein logisches Schlussfolgern in der Ontologie, sondern ein weniger aufwändiger Suchalgorithmus realisiert. Der Suchalgorithmus bewertet dabei alle Knoten ausgehend vom spezifizierten Startknoten mit Hilfe einer Evaluierungsfunktion. Dazu wird eine Breitensuche durchgeführt und Folgeknoten auf Basis der in Abschnitt 2.1 angesprochenen Relationen zwischen Konzepten und Ressourcen expandiert und bewertet. Der Algorithmus terminiert, wenn alle maßgeblichen Knoten analysiert wurden. Die Evaluierungsfunktion bewertet nicht nur den aktuell betrachteten Knoten, sondern auch den Pfad vom Startknoten.
Wir haben die Evaluierungsfunktion f wie folgt definiert: $f = a * depth + b * concept + c * relation$

a, b und c sind Parameter, um die drei verwendeten Faktoren zu gewichten:

- *depth* ist die Entfernung zum Startknoten, je weiter ein Knoten vom Startknoten im Baum entfernt ist, desto niedriger wird er gewichtet. Der Koeffizient für jedes Pfadsegment kann konfiguriert werden, ein sinnvolles Beispiel wäre es, mit 1 zu beginnen und bei jedem weiteren Schritt durch 2 zu teilen
- *concept* ist das Gewicht des Knotens selbst, in Abhängigkeit von seinem Typ. Ressourcen, die weniger häufig vorkommen, können z.B. höher gewichtet werden
- *relation* ist das Gewicht des Pfades zum Knoten, also der Relation, die der Algorithmus verfolgt hat, um zum aktuellen Knoten zu gelangen. Verschiedene Typen von Relationen können und sollten dabei unterschiedlich gewichtet werden

Je höher ein Knoten von der Funktion f bewertet wird, desto relevanter wird er von unserem System im aktuellen Kontext einschätzt. Alle genannten Parameter können derzeit in der Benutzerschnittstelle von SeMoDesk konfiguriert werden. Dies dient aber lediglich Testzwecken, ein Endbenutzer würde dies nicht machen müssen, sondern eine auf Basis des Anwendungsszenarios vorkonfigurierte Parametermenge verwenden. Zum Beispiel könnte man in einem Szenario mit vielen Nachrichten, aber weniger anderen Ressourcen, den Nachrichtenknoten ein niedrigeres Gewicht zuweisen. Genauso ist es wichtig, den verschiedenen Relationen verschiedenes Gewicht zuzuweisen. Zum Beispiel sind alle Nachrichten Instanzen einer gemeinsamen Oberklasse, wären darüber also relativ eng verwandt, was aber keine aussagekräftige Aussage für die Suche ist. Eine Verwandtschaftsbeziehung „related“ zwischen beispielsweise einem „Project“ und einem „Topic“, die der Benutzer manuell angelegt hat, ist dagegen viel aussagekräftiger und sollte hoch gewichtet werden.

4. Anzeige von ortsbezogenen Ressourcen und Points-of-Interests auf Karte

Abbildung 3: Zuweisung einer Aufgabe zu einem POI Typ (links), Zuweisung von Terminen zu Adressen (Mitte), Anzeige auf Karte (rechts)

Die bis jetzt in diesem Beitrag erläuterten Konzepte ermöglichen eine Verwaltung, Abfrage und Empfehlung von Ressourcen im persönlichen Informationsraum eines Benutzers. Desweiteren kann die persönliche Ontologie dazu genutzt werden, um zusätzliche Items zu empfehlen, die nicht explizit von Benutzer erfasst und

verwaltet wurden. Ein Anwendungsszenario ist, dass ein Benutzer Points-of-Interests (POIs) in der aktuellen geographischen Umgebung sucht, um spezifizierte Aufgaben ausführen zu können. Dazu haben wir die PIMO Ontologie um ein „POI" Konzept mit Subkonzepten wie „Tankstelle", „Kino" oder „Restaurant" erweitert. Der Benutzer kann in SeMoDesk Aufgaben, oder beliebige andere Ressourcen, diesen POI-Typen zuweisen, was in Abbildung 3, links, gezeigt ist. Zusätzlich können Termine, oder wieder beliebige andere Ressourcen, zu Adressen zugeordnet werden (Abb. 3, Mitte).

Nach Aufruf der SeMoDesk Kartenfunktion werden dann aktuell relevante POIs und Termine gemeinsam auf einer Karte angezeigt (Abb. 3, rechts). Die persönliche Ontologie dient somit auch als „semantisches Benutzermodell" das ausdrückt, was aktuell für den Benutzer relevant ist, und kann mit Ressourcen außerhalb des persönlichen Informationsraum verknüpft werden. Für die POI-Suche und Kartenanzeige wird in der aktuellen Implementierung Microsoft MapPoint verwendet, andere ähnliche Dienste könnten aber ohne größere Probleme eingebunden werden. Für die Kartenfunktion ist eine Internet-Verbindung nötig, während die restliche SeMoDesk Anwendung als eigenständige Applikation ohne Server abläuft. Um einen geeigneten Kartenausschnitt zu bestimmen, kann die aktuelle Position des Benutzers mittels eines ins Mobilgerät einbauten GPS-Empfängers oder einem anderen Lokalisierungsverfahren ermittelt werden.

5. Fazit

In diesem Beitrag haben wir einen Ansatz für mobiles persönliches Informationsmanagement auf Basis der Semantic Desktop Idee vorgestellt. Dabei können Benutzer eine persönliche Ontologie definieren und verwalten, um ihren Informationsraum zu strukturieren. Die Ontologie kann dann insbesondere dazu genutzt werden, um Items innerhalb und außerhalb des persönlichen Informationsraums zu empfehlen, die aktuell für den Benutzer interessant erscheinen. Die Ideen wurden in der SeMoDesk Anwendung für Microsoft Windows 5/6 PDAs implementiert und die Funktionalität getestet. Wir haben noch keine Benutzerstudie durchgeführt, um das System auch aus Benutzersicht evaluieren zu können, dies ist aber für die Zukunft geplant. Eine Integration und Synchronisation mit anderen, nicht-mobilen, Semantic Desktop Anwendungen ist prinzipiell sinnvoll und wäre relativ einfach machbar, da SeMoDesk auf bestehenden Ansätzen wie der PIMO Ontologie aufbaut. Eine Realisierung dieser Integration liegt aber nicht im Fokus dieser Forschungsarbeit.

Weitere Ansätze für zukünftige Arbeiten liegen insbesondere in einer Verbesserung der Kontextsensitivität des Systems. Wir arbeiten dazu an einer Integration mit einer RFID-Infrastruktur, um auch in Gebäuden eine Lokalisierung des Benutzers vornehmen zu können. Ein Beispiel wäre die Anzeige relevanter Dokumente, wenn ein Benutzer einen Besprechungsraum betritt. Dazu ist ein erweitertes Orts- bzw. Kontextmodell in SeMoDesk nötig, damit der Benutzer Ressourcen auf verschiedenen Ebenen lokalisieren kann (z.B. „Stadt", „Adresse", „Raum"). Darüberhinaus erscheint das automatische Erlernen von Beziehungen interessant, um den Benutzer von einer manuellen Eingabe gerade auf dem mobilen Gerät zu entlasten. Die eine Alternative für das Erlernen ist es, den Inhalt von Ressourcen zu betrachten und z.B. beim Auftreten gemeinsamer Schlüsselwörter in einem Dokument und einer Email-Nachricht, eine Verwandtschaftsbeziehung ableiten und dem Benutzer zur Aufnahme in das System vorschlagen zu können Eine andere Möglichkeit ist es, die Aktionen des Benutzers zu beobachten und daraus Beziehungen abzuleiten.

Referenzen

[1] Teevan, J., and Jones, W., *Personal Information Management*. Combined Academic Publ., 2008

[2] Sauermann, L., Grimnes, G.A., Kiesel, M., Fluit, C., Maus, H., Heim, D., Nadeem, D., Horak, B., and Dengel, A., "Semantic Desktop 2.0: The Gnowsis Experience". *Proc. 5th International Semantic Web Conference*, Springer LNCS 4273, 2006

[3] Sauermann, L., van Elst, L., and Dengel, A., "PIMO – a Framework for Representing Personal Information Models". *Proc. of I-Semantics' 07*, 2007

[4] Woerndl, W., and Woehrl, M., "SeMoDesk: Towards a Mobile Semantic Desktop", *Proc. Personal Information Management (PIM) Workshop*. CHI 2008 Conference, Florence, Italy, 2008

[5] Hristov, A., *A Context-Sensitive Recommendation System for a Mobile Semantic Desktop Application*. Diplomarbeit, Fakultät für Informatik, Technische Universität München, 2008

Praxisbericht Lokale News

Fabian Linke

YellowMap AG
Wilhelm-Schickard-Str. 12
76131 Karlsruhe

Abstract

Die YellowMap AG ist spezialisiert auf mobile Dienste mit modernster Karten-Technologie. Neben einem ausführlichen Branchenbuch sowie einer Filialsuche für Großkunden, betreibt die YellowMap AG verschiedene Forschungsprojekte.
Diese konzentrieren sich auf verschiedene Aspekte im Bereich der Location Based Services. Ferner dienen Forschungsprojekte dazu neue Anwendungen im mobilen Umfeld zu erkennen und die Grundlagen für eine wirtschaftliche Nutzung zu schaffen. Dies können einerseits Anwendungen für mobile Endegeräte andererseits aber auch browserbasierte standortbezogene Anwendungen sein.
Eine dieser Anwendungen ist in der Anzeige von lokalen News in einer Karte. Diese Nachrichten werden dynamisch generiert und angezeigt. Dadurch können schnell aktuelle Informationen und Ereignisse ansprechend präsentiert werden.
Diese Arbeit soll einen Überblick über die verschiedenen Forschungsaktivitäten geben und Ergebnisse präsentieren.

1. Wer ist YellowMap?

Unter dem Motto "Alles in Ihrer Nähe"; strebt die YellowMap AG eine führende Rolle im Markt für Location Based Services (positionsabhängige Dienste) an. Zu den Kernkompetenzen des in Karlsruhe ansässigen Unternehmens gehört die geografische Suche von Firmenadressen im Internet und auf mobilen Endgeräten.
YellowMap bietet das umfangreichste deutschsprachige Online-Branchenbuch YellowMap.de sowie eine marktführende standardisierte FilialFinder-Technologie für das Internet sowie für mobile Dienste. Umfassende Adressverzeichnisse, weltweite Straßenkarten und innovative Werkzeuge zur Kundengewinnung garantieren die konstante Qualität des Dienstes. Das Geschäftsmodell von YellowMap basiert auf der Zusammenarbeit mit starken Partnern im so genannten Marketingnetzwerk. Zusätzlich ist YellowMap an verschiedenen Forschungsprojekten beteiligt.

Der FilialFinder ist eine Karten- und Routenlösung zur Darstellung von Standorten und Filialen. Er ermöglicht Interessenten, sich mit wenigen Mausklicks über alle regionalen Geschäftsstellen zu informieren und den Weg zu jeder Niederlassung schnell und einfach zu finden.
Der Service basiert auf einer geographischen Umkreissuche, interaktiven Umgebungskarten und einem genauen Anfahrtsrouting mit Wegbeschreibung. Den FilialFinder bietet YellowMap weltweit im Umfang der Microsoft Virtual Earth Karten an. Mit mehr als 200 Kunden ist YellowMap ein marktführender Technologie-Anbieter für Filialsuche und Kundengewinnung im Internet.

Das Branchenbuch ist ein weiteres Standbein von YellowMap. Hier wird mithilfe von Partnern ein Marketingnetzwerk aufgebaut. Alle Unternehmer, Gewerbetreibende und Freiberufler können YellowMap als effiziente Werbeplattform nutzen. Durch preiswerte WerbeEinträge können sie Ihre Produkte und Dienstleistungen ausführlich im Branchenbuch präsentieren.

Die Forschung der YellowMap konzentriert sich auf lösungsverwandte Themen wie Semantic Web, Usability, Content-Context-Match, die Entwicklung mobiler Applikationen und vieles mehr. Um unsere Effizienz weiter zu verbessern, nehmen wir an europäischen und nationalen Forschungsprojekten teil. Die YellowMap kann zahlreiche Erfolge bei der Umsetzung von Forschungsergebnissen in Kundenvorteile aufweisen.
Wir haben immer ein offenes Ohr, wenn es um neue anspruchsvolle Partnerschaften geht.
YellowMap verfolgt das Ziel, kontextsensitive, mobile Internetdienste der nächsten Generation anzubieten. Dazu betreibt YellowMap zusammen mit führenden Unternehmen in Europa Forschungsaktivitäten, um in hohem Maße personalisierbare und intelligente Lösungen zu ermöglichen. Die YellowMap betreibt Forschung im sechsten und siebten Rahmenprogramm der Europäischen Kommission (IST - Information Society Technology). Darüber hinaus haben wir laufende Projekte in Programmen des Bundesministeriums für Wirtschaft und Arbeit (BMWA) und des Bundesministeriums für Bildung und Forschung (BMBF).
Einige dieser Forschungsprojekte werden nun genauer erläutert.[1]

2. Forschungsprojekt Pro Inno

In den letzten Jahrzehnten hat der Pkw-/Lkw-Verkehr um ein Vielfaches zugenommen, wie man im Alltag beobachten kann. Der Ausbau des Straßennetzes konnte nicht im gleichen Maße Schritt halten, so dass in den Ballungsräumen der Verkehr immer zähfließender wird. Das gewachsene Verkehrsaufkommen führt zu einer bedrohlichen Schadstoffbelastung unserer Umwelt. Jüngstes Beispiel für eine Konsequenz ist die Einführung einer Feinstaubplakette seit Anfang des Jahres. Betrachtet man z. B. während der Fahrt auf einer Autobahn die Auslastung der einzelnen Kraftfahrzeuge, so stellt man fest, dass diese meistens nur mit einer Person besetzt sind. Auch im Logistikbereich stellt die Auslastung, z. B. in Form von Leerfahrten, aufgrund gestiegener Treibstoffkosten und umfangreicher gesetzlicher Auflagen bei gleichzeitig zunehmendem europäischem Konkurrenzdruck, ein ernsthaftes Problem dar.
Ziel dieses Forschungsprojektes ist es, zur effektiveren Gestaltung des Individual- und Güterverkehrs, eine den spontanen, flexiblen Nutzungsgewohnheiten und -bedürfnissen gerecht werdende, Kommunikationsplattform zu entwickeln, die innerhalb weniger Sekunden den Anbieter einer Fahrt und den Suchenden zusammenbringt. Damit wird weltweit erstmals eine just-in-time, basierend auf dem aktuellen Kunden- (Mobiltelefon) Standort, Transportvermittlung als mobiler Service über das Handy angeboten. Um eine maximale Mobilität zu erreichen sollen erstmals auch Algorithmen für Teilstrecken und Umsteigemöglichkeiten in das zu berechnende Angebot implementiert werden. Alternative Angebote, z.B. des ÖPNV, sowie Bezahlschnittstellen sollen die automatisierte Transportvermittlung abrunden.

Ziel ist es, umgehend und passgenau an die zeitlichen Bedürfnisse des Kunden angepasst, den Fahrtanbieter und den Fahrtsuchenden zusammenzubringen. Dazu soll der Standort des Teilnehmers durch den Standort des Mobiltelefons in das System einfließen. Dadurch, dass Anbieter und Suchende ständig über das Handy ihren Standort auf Wunsch dem Server mitteilen können, kann kurzfristig ohne größeren Aufwand für die Teilnehmer eine Fahrt vermittelt werden. Dadurch sichert sich der Fahrer zusätzliche Einnahmen und der Fahrtsuchende bekommt mit maximaler Flexibilität ein günstiges Mitfahrangebot.
Im Gegensatz dazu muss bei bestehenden Lösungen, wie z. B. Internet-Mitfahrzentralen oder Kurierdiensten, mindestens 24 bis 48 Stunden vorher geplant werden und die Angebote stehen für spontane, kurzfristige Fahrten nicht zur Verfügung. Mitfahrzentralen bzw. Internetbörsen im Bereich Logistik stellen statische Systeme dar. Spätestens bis einige Stunden vor Abfahrt müssen die notwendigen Absprachen getroffen werden.

3. Forschungsprojekt ModiFrame: ein Framework für mobile Dienste

Zunehmend halten mobile Computer wie Notebooks und Handheld-Rechner Einzug in unseren Alltag. In Zukunft werden diese Geräte wesentlich leistungsfähiger werden und durch mobile Netzwerke ständig mit anderen Geräten oder stationären Netzen verbunden sein.
Die Entwicklung und Markteinführung eines mobilen Mehrwertdatendienstes ist aufgrund der Heterogenität mobiler Netze und Endgeräte sowie der spezifischen Merkmale solcher Dienste eine anspruchsvolle Aufgabe. Gerade für kleine und mittelständische Unternehmen (KMU) ist der entsprechende finanzielle und personelle Aufwand kaum tragbar. Dies ist einer der Gründe für den bis jetzt nur verhaltenen Erfolg mobiler Mehrwertdatendienste in Europa. Um dies zu ändern ist das Ziel von ModiFrame die Vereinfachung der Entwicklung und Bereitstellung mobiler Mehrwertdatendienste. Dazu wird ein Framework für mobile Dienste entwickelt und evaluiert, welches insbesondere für KMU den Betrieb eigener mobiler Dienste ermöglicht, ohne dass diese selbst die komplette dafür benötigte Infrastruktur betreiben müssen. Somit werden KMU in die Lage versetzt, ohne großes Risiko die Möglichkeiten mobiler Dienste zu testen und zu nutzen.

Basierend auf der ModiFrame Plattform kann beispielsweise eine Firma mit Außendienstmitarbeitern Anwendungen entwickeln, welche die Mitarbeiter unterstützen. Aufgrund der leichten Kombinierbarkeit verschiedener Dienste der ModiFrame Plattform muss dabei nicht die komplette Funktionalität selbst entwickelt werden, sondern es können vorhandene Dienste dafür verknüpft werden. Ein Beispiel hierfür ist eine Anwendung, welche auf CRM-Daten auf einem Unternehmensserver zugreift und diese Daten dann per Sprachausgabe oder auf einem mobilen Gerät den Mitarbeitern zur Verfügung stellt. Dazu kann ein CRM-Dienst mit einem Dienst, welcher die Daten an den Benutzer übermittelt, kombiniert werden. Der CRM-Dienst greift auf den Unternehmensserver zu und stellt die Daten zur Verfügung. Diese können dann beispielsweise auf einem mobilem Endgerät per grafischer Oberfläche angezeigt werden. Bei einer reinen Sprachkommunikation kann ein Dienst genutzt werden, der per Spracherkennung die Anfrage erkennt und per Text-to-Speech die Ergebnisse dem Benutzer in natürlicher Sprache mitteilt.

Eine andere mögliche Anwendung ist die vereinfachte Hotelsuche und –buchung für einen Außendienstmitarbeiter. Dieser kann aus dem Auto während der Fahrt zwischen Kundenterminen per Telefon seine aktuelle Position bestimmen lassen. Danach werden ihm alle Hotels in einem bestimmten Umkreis genannt, welche er nach verschiedenen Kriterien, wie z.B. Hotelkategorie oder Preis auswählen kann. Bei dem gewünschten Hotel kann direkt eine Buchung vorgenommen werden. Hier wird ebenfalls bei der Entwicklung der Anwendung auf vorhandene Dienste zurückgegriffen. Zunächst erkennt ein Spracherkennungsdienst die gesprochene Anfrage. Daraufhin wird die Ortung durch einen Lokalisierungsdienst durchgeführt. Die dadurch ermittelten Koordinaten werden für eine geografische Branchensuche genutzt. Ein weiterer Dienst übernimmt dann die eigentliche Hotelreservierung bzw. –buchung.

Mit der ModiFrame Plattform ist ein kleines oder mittleres Unternehmen in der Lage, mobile Anwendungen mit geringem Entwicklungsaufwand selbst zu entwickeln. Damit muss es die Entwicklung nicht auslagern, was mit enormen Kosten verbunden wäre. Außerdem muss sich ein KMU nicht um die Bereitstellung der Dienste bei den jeweiligen Mobilfunkbetreibern kümmern, da auch diese Aufgabe von ModiFrame übernommen werden kann.[2]

4. Forschungsprojekt LOMS - Local Mobile Services

In Zusammenarbeit mit Partnern aus Belgien, Deutschland, Frankreich und Spanien werden im Projekt LOMS Methoden und Tools zur Entwicklung, Durchführung und Nutzung von Context Aware Mobile Services bereit gestellt.
Aufgrund der Allgegenwärtigkeit von IT-Diensten steigen zusehends die Erwartungen der Nutzer hinsichtlich eines einfachen und schnellen Zugangs zu Adhoc-Diensten. Endnutzer konsultieren verschiedenste mobile Dienste je nach aktueller Situation. Jedoch verlangen sie nach einer gewissen Konformität der angebotenen Dienste. Die Vision von LOMS ist die Entwicklung eines Systems, welches selbst den kleinsten Wettbewerber dazu befähigt, wertvolle und ortsbezogene Dienste anzubieten.
Die Zielsetzung von LOMS besteht darin, eine Open-Service-Architektur zu definieren und zu entwickeln, die es ermöglicht, neue innovative, ortsbezogene, mobile Dienste einfach zu entwerfen und zu realisieren. Verfolgt wird vor allem die Herabsetzung der Eintrittsschwelle für kleine und mittelständische Unternehmen sowie nicht gewerblichen Nutzern, vordefinierte mobile Dienste anzubieten. Mit dem Projekt LOMS sollen diese Grenzen überwunden werden, um eine Zunahme an ortsbezogenen, mobilen Internetdiensten zu ermöglichen.
Im Gegensatz zum oben genannten ModiFrame Projekt liegt hierbei der Schwerpunkt auf der Erstellung von geeigneten Rollenmodellen und Plattformen, um die Dienste möglichst effizient zu implementieren.
Ein Labordemonstrator von LOMS wurde mit Hilfe des Partners Suedwest Presse zu lokalen News weiterentwickelt.

Lokale News sind eine Anwendung von mit Zusatzinformationen angereicherter Karten. Konkret werden hierbei an dem jeweiligen Ort, an dem ein Ereignis stattfindet (oder stattgefunden hat) ein Marker in der Karte angezeigt. Beim Selektieren des Markers mit dem Mauszeiger öffnet sich ein Pop-Up Fenster, in dem zusätzliche Informationen angezeigt werden. Diese Informationen können Texte, Tabellen, Bilder oder auch Videos sein.
In Zusammenarbeit mit unserem Partner Suedwest Presse wurde die lokale News Anwendung in www.swp.de integriert. Direkt auf der Startseite erfährt der User Meldungen in und um Ulm - angezeigt und genau platziert auf einer Landkarte. Bei Bedarf kann der User in die Karte bis auf Straßenniveau hinein zoomen. Mit nur einem Klick erhält er die komplette Meldung sowie dazugehörende Bilder. Über ein eigenes Nachrichtensystem pflegen die Redakteure des Nachrichtenportals ganz einfach die lokalen News ein und fügen eine Adresse hinzu, wodurch der Ort des Geschehens lokalisiert werden kann. So ist der Nutzer nicht nur informiert, was gerade in und um Ulm passiert, sondern weiß genau, wo genau sich das Ereignis abspielte.
Die Zusammenarbeit basiert dabei auf einem Rollenmodell, welches ursprünglich aus dem Forschungsprojekt LOMS stammt. Es sieht eine Unterteilung der Zuständigkeiten in einerseits den Contentprovider und andererseits den Plattform- & Dienstanbieter vor. Suedwest Presse übernimmt die Aufgabe des Contentproviders, da sämtliche Inhalte und das Layout beim Partner liegen. YellowMap stellt lediglich die Karte und die zugehörigen JavaScript-Funktionen zur Verfügung.
Aufgrund dieser Trennung kann der Partner Layout und Inhalt der einzelnen Pop-Ups sowie die Icons der jeweiligen Nachrichten frei gestalten. Die Übergabe der Informationen an YellowMap erfolgt dabei per mit GeoRSS versehenem XML. Der Partner erzeugt dazu eine XML-Datei, in der sich die Inhalte und das Layout der Popups sowie ein Verweis auf das für die jeweilige Nachricht zu benutzende Icon befinden. Die GeoRSS Daten werden mit Hilfe eines WebServices von YellowMap erzeugt und in das XML-Dokument eingefügt. Beim Aufruf wird dem Kartenmodul eine oder mehrere XML-Dateien mit Nachrichten übergeben. Diese werden dynamisch geparst und danach dargestellt. Es können dabei auch verschiedene Sortierungen der übergebenen Daten vorgenommen werden.
Die lokalen News sind eine perfekte Ergänzung zum Portfolio von YellowMap. Sie geben zusammen mit dem Branchenbuch unseren Partner die Möglichkeit, ihre Inhalte schnell und übersichtlich online zu präsentieren.

5. Quellen

[1] http://www.yellowmap.com
[2] http://www.modiframe.de
[3] http://www.dloms.org

Ein skalierbares Umgebungsmodell für ortsabhängige Anwendungen

Frank Müller, Steffen Meyer, Stephan Haimerl, Thorsten Vaupel, Kitti Zahonyi

Fraunhofer-Institut für Integrierte Schaltungen IIS
Nordostpark 93
90411 Nürnberg

Abstract

Ziel unserer Arbeit war es, ein Umgebungsmodell zu entwerfen, das die gemeinsamen Anforderungen typischer, ortsabhängiger Anwendungen erfüllt und dabei besonderen Schwerpunkt auf die Verarbeitbarkeit auf mobilen Geräten legt. Die Eigenschaften des Umgebungsmodells umfassen: die Möglichkeit, beliebige logische Einheiten (z. B. Stadtteile, Straßen, Gebäude, Räume, Bereiche) mit ihren geographischen Ausdehnungen sowie Übergängen (z. B. Türen) abzubilden, die wechselseitige Zuordnung von eindeutigen semantischen Bezeichnungen zu geographischen Positionen, eine gute Skalierbarkeit (Modellierung auf Stadtebene bis zur Schublade im Schreibtisch), eine gute Partitionierbarkeit (feingranulare Aufteilung in Teilmodelle), die effiziente Speicherung anwendungsspezifischer, ortsabhängiger Daten, die Möglichkeit der effizienten Wege- und Entfernungsberechnung sowie die Möglichkeit der Visualisierung und Wegführung auf dem mobilen Endgerät.

Der Artikel beschreibt das entwickelte Umgebungsmodell in Struktur und Funktionalität und spiegelt die Leistungsfähigkeit am Beispiel eines Informationssystems für mobile Endgeräte in der Nürnberger Innenstadt auf Basis der WLAN-Lokalisierung wider. Das Informationssystem nutzt dabei neben der Visualisierungsfähigkeit das Modell insbesondere dazu, Daten speichereffizient zu hinterlegen (z. B: Restaurant- und Apothekeninformationen, Referenzdaten für die WLAN-Lokalisierung), Kontextwechsel zu erkennen (Stockwerkswechsel, innen/außen), semantische Positionen zu verarbeiten (z. B. „Eingangsebene des Schürstabhauses“) sowie die Partitionierungsfähigkeit des Modells nach Bedarf (nur benötigte Ausschnitte werden gehalten).

1. Ortsabhängige Anwendungen

Der Erfolg ortsabhängiger Anwendungen hängt maßgeblich von der Verfügbarkeit einer kostengünstigen Lokalisierung für mobile Endgeräte ab. Zwar nimmt die Verbreitung von GPS-Empfängern zu, doch können die Anforderungen vieler, mobiler Anwendungen hinsichtlich der Zeitdauer bis zur ersten Position, der Genauigkeit, der Verfügbarkeit in Gebäuden sowie des Energieverbrauchs nur unzureichend erfüllt werden. Die Nutzung von GSM/UMTS-Funkzellen zur Lokalisierung ermöglicht lediglich eine Grobortung im Außenbereich. Daher ist es nachvollziehbar, dass bisherige ortsabhängige Anwendungen und Dienste sich hauptsächlich auf den Außenbereich beschränken. Neue Ansätze nutzen die weit verbreiteten WLAN-Funkstationen für die Lokalisierung mobiler Endgeräte. Dabei werden die Signalstärken mehrerer Stationen ausgewertet und zur Positionierung genutzt. Bedeutender Vorteil dieses Ansatzes ist die Verfügbarkeit im Innen- und Außenbereich; damit ist der Grundstein für eine neue Generation ortsabhängiger Anwendungen gelegt.
Applikationen für die nahtlose Nutzung kombinierter Innen- und Außenbereiche unterscheiden sich deutlich von reinen Außenanwendungen. Während außerhalb von Gebäuden häufig ein globales, zweidimensionales Koordinatensystem ausreicht, ist innerhalb ein lokales, vorzugsweise kartesisches, metrisches Koordinatensystem intuitiver und performanter. Zudem gewinnt die Höhe als dritte Dimension zur Unterscheidung von Stockwerken große Bedeutung. Zusätzlich sind für viele Anwendungen im Innenbereich beschreibende Informationen (z. B. Ebene, Raumbezeichnung) wichtiger als geometrische Koordinaten. Die Kenntnis über die Struktur der Umgebung ist damit essentiell; auch die Berechnung von Laufwegen und Entfernungen hängt maßgeblich davon ab.
Das im Folgenden vorgestellte Umgebungsmodell ist in Form einer Java-Bibliothek realisiert und stellt neben der reinen Aufnahme- und Strukturierungsfunktionalität geographischer Daten, Basisalgorithmen für verschiedenste Anwendungsbereiche bereit.

2. Struktur des Umgebungsmodels

Den Kernbaustein des Umgebungsmodells bildet die Entität. Eine Entität ist ein reales oder abstraktes Objekt, das eindeutig identifiziert werden kann. Entitäten stehen in einer hierarchischen Beziehung zueinander und bilden Baumstrukturen (siehe Abbildung 1).

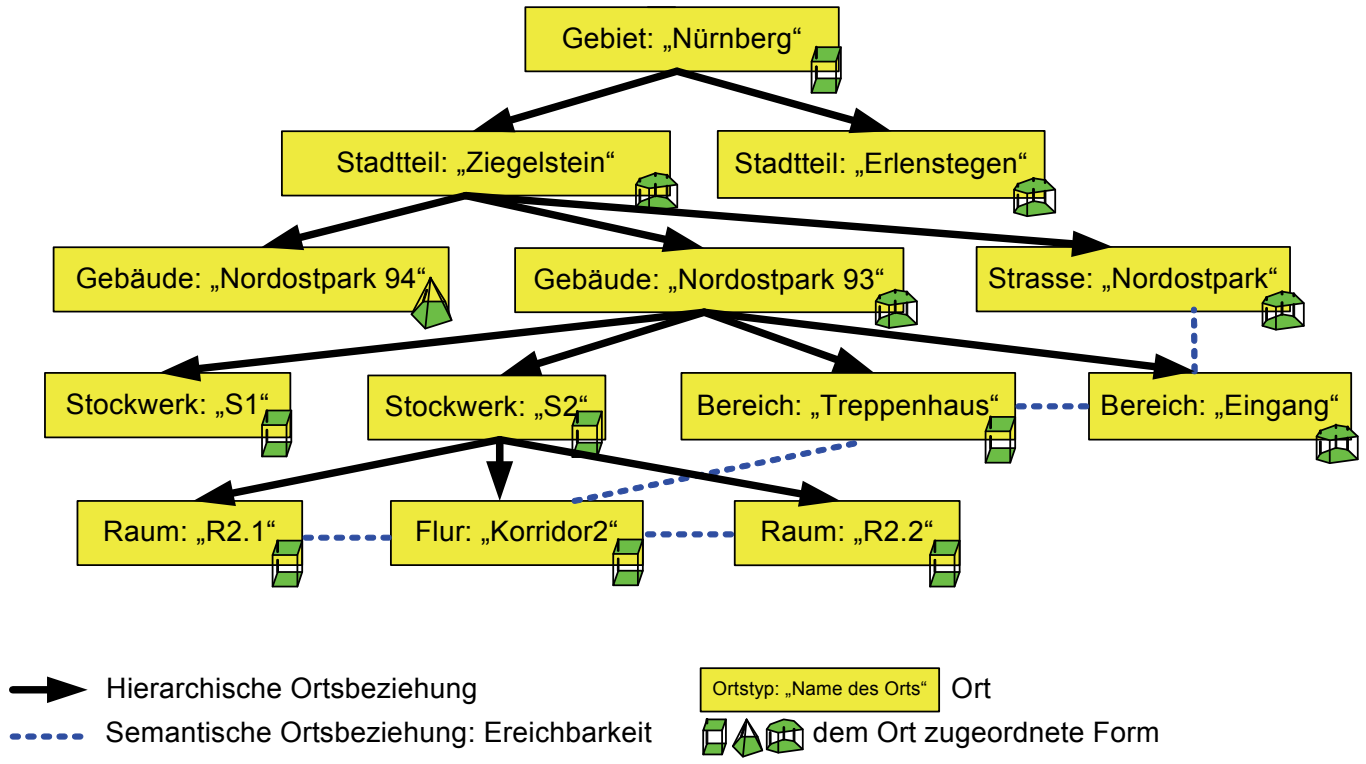


Abbildung 1: Beispiel für eine Ortshierarchie

Ein Ort ist eine spezielle Entität, welche eine geographische Position und Ausdehnung besitzt. Jedem Ort ist eine geometrische Form zugeordnet. Für die Definition dieser Form stehen verschiedene Grundbausteine zur

Verfügung: z. B. Quader, Polyeder oder Pyramide (siehe Abschnitt 2.1). Die Baumstruktur einer Ortshierarchie unterliegt folgenden Regeln:

- „Child-Parent-Containment“-Regel: Die geometrische Form eines eingeschachtelten Ortes muss komplett in dem übergeordneten Ort enthalten sein.
- „Child-Child-Not-Overlapping“-Regel: Die geometrischen Formen von Orten auf gleicher Baumebene dürfen sich nicht überschneiden, jedoch berühren.

Orte existieren in verschieden Ausprägungen, die ihre logische Bedeutung definieren: z. B. Raum, Gebäude, Straße oder Gebiet. Jeder Ort besitzt einen Namen (ShortLocationIdentifier, SLI), welcher bei einer Konkatenation aller Namen (Punkt-getrennt) bis zur Wurzel des Ortsbaumes den Ort eindeutig identifizieren muss (LongLocationIdentifier, LLI), beispielsweise „Nürnberg.Ziegelstein.Nordostpartk93“. An der Wurzel eines Ortsbaums steht immer ein Entität vom Typ: Gebiet. Im Gegensatz zu anderen Ortstypen definiert ein Gebiet ein lokales, kartesisches Koordinatensystem, auf welches die Koordinaten der eigenen Form und der Formen eingeschachtelter Orte bezogen sind. Ein Gebiet wird durch zwei WGS84-Koordinaten global georeferenziert: einem Ursprung (untere linke Ecke) und einem zweiten Punkt auf der positiven x-Achse des lokalen kartesischen Koordinatensystems (siehe Abbildung 2).

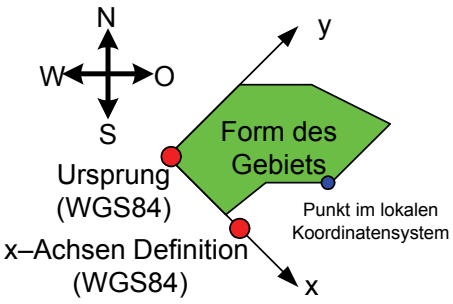


Abbildung 2: Definition von lokalen Koordinatensystemen durch Gebiete

Zusätzlich zu den hierarchischen Beziehungen, können Entitäten (bzw. Orte) optional an beliebig vielen semantischen Beziehungen partizipieren. Für Routingverfahren in Navigationsanwendungen ist es beispielsweise wichtig, die Übergänge und Hindernisse zwischen benachbarten Orten zu kennen. Diese Zusammenhänge werden durch spezielle semantische Beziehungen (Erreichbarkeits- und Hindernisbeziehung) ausgedrückt (siehe Abbildung 1). Die konkreten Übergangs- bzw. Hindernisbereiche lassen sich durch die Angabe von hierarchisch eingeschachtelten Orten genauer spezifizieren.

2.1.Formen

Für die Definition der Form eines Orts stehen mehrere dreidimensionale Grundbausteine zur Verfügung, die sich durch Komposition und Subtraktion zu komplexeren Formen kombinieren lassen. Die Grundbausteine lassen sich konzeptionell in zwei Gruppen einteilen. Zum einen in eine Menge von Formtypen, die durch eine zweidimensionale Grundfläche und zwei z-Koordinaten (obere und unter z-Lage der Grundfläche) definiert sind. Als Grundfläche stehen hierfür folgende Flächen zur Verfügung: Dreieck, Rechteck, Polygon (siehe Abbildung 3).

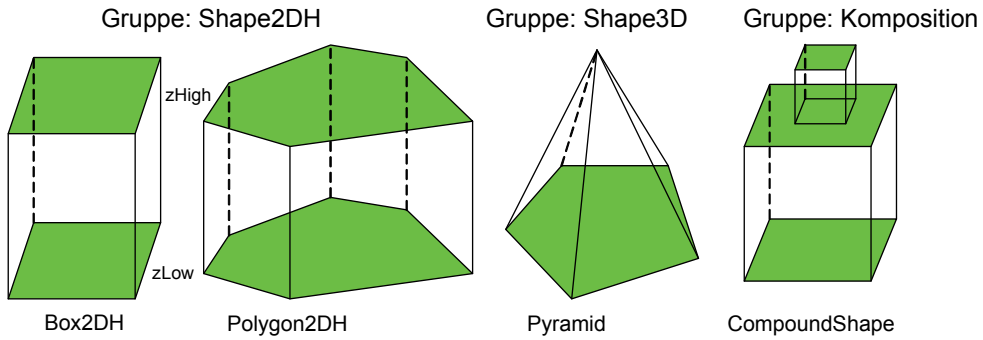


Abbildung 3: Formtypen

Die zweite Gruppe beinhaltet Formtypen, die nicht über eine simple Höhenangabe definiert werden können (z. B. Pyramide oder Polyeder). Über eine Kompositions- oder Subtraktionsform können aus elementaren Formen komplexere Formen erzeugt werden.

3. Kernfunktionalität des Umgebungsmodells

Die Java-Implementierung des Umgebungsmodells stellt einige Basisalgorithmen basierend auf Entitäten, Orten und Formen bereit. Elementar ist die Navigationsfähigkeit und Veränderbarkeit der Baumhierarchie: Für jede Entität ist neben dem Hinzufügen und Entfernen auch die Abfrage eingeschachtelter Entitäten und der umfassenden Entität möglich. Jede Form erlaubt die Überprüfung, ob eine 3D-Koordinate geometrisch in dieser Form enthalten ist (*containsPoint*). Darauf basierend, erlaubt eine Ortshierarchie das Finden des am tiefsten eingeschachtelten Orts, dessen Form eine bestimmte 3D-Koordinate enthält (*findEntity*).
Zwischen zwei beliebigen Formen können die räumlichen Beziehungen Enthaltensein (*contains*), Überlappung (*overlaps*), Berührung (*touches*) und Überscheidung[1] (*intersects*) berechnet werden. Dies ermöglicht die Verifikation erstellter Ortshierarchien (*verifyEntity*) bezüglich der bereits erwähnten „Child-Parent-Containment"- und „Child-Child-Not-Overlapping"-Regeln. Jede Form kann ihre Oberfläche als Menge von planaren 3D-Polygonen berechnen. Bei der Oberflächenberechnung von komplexeren Formen, entstanden durch Komposition und Subtraktion elementarer oder komplexer Formen, werden sich berührende Teilflächen automatisch eliminiert (siehe Abbildung 4). Dies und die weitere Zerlegbarkeit eines planaren 3D-Polygons in konvexe 3D-Polygone oder Dreiecke erleichtert die 3D-Visualisierungsfähigkeit des Umgebungsmodells, da viele 3D-Graphikschnittstellen am Effizientesten mit Dreiecksmengen umgehen können.
Das Umgebungsmodel erlaubt die Hinterlegung beliebiger orts- und applikationsabhängiger Daten; unter einem definierten Inhaltsschlüssel kann für jede Entität ein beliebiges Datenobjekt gespeichert werden. Bei den Inhaltsschlüsseln ist zwischen Hash- und Feldschlüsseln zu unterscheiden, die sich bzgl. Hauptspeicherbedarf und Zugriffsgeschwindigkeit unterscheiden. Ein Feldschlüssel ist zu bevorzugen wenn anzunehmen ist, dass bei einem Großteil der Entitäten ein Datenobjekt unter diesem Schlüssel abgelegt werden wird. Jede Entität reserviert automatisch ein Array mit der Größe der Anzahl der im konkreten Umgebungsmodell definierten Feldschlüssel. Jedem dieser Feldschlüssel ist ein gleich bleibender Index in diesem Array zur Speicherung einer Referenz auf ein Datenobjekt zugeordnet. Dies ermöglicht einen schnellen Datenzugriff, hat jedoch zum Nachteil, dass der Platz für die Referenz reserviert werden muss – unabhängig davon, ob tatsächlich Daten hinterlegt sind oder nicht. Im Gegensatz dazu wird im Umgebungsmodel global pro Hashschlüssel eine eigene Hashtabelle verwaltet, in welcher die Datenobjekte mit der Entität als Schlüssel hinterlegt sind. Die Datenspeicherung mit Hashschlüsseln ist zu bevorzugen falls nur für wenige Entitäten (im Vergleich zur Gesamtmenge) ein Datenobjekt unter diesem Schlüssel gespeichert wird. So muss für nicht existierende Datenobjekte kein zusätzlicher Speicher reserviert werden.

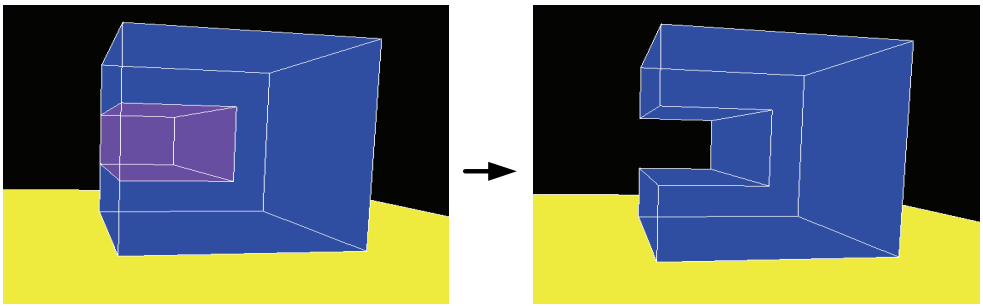


Abbildung 4: Automatische Oberflächenberechnung bei Subtraktionsformen

4. Anwendungsbeispiele

Das Umgebungsmodell ist integraler Bestandteil des vom Fraunhofer IIS entwickelten WLAN-Lokalisierungssytems. Neben der Speicherung der geographischen Daten von modellierten Städten, wird das Umgebungsmodell unter anderem verwendet um Referenzdaten von WLAN-Basisstationen zu vermerken. Daten für die, an einem speziellen Punkt empfangbaren, WLAN-Basisstationen werden unter einem speziellen

[1] Entspricht Überlappung oder Berührung.

Inhaltschlüssel direkt in dem logisch zugehörigen Ort hinterlegt. Um die Möglichkeiten der WLAN-Lokalisierung zu demonstrieren, wurde die Anwendung „Mobile Locator“ entwickelt. Die graphische Oberfläche dieser Software erlaubt die Visualisierung der gegenwärtig ermittelten Position auf Stadt-, Gebäude- oder Stockwerkskarten.

Abbildung 5: Software "Mobile Locator"

Referenzen auf das zugehörige Kartenmaterial werden im zugehörigen Ort unter einem Inhaltsschlüssel hinterlegt. Die Anwendung erlaubt eine ortsgebundene Suche nach Sehenswürdigkeiten oder Restaurants. Sogenannte „Points-Of-Interest (POI)“-Daten sind im Umgebungsmodell als weitere applikationsbezogene Daten über den Inhaltsschlüsselmechanismus hinterlegt.

Infrared-based position determination for augmented cognition utilization in miniaturized traffic test facilities

Andreas Lehner

Thomas Strang

Matthias Kranz

Cristina Rico García

German Aerospace Center DLR

Institute of Communications and Navigation

Muenchner Str. 20

D-82234 Wessling, Germany

Abstract

New infrastructure-less traffic collision avoidance systems are currently developed for road and railway transportation. Similar to maritime and air transport a significant increase in safety is expected by exploiting direct vehicle-to-vehicle communication to detect possible collision threats and to immediately warn the driver. Augmented cognition is a key feature to expand the driver's "view" or to support the self awareness of intelligent onboard units in such critical situations. To design new control and collision avoidance systems and to assess their performance, it is essential to emulate the interaction of centralized traffic management components with partly autonomous vehicles or their drivers. For this purpose test facilities can be operated, e.g. in case of railway transportation using train models to reduce costs and to avoid substantial damage. Concomitant with this approach, a millimetre level localisation of vehicle models is required for small-scale indoor test environments. In this paper we present an infrared based positioning concept that allows simulation of the vehicle's and driver's access to additional traffic relevant information like position or velocity of vehicles in the vicinity.

1. Introduction

So far small-scale traffic control test facilities for railway transportation feature fully equipped control centres, which can be operated by humans in order to investigate and improve safety and economy relevant operation mechanisms. Train models are typically steered from a central unit and their position is detected by sensors in the tracks. Alike, in reality train positions are determined using transponders at decisive points in the railway network.
The historically grown centralized traffic control in real railway transportation still suffers three significant train accidents in Europe every day [1]. The concept of direct vehicle-to-vehicle communication adds here substantial benefit by providing the most relevant traffic information very fast directly to the driver. Moreover in numerous cases such a safety overlay service can prevent collisions with obstacles that cannot be sensed by the control centre, such as construction vehicles, construction workers or pedestrians and vehicles on level crossings, if they are equipped with transceivers. Another important feature in the context of pervasive computing is the general possibility to collect and distribute information about hindrances or damages, but also about delays, weather conditions, or numbers of passengers.
The basis for such an infrastructure-less collision avoidance system in real, as well as in a small-scale model testbed, is the precise position determination of vehicles and the transmission of this information to other vehicles in the vicinity. In this way a joint situation awareness is established and each vehicle is able to detect a collision threat and advise the driver in how to resolve the critical situation. As described in [2], a reliable and accurate position solution for real railway vehicles can be achieved by fusion of multiple sensors, e.g. a GNSS (Global Navigation Satellite System) receiver, an odometer and an eddy current sensor. While GNSS provides reasonable absolute position information with high long term stability, the combination with the short term stable odometer allows very accurate position determination along the track. In order to resolve the track on multi track lines with high reliability the eddy current sensor is used to detect which branch is taken when the train passes a switch.
In the next section we will explain how the functionality of existing miniaturized traffic control test facilities [3] can be expanded to integrate autonomous collision avoidance systems in the individual vehicles. For a close to reality approach, this can be achieved by accurate determination of the position on board of the models and transmission of the data to other intelligent vehicles.

2. System architecture

Standard position determination in test facilities is realized infrastructure based. That means the position information of vehicles is e.g. collected by sensors on the tracks and analyzed by a central control unit. Our concept provides the information onboard the models, as it will be done in future vehicle-to-vehicle communication systems, including rail-based systems.

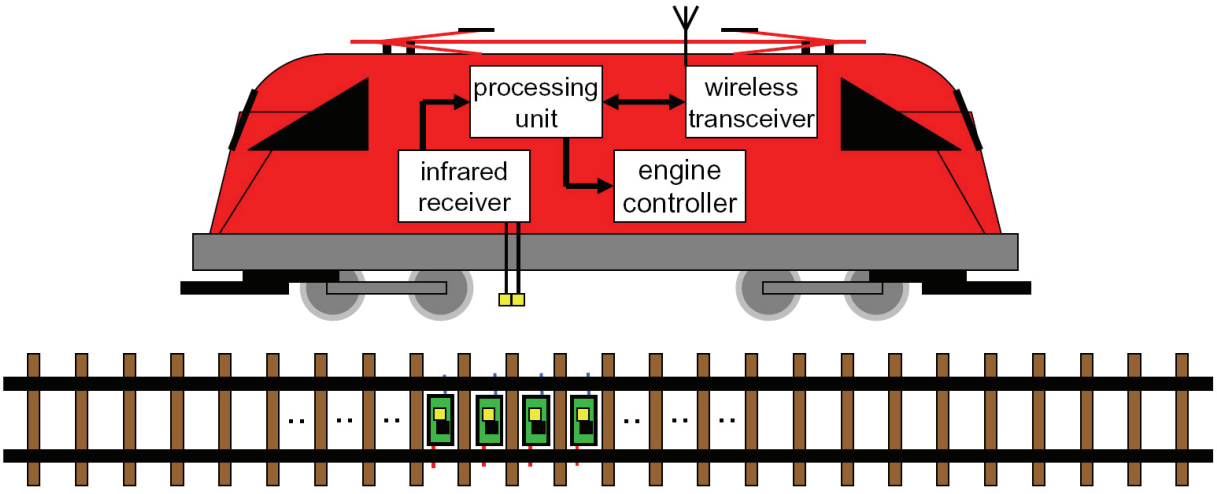


Figure 1.
Railway Collision Avoidance System (RCAS) testbed with a tiny vehicle model acquiring position information through infrared sensors.

Figure 1 illustrates the system setup for our Railway Collision Avoidance System (RCAS) testbed with tiny vehicle models exploiting ad-hoc vehicle-to-vehicle communication. For position determination a series of infrared transmitters is integrated in the tracks. Each transmitter broadcasts a specific address that is assigned to a predefined location on the track map, which is known by the processing unit on the vehicle. To precisely determine position, each vehicle is equipped with a row of infrared sensors. By appropriate selection of the transmitter diode beam width and the distances between transmitters and sensors, respectively, an accuracy down to millimetre range can be achieved. Besides providing the position information directly to the model train, one of the main advantages of this approach is, that this millimetre precision on the small model railways allows emulation of accurate GNSS based positioning that shall be used for the real anti-collision systems in the future. Moreover, as each infrared transmitter address can be mapped to a known WGS84 coordinate on the track map, it is also possible to apply error models for reproduction of characteristic degradation of position accuracy by shadowing effects or multipath. For example, if the model train enters a tunnel, the variance of the error model can be increased when mapping the corresponding addresses to WGS84 coordinates. On the other hand, a very precise rail selective position information can be passed to the train, if a certain address is specified to model a location-transponder on the track.
Another important property of the selected infrared-based positioning solution is the provided absolute position information. In contrast to distance measurement techniques like e.g. wheel sensors, the absolute position information is available under all circumstances, even if a vehicle derails and is placed anywhere else on the test platform.
Existing infrared hardware e.g. by [4] can be adapted to realize the depicted miniaturized positioning system. For cost efficient solutions fully integrated transmitters can be build into the track and can be directly powered through the rails.

In order to run and test collision avoidance algorithms, the position data is read by an onboard processing unit and broadcasted to other vehicles on the test facility using WiFi transceivers. Together with the received information from other vehicles, the processing unit detects potential collision threats, warns the driver or even takes over control to stop the vehicle.
Figure 2 illustrates the collision avoidance test facility with a selected scenario setup for railway transportation. In addition to existing facilities the RCAS system approach is fully reproducible. The collective interaction of the centralized control mechanisms and the onboard collision avoidance strategies, including the human interfaces to the control centre staff and the train drivers, can be tested, analysed and improved. In order to emulate a realistic environment for the train driver, the image of a miniaturized camera (onboard the model) can be visualized in the rebuilt driver's cabin, where he can access all instruments and remotely steers the model.

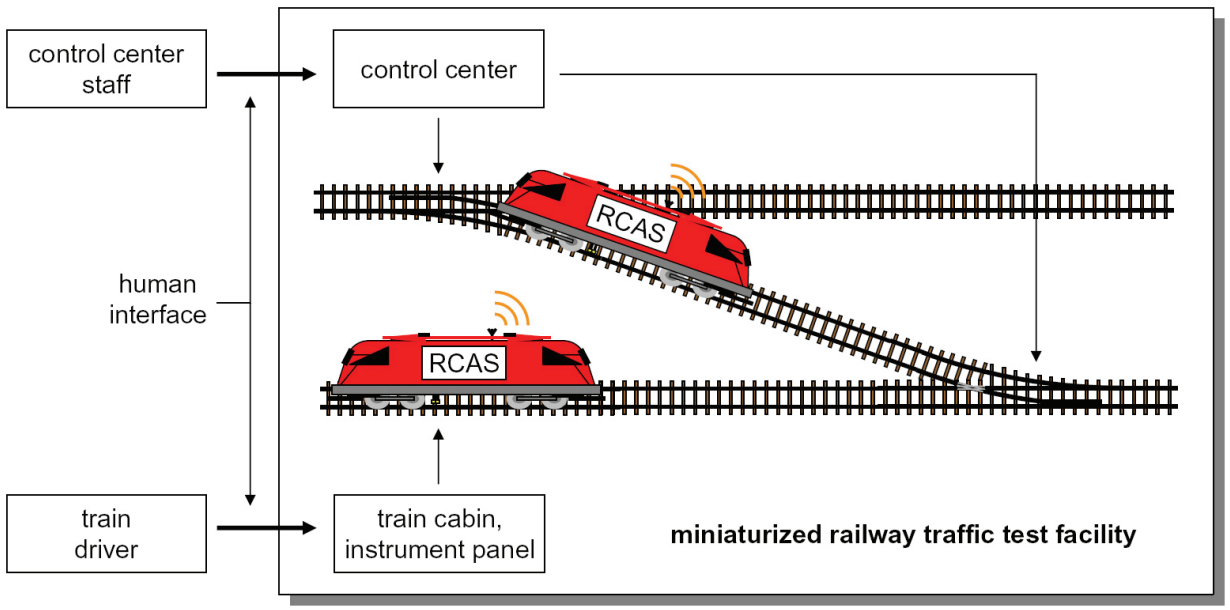


Figure 2.
System architecture of the collision avoidance test facility with a selected scenario setup for railway transportation.

3. Related work

A wide range of positioning technologies is available for various industrial, safety of life, or location based service applications in indoor environments. Just as well different types of sensors are in use to determine the position of vehicles on model railways, but normally this position information is quite rough and it is only provided to the central control unit. Rail occupation sensors or reed contacts only allow replay of railway operation as practised today.
More sophisticated position detection approaches make use of camera systems that track reflective markers on moving objects like the motion capture system from [5]. Similar if velocity information of objects is available in a network and the position of static reflective markers on a test facility are known, these objects can be virtually mapped to the tracks and their location can be further processed or visualized as demonstrated in [6]. Again the positions are determined on the infrastructure side and not on the vehicles' as aimed for.
Wide area inertial-optical motion tracking systems with integrated image processing abilities [7] are typically used in mobile augmented reality applications. Nevertheless, compared to the proposed infrared-based approach the complexity and costs are much higher, and the larger size limits the applicability to large vehicle models. Augmented reality solutions also require steady lightning conditions and do not scale well enough when it comes to test a large number of trains. Occlusion as in tunnels is problematic for vision-based approaches, too.
Advantageous when considering positioning by RFID technology, is the fact that passive tags can be integrated into tracks. On the other hand position accuracy is some orders lower than for the proposed infrared concept, just as it is the case for ultra wide band (UWB) sensors [8], which are limited to decimetres. This would result in a loss of compatibility and thus would not allow transferring results from the model system to real railway systems.

4. Conclusions and outlook

Precise onboard knowledge of position, and the distribution of this information through inter-vehicle communication networks, is the cornerstone to increased safety. At the same time efficiency can be increased by minimizing follow up distances. But also inter-connections to other train control systems and to travel information systems offer new services.
To develop and test new collision avoidance systems, taking interaction with operation centres, controllers and drivers into account, miniaturized test facilities can be equipped with wireless communicating vehicles. A precise, low cost infrared sensor based system can be used to reproduce the great position accuracy of real world multi-sensor-fusion positioning systems. By applying error models different position accuracies can be reproduced to investigate the influence on the system performance.
Currently a test platform is under construction (see Figure 3) which shall demonstrate the functionality of collision avoidance in railway transportation. This miniaturized railway traffic testbed will be operational in autumn 2008 and will allow to test and develop collision avoidance algorithms that augment the train driver's cognition by warning him from dangers he cannot see.

Figure 3.
Platform for the demonstration and test of collision avoidance in railway transportation with an integrated infrared sensor based positioning system.

References

[1] Safety Database Project Team (UIC-SDB), “State of the Art”, The UIC Safety Data Base (UIC-SDB), Paris, 2006

[2] K. Hartwig, M. Grimm, M. Meyer zu Hörste, K. Lemmer, “Requirements for Safety Relevant Positioning Applications in Rail Traffic – a demonstrator for a train borne navigation platform called ″DemoOrt″ ”, National Research Council Canada: 7th World Congress on Railway Research WCRR, Montréal, Canada, 2006

[3] Railway service laboratory (EBL), Technical university of Dresden, http://www.tu-dresden.de/vkivb/ile/home.htm, 2008

[4] Uhlenbrock Electronics GmbH, Bottrop, Germany, “LISSY Locomotive Individual Steering System”, http://www.uhlenbrock.com/3/9/1/I2777817-026.apd/Bes68000e.pdf, 2008

[5] Vicon Motion Systems, http://www.vicon.com/products/viconmx.html, 2008

[6] D. Wagner, T. Pintaric, F. Ledermann and D. Schmalstieg, “Towards Massively Multi-User Augmented Reality on Handheld Devices”, Proceedings of the Third International Conference on Pervasive Computing, Munich, Germany, 2005

[7] E. Foxlin, L. Naimark, “VIS-Tracker: A Wearable Vision-Inertial Self-Tracker”, Proceedings of the Virtual Reality Conference, Los Angeles, USA, 2003

[8] Ubisense ultra wide band (UWB) positioning technology, http://www.ubisense.net/, 2008

LBS_2.0 - Realisierung von Location Based Services mit user-generated, collaborative erhobenen freien Geodata

Pascal Neis und Alexander Zipf

Lehrstuhl Kartographie, Department of Geography, University of Bonn, Germany

http://www.geographie.uni-bonn.de/karto

http://www.OpenRouteService.org

1. Einleitung

"There will, if necessary, be a grass-roots remapping." Tim Berners-Lee

Von Benutzern freiwillig beigesteuerte Inhalte sind der Grundpfeiler erfolgreicher Anwendungen im Web 2.0. Ein Paradebeispiel ist die Online-Enzyklopädie-Wikipedia. Trotz (oder gerade wegen?) ihrer offenen Architektur (jeder Benutzer darf jeden Artikel ändern), zeichnet sie sich durch hohe Qualität und Aktualität der enthaltenen Informationen aus. Auch Geoinformationen können, insbesondere dank GPS-Technologie, durch die freiwillige Kollaboration vieler Individuen generiert werden – Volunteered Geography. Ein prominentes Beispiel ist das OpenStreetMap-Projekt (OSM). Hier entsteht seit 2004 eine frei verfügbare, von jedermann editierbare Weltkarte, die besonders in einigen deutschen Städten bereits höheren Detailreichtum (Fußwege, Radwege) bietet als amtliche oder kommerzielle Kartenwerke.

2. OpenRouteService.org

Vor kurzem wurde mit OpenRouteService.org von Neis (2008) ein zunächst deutschlandweiter Routenplaner vorgestellt, der einerseits komplett auf den offenen Standards des Open Geospatial Consortiums (OGC) basiert und als Datenbasis die freien Geodaten von OpenStreetMap - der feien Wiki-Weltkarte - nutzt. OpenRouteService ist aber mehr als ein einfacher Routenplaner, da er mehrere Dienste der OGC Open Location Services Initiative beinhaltet, die Funktionen wie Adresssuche, Gelbe -Seiten-Dienst (Umgebungssuche). Insbesondere bietet er damit echte LBS-Funktionalität auf Basis von user generated content sowohl für Karten und Routing als auch POI-Suche (Directory Service). Daneben wurden auf diesen Basisdiensten weiterführende Dienste realisiert, wie z.B. einen Erreichbarkeitsdienst, der das in einer gegebenen Zeit erreichbare Gebiet ausgibt.

In den letzten Wochen wurden zahlreiche Erweiterungen realisiert und neben der Version für Deutschland auch eine Variante für das Urspungsland von OpenStreetMap das Vereinigte Königreich von Großbritanien umgesetzt. Eine weitergehende räumliche Ausdehnung auf weitere Staaten erfolgt in Kürze. Neben der Web-GUI unter http://www.openrouteservice.org wurden auch schon erste mobile Clients für Smartphones / PDAs mit Web-Browser oder über Java Midlet-Technologie entwickelt. Im Rahmen eines Google Summer –Of-code-Projektes entstand zudem ein erster Client für die Google Android Plattform.

Im Moment bietet OpenRouteService.org die folgenden Basisdienste:

- OpenLS Directory Service – Ortsbezogene "Gelbe Seiten" - Suche (Online-Branchenverzeichnis), d.h Umkreissuche nach Restaurants, Geschäften, Hotels, Parkplätzen etc.
- Accessibility Analysis Service (AAS) - Berechnung eines vom ausgewählten Punkt aus in einer vorzugebenden Zeit erreichbares Gebiet als Polygon, auf Basis des Straßennetzes
- OpenLS Location Utility Service – Geocoding & Reverse Geocoding, d.h. Konvertierung von Adressen (auch Freiform) in Geometrien und umgekehrt
- OpenLS Presentation Service – Karten mit eingezeichneten Routen, POIs etc. (Z.Zt. nur interne Nutzung)
- OpenLS Route Service – Routenplanung auf Netzwerkgraphen nach diversen Kriterien
    - Autofahrer: fastest
    - Autofaher: shortest
    - Radfahrer
    - Fußgänger
    - Erweiterungen sind in Arbeit
- ein OGC WMS (Web Map Service) mit SLD Unterstützung wurde aufgesetzt und kann bei Bedarf eingebunden werden.
- intern wird ein OGC WFS (Web Feature Service) verwendet, in dem die Daten von OpenStreetMap thematisch aufbereitet und strukturiert vorliegen. Somit könnten die Daten von OpenStreetMap auch per GML ausgeliefert werden.
- Die Integration der Daten in einen OGC Katalogdienst (CS-W, Catalogue Service Web) ist in Arbeit.

Neben den normalen Fahranweisungen können mittlerweile auch erweiterte Sprachanweisungen ausgegeben werden (Auswahl: 'de*' oder 'en*'). Dabei werden in der Fahranweisung zusätzliche Objekte, die an Kreuzungen vorhanden sind, erwähnt, z.B. Ampeln oder kleine Kreisel. Dies wird zurzeit erweitert um Objekte, die in der näheren Umgebung von Entscheidungspunkten, z.B. Kreuzungen, liegen. Wenn also eine weitere Sprachanweisung nötig wird werden zusätzliche POIs (Points of Interest) als Art Landmarke in die Navigationsanweisung eingebunden.

Des Weiteren ist der Download der berechneten Route als GPX-Datei möglich (intern auch KML). Intern sind noch weitere Optionen realisiert (wie z.B. Zwischenpunkte beim Routing setzen, erweiterte Ausgabemöglichkeiten bei der Erreichbarkeitsanalyse etc.), die noch nicht in der Benutzeroberfläche sichtbar sind. Als nächste Funktionen werden Höhenprofile der berechneten Touren integriert, sowie an der Berücksichtigung dynamischer TMC Informationen (Staus, Baustellen etc.) gearbeitet. Die folgende Abbildung 1 zeigt die Architektur von OpenRouteService.org.

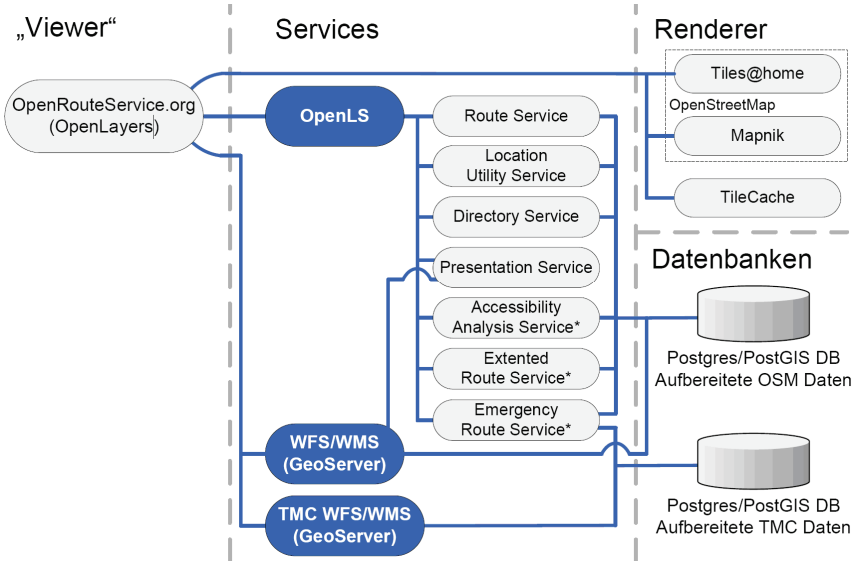


Abbildung 1: Service-Architektur von OpenRouteService.org

OpenRouteService wird laufend weiterentwickelt (z.B. Routenabhängiges Höhenprofil auf Basis von SRTM-Geländemodellen) und wurde schon erfolgreich in diverse Produkte und Open Source Projekte integriert. Der Route Service wurde auch schon mit Straßendaten kommerzieller oder öffentlicher Anbieter betrieben.

3. Fallbeispiel Katastrophenmanagment

Am Beispiel einer vor kurzem realisierten Erweiterung soll der Vorteil der Nutzung offener Daten und offener Standards am Beispiel von OpenRouteService veranschaulicht werden.

Nachdem Hurrikan „Ike“ Haiti verwüstet hat und über hundert Menschen getötet und zehntausende obdachlos wurden, ist die Lage in Haiti auch Wochen nach dem Hurrikan weiter angespannt. Durch Überflutungen und Zerstörung zahlreicher Brücken und Straßen ist der durch die UN organisierte Hilfseinsatz für die über 650.000 betroffenen Menschen weiter äußerst schwierig. Um eine Hungersnot und Epidemien zu verhindern, müssen Lebensmittel, Medikamente und weitere Güter zu den Betroffenen gebracht werden und der Wiederaufbau der Infrastruktur organisiert werden. Im Rahmen des vom GLCSC (Global Logistics Cluster Support Cell) koordinierten Einsatzes sind mehrere UN Organisationen (OCHA, WFP, WHO, UNICEF) als auch NGO's (World Vision, MSF) aktiv beteiligt.

Das UN Joint Logistics Center (UNJLC) – eine von der World Food Program (WFP) getragenen Einheit – ist dabei für die Themengebiete GIS und Transport/Logistik zuständig und koordiniert die Aktivitäten diesbezüglich. Dabei definiert und realisiert UNJLC mit UN SDI-T (T für Transportation) den für Transport zuständigen Teilbereich der im Aufbau befindlichen United Nations Spatial Data Infrastructure (UNSDI).
Für den aktuellen Einsatz relevant sind dabei insbesondere aktuelle Straßenzustandskarten und Informationen über Hindernisse, Gefahrengebiete und die sonstige Infrastruktur. Mehrere NGOs wie z.B. CartONG.org unterstützen bei der Erhebung dieser Daten, analysieren sie und stellen die Ergebnisse den beteiligten Hilfsorganisationen zur Verfügung. Im Rahmen dieser Bemühungen wurde der Lehrstuhl Kartographie des Geographischen Instituts der Universität Bonn von Seiten des UNJLC angefragt, ob eine Unterstützung bei der Realisierung eines dringend benötigten Routenplaners, der die aktuelle Straßensituation auf Haiti berücksichtigt, möglich ist.

In einer spontanen Hilfsaktion mehrerer freiwilliger Mitarbeiter des Lehrstuhls wurde daher der bestehende Routing Service „OpenRouteService.org“ (ORS) funktional und bzgl. der Daten erweitert und umgeschrieben. So konnte innerhalb weniger Tage eine erste spezielle Version zur Unterstützung der Katastrophenhilfe in Haiti veröffentlicht werden. Dieser wurde dem UNJLC zur Verfügung gestellt und wird nun von diesem und den beteiligten Einsatzkräften zur Optimierung der Hilfslogistik genutzt.

Eine wesentliche Hürde dabei war zunächst die Verfügbarkeit eines geeigneten Datensatzes des Verkehrsnetzes. Während die freie Wiki-Weltkarte OpenStreetMap (OSM) in vielen europäischen Ballungszentren schon eine außerordentlich gute Abdeckung aufweist und auch in verschiedenen Städten in Entwicklungsländern teilweise schon eine erstaunliche Qualität bietet, ist die Situation in dünner besiedelten Gegenden oft noch unzureichend. In Haiti waren die bislang durch OpenStreetMap abgedeckten Datensätze äußerst spärlich, so dass kurzfristig eine Alternative gefunden werden musste. Grundsätzlich bestand seitens des UNJLC aber schon die Idee die Infrastruktur von OpenStreetMap zu nutzen, um die laufende Aktualisierung und Verbesserung der Geodaten von Haiti schnell und unkompliziert umsetzen zu können. Bedingt durch die freie Lizenz von OpenStreetMap dürfen aber nur „freie“ Geodaten in die OSM-Datenbank eingespielt werden. Ein derartiger Datensatz stand zunächst nicht zur Verfügung. Stattdessen wurde nach Tests mit verschiedenen Alternativen, die zum Teil mangels ausreichender Topologie nicht routingfähig waren, ein Straßennetz einer Umweltorganisation ohne Umweg über OSM direkt per Hand in die Datenbank von OpenRouteService geladen. Zuvor mussten noch einige Attribute angepasst werden. Mittels dieses Datensatzes war wenigstens ein grundsätzliches Routing auf Basis der Dienste von OpenRouteService in Haiti möglich. Dieser Datensatz muss nun regelmäßig – voraussichtlich täglich - aktualisiert werden, um den wechselnden Straßenzustand abzubilden. Sobald wie möglich soll dieses Verfahren durch eine Variante auf Basis von OpenStreetMap abgelöst werden. Hierzu wird ein freier Straßendatensatz erstellt und soll bald über die Infrastruktur von OpenStreetMap zur Verfügung stehen. Dieser kann dann von jedermann über die bekannten Werkzeuge von OpenStreetMap verbessert und für eigene Anwendungen genutzt werden. Dies ist besonders für professionelle Anwendungen wie das Katastrophenmanagement eine wichtige Voraussetzung und macht OSM so attraktiv im Gegensatz zu auf Google Maps oder Earth basierenden Ansätzen, bei denen die Karten nur als Hintergrundinformation dienen, aber sonst nicht wirklich genutzt werden können.

Zusätzlich wurde für die Übergangszeit bis ein geeigneter freier Datensatz über OpenStreetMap verfügbar ist, ein weiterer Datensatz mit Ortschaften in die OpenRouteService-Datenbank für Haiti geladen. Damit konnten weitere Funktionen wie z.B. eine Ortssuche realisiert werden. Mit beiden Datensätzen wurde zudem ein erster einfacher Web Map Service (OGC WMS) unter Verwendung von Styled Layer Descriptors (OGC SLD) für das Styling umgesetzt. Dieser WMS wird in der ORS-Version der Haiti-Hilfsaktion sowieso benötigt, da die üblichen Renderer von OpenStreetMap wie Mapnik neu eingespielte Geodaten in der Regel nicht direkt in ihrer Karte anzeigen. Hier wird aber eine mindestens tägliche Aktualisierung benötigt. Die Aktualisierung der Grunddaten und ihrer Attribute geschieht durch die Mitarbeiter der Hilfsorganisationen. Die resultierenden neuen Geodaten werden dann täglich seitens des Lehrstuhls für Kartographie in die jeweiligen eigenen Datenbanken für die relevanten Dienste (routing, geocoding, map service) übernommen.

Für die Hilfsaktion in Haiti ist eine Fähigkeit von OpenRouteService.org besonders relevant: als gesperrt definierte Gebiete oder Straßenabschnitte können beim Routing berücksichtigt, also umfahren Gebiete werden. Die Grundfunktion hierfür ist schon im verwendeten Standard, dem OGC Location Services Route Service, über die Unterstützung sogenannter „AvoidAreas“ angelegt. OpenRouteService.org bietet die Nutzung derselben über zwei alternative Wege an: Der erste ist die - auch über die Bedienoberfläche von OpenRouteService.org verfügbare - Option als Benutzer selbst ständig interaktiv in die Karte die zu umfahrenden „AvoidArea“ als Polygone einzuzeichnen. Diese werden beim folgenden Routing dieses Nutzers berücksichtigt, sind aber für andere Nutzer nicht sichtbar, da sie nur im Client des jeweiligen Anwenders verfügbar sind. Daher wurde die zweite Option geschaffen: Mitarbeiter der Hilfsorganisationen können derartige AvoidAreas selbst über die Web-Oberfläche als Geodatensatz in eine vom Lehrstuhl Kartographie bereitgestellte Geodatenbank einpflegen. Diese Gebiete stehen dann allen Nutzern des Dienstes zur Verfügung und können nach Aktivieren einer entsprechenden Option beim Routing berücksichtigt werden. Diese Variante mit einer zentralen AvoidArea-Geodatenbank wird in ähnlicher Weise schon im normalen OpenRouteService für die Berücksichtigung von Stau- und Baustelleninfos über RDS-TMC genutzt (prototypisch für NRW und Bayern). Für den Haiti-Routing-Dienst werden zusätzlich spezielle Attribute unterstützt. Zukünftig sollen weitere sinnvolle Funktionen für die Webanwendung hinzukommen, um diese Informationen auch leicht auf der Karte annotieren und darstellen zu können. Sicherlich sind noch zahlreiche Verbesserungen und Ergänzungen nötig und sinnvoll. Erste Rückmel-

dungen aus Haiti bestätigen aber, dass der Dienst schon in dieser ersten prototypischen Ausbaustufe eine wertvolle Hilfe für die Arbeit vor Ort darstellt.

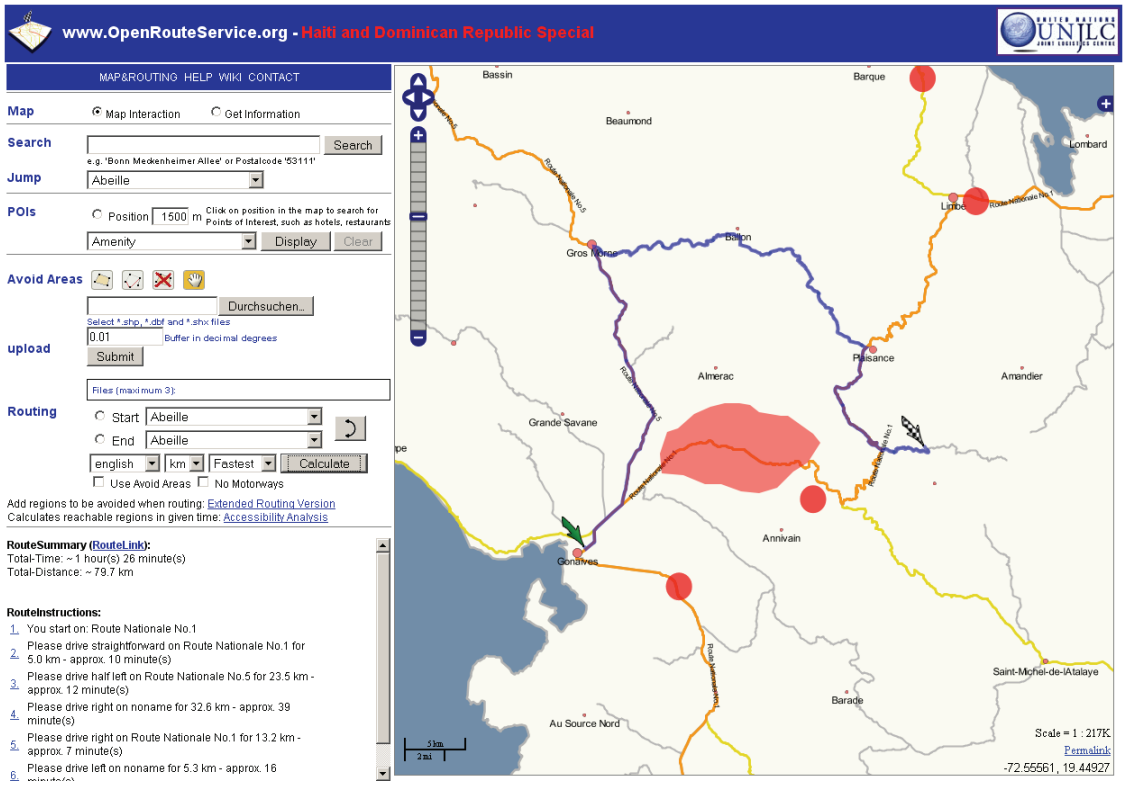


Abbildung 2: Spezielle Version des OpenRouteService für Haiti-Einsatz des UNJLC

4. Zusammenfassung und Ausblick

Der ORS-Haiti verdeutlicht an einem am Ernstfall erprobten Realweltbeispiel, dass auf offenen Daten und offenen Standards basierende Webdienste wie OpenRouteService.org einen sinnvollen Ansatz für die Realisierung von Geoinformationsanwendungen darstellen können - insbesondere auch für das Katastrophenmanagement. Die Einbeziehung der weltweiten OpenStreetMap-Community und deren Infrastruktur für das Erheben und Aktualisieren von Geodaten und der so gewonnenen zusätzlichen aktuellen Datensätze führen zu einer neuen und offenen Prosumer-orientierten GDI 2.0. Diese ergänzt die klassischen Ansätze in innovativer Weise. Somit kann die Kreativität des GeoWeb2.0 auch für ernsthafte Anwendungen wie das Katastrophenmanagement nutzbar gemacht werden. Andererseits wird gerade hier auch weiterhin auf amtliche Geodaten – und GDI-Dienste zurückgegriffen werden. Dies wird durch die zunehmende Standardisierung im Rahmen der Realisierung von GDIs immer einfacher. Da OpenRouteService sowohl die GDI-Standards des OGC nutzt, als auch einen Prototypen für einen Mehrwertdienst auf Basis von durch Prosumer erzeugten Geodaten darstellt, bietet er eine zweiseitig interoperable Brücke zwischen diesen bis vor kurzem noch relativ separaten Welten.

5. Zitierte und weiterführende Literatur:

- Coast S. (2007). OpenStreetMap. Workshop on Volunteered Geographic Information, December 13-14, 2007 http://www.ncgia.ucsb.edu/projects/vgi/
- Neis, P. (2008): Location Based Services mit OpenStreetMap Daten. Masterarbeit. FH Mainz. Betreuer A. Zipf.
- OpenLS: OGC Open Location Services Version 1.2 http://www.opengeospatial.org/standards/ols
- OpenStreetMap: www.openstreetmap.org
- ORS – OpenRouteService.org

Data Mining und natürlichsprachliche Verbalmorphologien

Prof. Dr. Alfred Holl

Gordon Zimnik, B. Sc.

Georg-Simon-Ohm-Hochschule Nürnberg
Fakultät Informatik

Abstract

Die Analyse von Daten mit dem Ziel der Mustererkennung ist vor allem in der Betriebswirtschaft weit verbreitet (z.B. für Verkaufsdaten). Aber auch Sprachwissenschaftler führen Datenanalysen durch, wenn sie grammatische Regeln aufstellen. Die Informatik kennt unter dem Namen Data Mining viele verschiedene Vorgehensweisen zur Datenanalyse, darunter Clusteranalyse-Methoden. Ein solcher Algorithmus kann in der Linguistik einzelsprachunabhängig für die Analyse von flexionsmorphologischen Systemen, meist Verbalsystemen, eingesetzt werden. Der jeweils entstehende Regelapparat in Gestalt eines Verbregisters besteht aus morphologisch homogenen Clustern, die rückläufig ähnliche, morphologisch analoge Vertreter der Wortart Verb enthalten.
Diese Regeln liefern sprachdidaktisch verwertbare, quantitative und strukturelle Aussagen über flexionsmorphologische Systeme. Eine wichtige Anwendung ist es, ein beliebiges Verb automatisch seinem Cluster zuzuordnen, d.h. die Konjugationsklasse des Verbs zu ermitteln. Einen dafür geeigneten Algorithmus zu entwickeln, bildet den Gegenstand dieser Untersuchung. Die Entwicklung geschieht linguistisch motiviert, mit englischen Beispielen illustriert und informationstechnisch dokumentiert. Dabei werden Prinzipien des Software Engineerings im Rahmen eines Phasenkonzepts konsequent eingehalten.
Diese Untersuchung ist die Erweiterung einer Bachelorarbeit an der Fakultät Informatik im SoSe 2008. Sie steht in der mehrjährigen Tradition von Forschungsprojekten und Publikationen zum Thema „Data Mining flexionsmorphologischer Systeme“ an der Ohm-Hochschule. Angewendet auf die Verbalsysteme verschiedener Schulsprachen, wurden die Resultate dieser Studie erstmals anlässlich der Technikmeile Ende Juli 2008 einer breiten Öffentlichkeit vorgestellt.

Inhaltsverzeichnis

1. Einleitung

In dieser Untersuchung wird Data Mining als Teildisziplin der Informatik mit der linguistischen Analyse flexionsmorphologischer Systeme in natürlichen Sprachen verbunden. Datenanalyse in der Linguistik ist nichts grundsätzlich Neues. Jeder Linguist, der morphologische, syntaktische, phonologische oder ähnliche Regeln aufstellen will, führt eine Datenanalyse durch, wenn er linguistisches Material (Texte, Grammatiken, Wörterbücher) analysiert. Das gilt insbesondere für morphologische Systeme. Jede bestehende Liste unregelmäßiger Verben ist beispielsweise aus einem Datenanalyseprozess hervorgegangen.
Diese Untersuchung nimmt durch Data Mining ermittelte Register flexionsmorphologischer Regellexeme für jeweils eine Wortart einer natürlichen Sprache zum Ausgangspunkt. Ziel ist es, einen Algorithmus zu entwickeln, der zu einem beliebigen Suchlexem dessen Regellexem bzw. Regellexeme liefert. Die hier verwendete Terminologie wird in 3.2 genau definiert.
Kapitel 2 beschreibt die Grundlagen und Zielsetzungen, welche die Basis für die Konzipierung und spätere Implementierung des Algorithmus bilden, und erklärt die dahinter stehende Forschungsintention. In Kapitel 3 und 4 wird ein strukturierter Software-Entwicklungsprozess verfolgt (Abb. 1.01). Kapitel 3 zeigt schrittweise die Konzeption des Algorithmus. In Kapitel 4 wird die technische Realisierung des Algorithmus im Detail erklärt. Kapitel 5 zeigt die Möglichkeiten und Einsatzgebiete dieser Forschungsergebnisse.

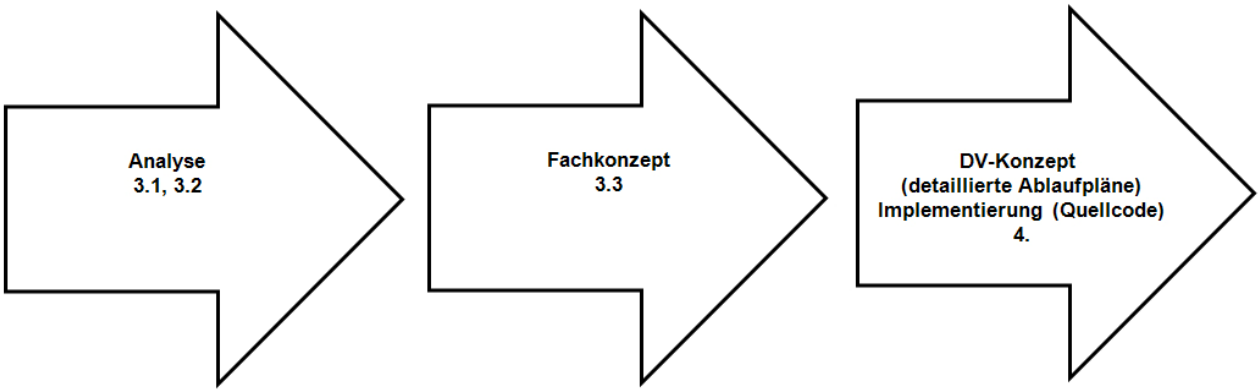


Abbildung 1.01: Phasen des Software-Entwicklungsprozesses
(vgl. Mayr / Hess, 2008, 381, Abb. 4)

2. Grundlagen und Zielsetzungen

In 2.1 wird die Historie dieser Art von Sprachforschung und deren Zielverfolgung erörtert. 2.2 verdeutlicht die über Jahrzehnte reichende Forschungsintention von Alfred Holl.

2.1 Historie

Die ersten Vorarbeiten zu dieser Thematik begannen mit einer Dissertation über Verbalsysteme in der Latein-Romania (Holl 1988), damals noch mit der manuellen Durchführung formalisierbarer Analysealgorithmen. Durch die Finanzierung der bayerischen HighTech-Offensive und der Staedtler-Stiftung Nürnberg war es an der Fakultät Informatik der Georg-Simon-Ohm-Hochschule Nürnberg möglich, diesen Forschungsansatz in mehreren Teilprojekten bis heute weiterzuführen (vgl. 3.1).

2.2 Grundidee und Motivation[1]

Zur Analyse morphologischer Systeme werden Verfahren des Data Mining verwendet. „Data Mining bedeutet buchstäblich Schürfen oder Graben in Daten ... Die Ergebnisse lassen Muster in Daten erkennen, weswegen Data Mining auch als Datenmustererkennung übersetzt wird“ (Alpar / Niedereichholz, 2000, 3). Man sucht nach nicht bekannten, versteckten Zusammenhängen und Ähnlichkeiten, deren Kenntnis einen wirtschaftlichen oder wissenschaftlichen Nutzen verspricht. Für die Disziplin der Datenanalyse finden sich statt des Ausdrucks „Data Mining“ – entsprechend der bekannten Instabilität der Terminologie in der Informatik – weitere, wie „Information Mining“ oder „Knowledge Discovery in Databases (KDD)“, ohne dass eine genaue definitorische Abgrenzung möglich wäre. Der gesamte Bereich umfasst heute eine Vielzahl unterschiedlicher, teils statistischer Methoden, die verbreitet zur Analyse großer Datenbestände eingesetzt werden, etwa im Marketing, um Regelmäßigkeiten im Kundenverhalten festzustellen oder um aus potentiellen Kunden besonders viel versprechende herauszufiltern.

Die im Rahmen dieser Forschungsarbeit zu verwendenden Datenanalyse-Verfahren gehören zu der variantenreichen Gruppe der Clusteranalyse-Algorithmen. Unter Clustern versteht man in diesem Zusammenhang ganz einfach Mengen von Datensätzen. Ein Datensatz ist ein Tupel (eine Aneinanderreihung) zusammengehöriger Einzeldaten, im vorliegenden Zusammenhang die Schlüsselformen eines Lexems[2], etwa die Stammreihe eines Verbs (im Deutschen z.B. *gehen, ging, gegangen*) mit dem Verweis auf ein Musterparadigma, ggf. mit Angabe der Bedeutung, falls der Konjugationstyp von ihr abhängt (z.B. bei *schleifen*, *schliff*, *geschliffen* ‘schärfen‘ vs. *schleifen*, *schleifte*, *geschleift* ‘zerstören‘). Ziel einer Clusteranalyse ist es immer, Cluster mit möglichst ähnlichen Elementen (Datensätzen) zu ermitteln, mit anderen Worten „Daten (semi-)automatisch so in Kategorien, Klassen oder Gruppen (Cluster) einzuteilen, dass Objekte im gleichen Cluster möglichst ähnlich und Objekte aus verschiedenen Clustern möglichst unähnlich zueinander sind“ (Ester / Sander, 2000, 45). Ähnlichkeit ist den jeweiligen Anforderungen entsprechend zu definieren.

Auch bisherige ausführliche Darstellungen von Verbalsystemen sind Produkte „manueller“ Clusteranalyse. Wird etwa zu jedem Musterverb eine Liste aller Verben mit den gleichen Unregelmäßigkeiten angegeben oder in einem Gesamtverzeichnis aller Verben zu jedem das Sigel des zugehörigen Musterverbs, so werden – im Sprachgebrauch des Data Mining – Cluster aus jeweils morphologisch analogen Verben gebildet. Morphologische Analogie bedeutet das Vorhandensein der gleichen morphologischen Eigenschaften, bei Verben also der gleichen Konjugationsmerkmale (Stammalternanzen und Personalendungen). Doch was nützen derartige Cluster unter einem sprachdidaktischen Blickwinkel? Sie sind so umfangreich und unstrukturiert, dass es unmöglich ist, sie auswendig zu lernen.

Sprachstudierende gehen in ihren Lernprozessen daher andere Wege. Sie wollen und müssen ihren Lernaufwand minimieren und würden am liebsten an der lexikalischen Grundform eines Verbs, häufig dem Infinitiv Präsens Aktiv[3], auch dessen Unregelmäßigkeiten ablesen können, so wie man tatsächlich den Konjugationstyp regelmäßiger Verben in romanischen Sprachen an der Infinitivendung erkennt. Ist nun aber mühselig ein Averbo[4] gelernt, so versucht der Lernende in einer ersten Stufe, aus diesem Wissen größtmögliches Kapital zu

[1] Grundidee und Motivation wurden aus (Holl, 2006, 14-18) übernommen.

[2] Wir verwenden den Terminus *Lexem* als Bezeichnung für ein „Wort“ in der Form, wie es in einem Lexikon eingetragen ist. Ein Lexem ist eine abstrakte Basiseinheit des Lexikons, die in verschiedenen Flexionsformen auftreten kann. Es wird durch seine lexikalische Grundform (Lemma) repräsentiert; bei Verben ist dies meist der Infinitiv Präsens Aktiv.

[3] Um die Flexionsformbestimmungen handlicher zu gestalten, wird die Angabe „*Person*“ stets weggelassen. Bei der Diathese wird „*Passiv*“ stets genannt, „*Aktiv*“ nicht immer. Statt „*Infinitiv Präsens Aktiv*“ sagen wir kurz „*Infinitiv*“ oder „*Infinitiv Präsens*“.

[4] Unter dem *Averbo* eines flektierbaren Lexems versteht man die geordnete Menge seiner Flexionsformen.

schlagen und es auf weitere „ähnliche“ Verben anzuwenden, die er dann zu den gleichen morphologischen Eigenschaften „verurteilt“. Als tertium comparationis[5] zur „Feststellung“ der Ähnlichkeit wird die rückläufige Ähnlichkeit gewählt.

Die rückläufige Ähnlichkeit (Ausgangsgleichheit) manifestiert sich in einem gemeinsamen Ausgang in der lexikalischen Grundform, wobei hier der Terminus „Ausgang“ nicht als morphologische Kategorie im Sinne von „Endung“ verstanden wird, sondern ganz einfach als Buchstabenfolge am Ende eines Wortes, deren Länge jeweils pragmatisch festgelegt wird. Ein *n*-stelliger Ausgang sei definiert als Ausgang der Länge *n*, d.h. als die letzten (schließenden) *n* Buchstaben einer lexikalischen Grundform.

Als Beispiele für die genannte Lernstrategie seien die deutschen Verben *gehen*, *flehen* und *drehen* genannt. Bei ihnen besteht Ähnlichkeit in den letzten 4 Buchstaben, dem Ausgang –*ehen*. Würde ein Sprachlernender das Verb *flehen* mit seinen Schlüsselformen (*flehen*, *flehte*, *gefleht*) als erstes dieser Gruppe lernen, würde er diese Regelmäßigkeit auf die beiden anderen übertragen wollen. Für *drehen* (*drehte*, *gedreht*) stimmt dieser Transfer, für *gehen* (*ging*, *gegangen*) allerdings nicht. Letzteres muss als explizite Ausnahme der Verben mit dem Ausgang –*ehen* gelernt werden.

Diese Strategie stimmt bei jedem Simplex und seinen präfigierten Verben (mit wenigen Ausnahmen), bei den regelmäßigen und einem Teil der unregelmäßigen Verben (im Deutschen bei etwa einem Neuntel [Holl, 2002, 159]), aber nicht durchgängig, was eine erhebliche Fehlerquelle bedeutet. Diese Strategie kann also hilfreich oder irreführend sein. Erst wenn sie scheitert, werden in einer zweiten Stufe weitere Verben mit ihren Konjugationsbesonderheiten intensiv gelernt.

Hier treten zwei konkurrierende Formen von Ähnlichkeit auf: Die rückläufige Ähnlichkeit der lexikalischen Grundform und die morphologische Analogie. Diese sind keineswegs deckungsgleich, vor allem kann man bei einer Wortart nicht von der rückläufigen Ähnlichkeit zweier lexikalischer Grundformen auf die morphologische Analogie der zugehörigen Lexeme schließen. Nun ist das der genannten Strategie zugrunde liegende analogische Denken (der Analogieschluss von partieller auf totale Ähnlichkeit) ein allgemeines, wesentliches – häufig unbewusstes – Grundprinzip menschlichen Lernens und Denkens. Es kann also nicht einfach ausgeschaltet werden, sondern man muss bewusst damit umgehen, der Sprachlernende ebenso wie der Sprachlehrende. Es ist am besten, dem Sprachlernenden von vornherein zu zeigen, in welchen Fällen rückläufige Ähnlichkeit morphologische Analogie impliziert und in welchen nicht. Das ist in Holl, 2002, 151-158 ausführlich gezeigt.

Es ist daher der bisherigen Form der Clusteranalyse morphologischer Systeme, die für Nachschlagewerke weiterhin ihren Sinn behält, eine zweite hinzuzufügen, die ganz gezielt den Spezifika analogischen Denkens auf der Basis rückläufiger Ähnlichkeit sprachdidaktisch Rechnung trägt. Ziel ist die automatisierte explizite Ermittlung homogener Cluster ausgangsgleicher Lexeme einer bestimmten Wortart. Ein Cluster heiße morphologisch homogen (im Folgenden kurz homogen), wenn alle seine Lexeme morphologisch analog sind. Das bedeutet, dass z.B. die Verben *rasieren*, *organisieren*, *charakterisieren* usw., d.h. diejenigen mit dem Ausgang –*sieren*, ein homogenes Cluster bilden, da sie die gleichen Konjugationsmerkmale haben.

Für diese Form der Clusteranalyse wird die Menge aller Lexeme einer Wortart – in Gestalt ihrer lexikalischen Grundformen – nicht ungeordnet oder in der üblichen alphabetischen Sortierung von links betrachtet, sondern als rückläufig, also von rechts alphabetisch sortierte Menge. Dadurch werden ausgangsgleiche lexikalische Grundformen benachbart angeordnet. Geordnete Grundmengen liegen in uns aus der Informatik bekannten Clusteranalysen nicht vor, so dass dort übliche Algorithmen nicht verwendet werden können.

[5] Tertium comparationis (lat. „das Dritte des Vergleichs“) ist die Eigenschaft oder Dimension, die zwei zu vergleichende Gegenstände gemeinsam haben und die den Vergleich erst ermöglicht.

Bei der Entwicklung von Clusteranalyse-Verfahren ist generell zwischen zwei Grundtypen zu unterscheiden: „Bei den meisten Varianten wird so verfahren, dass entweder jedes zu gruppierende Objekt als ein Anfangscluster oder alle Objekte als ein Cluster gewählt werden. Danach werden die Anfangscluster zusammengefasst oder das alle Objekte umfassende Cluster aufgespaltet. In beiden Fällen geschieht das so, dass die Abstände zwischen den Elementen eines Clusters möglichst gering werden“ (Alpar / Niedereichholz, 2000, 11). Im vorliegenden Fall ist diese Abstandsbedingung sehr einfach; sie besagt, dass größtmögliche Cluster ermittelt werden, die morphologisch analoge, ausgangsgleiche Lexeme enthalten. Die Strukturierung des allumfassenden Anfangsclusters wird auch als Top-down-Clusteranalyse bezeichnet (divisiv), die Zusammenfassung ähnlicher einelementiger Anfangscluster als Bottom-up-Clusteranalyse (agglomerativ). Bei dem vorliegenden Projekt wird die erste Variante verwendet, da ihre Algorithmen einfacher und besser implementierbar sind.

Als Ergebnis erhält man eine stabile, weitgehend objektivierbare Basis, auf der die Nachbearbeitung aufbauen kann.

3. Konzept des morphologischen Data Minings

3.1 gliedert den betrachteten Forschungsansatz in einen analytischen und einen synthetisch-generativen Teilbereich und zeigt den jeweiligen aktuellen Forschungsstand. 3.2 analysiert anhand des englischen Verbalsystems als Beispiel die Register, welche den Ausgangspunkt für den zu entwickelnden Algorithmus bilden, und gibt eine detaillierte Darstellung der Phasen eines Data-Mining-Prozesses. 3.3 befasst sich mit der Modellierung des synthetischen Algorithmus und erklärt dessen Funktionsweise. In 3.4 werden Konventionen für die Register der Verbalsysteme festgelegt, welche die bereits entwickelten Register optimieren und richtungsweisend für die Erstellung zukünftiger sind. Im Zusammenhang damit steht die Idee, den Algorithmus so zu konzipieren, dass er unabhängig von einem speziellen Verbalsystem und dem damit verbundenen Register Anwendung finden kann.

3.1 Stand der Forschung

Der vorliegende Forschungsgegenstand untergliedert sich in zwei Teilbereiche, einen analytischen und einen synthetisch-generativen. 3.1.1 beschäftigt sich mit dem analytischen Teilbereich, d.h. mit der Entwicklung von Registern (von Regellexemen) für die unterschiedlichsten Verbalsysteme durch morphologisches Data Mining. 3.1.2 und 3.1.3 beziehen sich auf die synthetisch-generativen Teilbereiche. 3.1.2 betrifft das Auffinden von passenden Regellexemen zu einem Suchlexem. 3.1.3 gibt einen Ausblick auf die Erzeugung der Schlüsselformen des Suchlexems.
Diese Untersuchung beschäftigt sich schwerpunktmäßig mit dem ersten Schritt des synthetisch-generativen Teilbereichs.

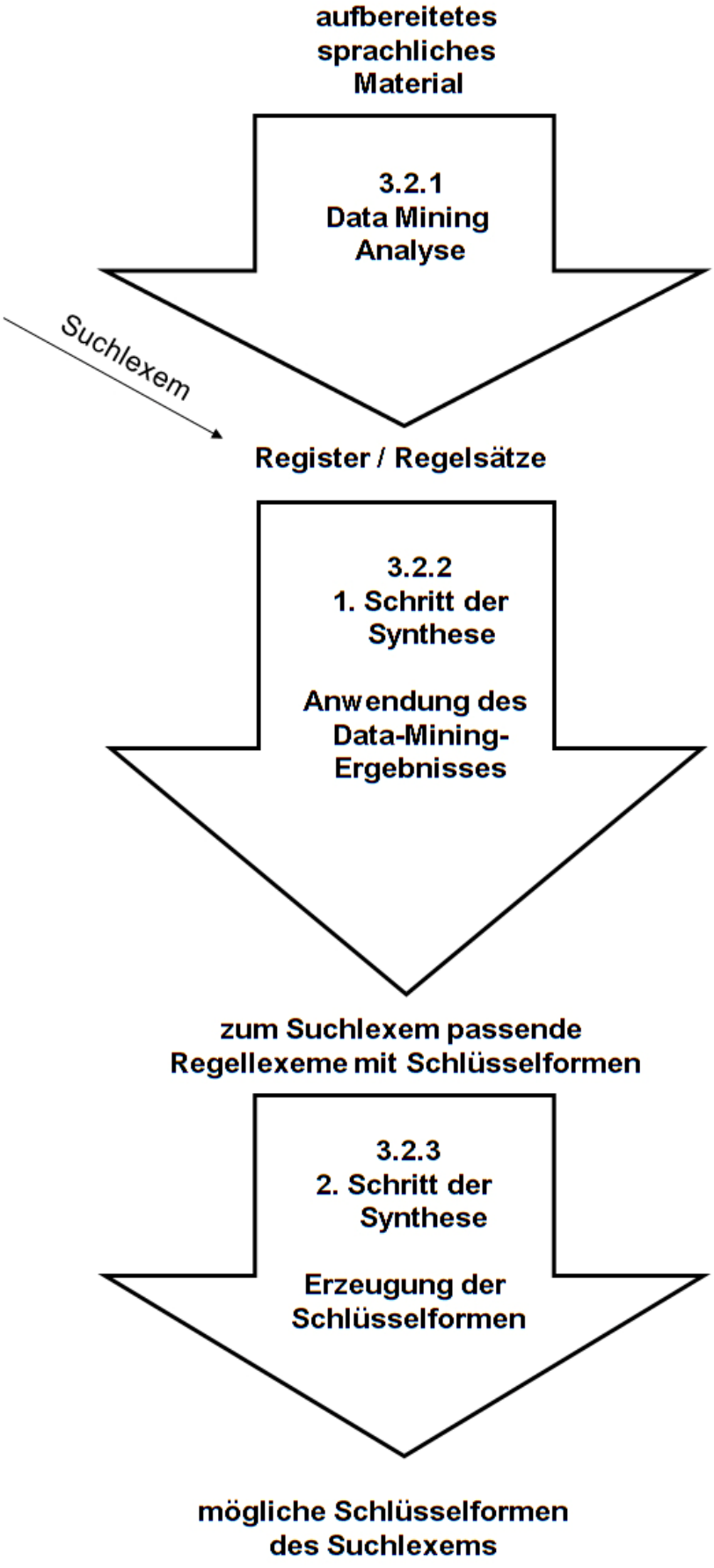


Abbildung 3.01: Verarbeitungsprozess

3.1.1 Analytischer Teilbereich: Ermittlung morphologisch homogener Cluster

In diesem Forschungsbereich werden morphologische Analogieregeln auf der Basis der rückläufigen Ähnlichkeit (Ausgangsgleichheit) von Präsensinfinitiven als den lexikographischen Grundformen konstruiert. Die Datenanalyse erfolgt mittels Clusteranalyse, einer Methode des Data Mining (vgl. 2.2).
Die hierbei entstehenden verdichteten sprachabhängigen Register bestehen aus morphologisch homogenen Clustern, die rückläufig ähnliche, morphologisch analoge Vertreter der untersuchten Wortart enthalten. Diese Register erlauben sprachdidaktisch verwendbare, quantitative und strukturelle Aussagen über flexionsmorphologische Systeme (vgl. Holl, 2003, 107-119).
Die Datenanalyse, Aufbereitung und Auswertung wurden bereits für folgende Verbalsysteme erfolgreich durchgeführt:

- Latein (1988)
- Katalanisch (1988)
- Portugiesisch (1988)
- Rumänisch (1988)
- Italienisch (1988, 2002)
- Spanisch (1988, 2002)
- Französisch (1988, 2002, 2003)
- Deutsch (2002, 2004)
- Russisch (2004)
- Neugriechisch (2006)
- Altgriechisch (2006)
- Englisch (2002, 2007)
- Schwedisch (2002, 2007)
- Kroatisch (2008)
- das schwedische Substantiv (2007)

Die so gewonnenen Register bilden für die in 3.1.2 und 3.1.3 dargelegten synthetisch-generativen Teilbereiche die Ausgangsbasis.

3.1.2 Erster Schritt des synthetisch-generativen Teilbereichs: Auffindung passender Regellexeme

Auf der Basis einheitlicher Regel-Konventionen (vgl. 3.4) für die Register wird mittels der in der Informatik zur Verfügung stehenden Funktionselemente ein Algorithmus modelliert (erste Gedanken zu dieser Art der Formalisierung finden sich in Holl, 1988, 184), welcher zu einem Suchlexem die entsprechenden Regellexeme eines Registers findet.
Um Korrektheit und Sprachunabhängigkeit zu überprüfen, wird dieser Algorithmus in einer ersten Version ("SMIRT Ver. 1.0")[6] realisiert.

[6] Der Name "SMIRT" ist abgeleitet von der englischen Wortkreation "Smirting", welche aus den Worten "Smoking and Flirting" gebildet wurde. Smirting bezeichnet das Flirten beim Tabakkonsum vor Gebäuden wie Büros oder Restaurants, in denen ein Rauchverbot gilt. Der Algorithmus wie auch die englische Wortkreation sind zu einer Zeit entstanden, als zunehmend in Europa ein generelles Rauchverbot eingeführt wurde. Die gleiche Temporalität und die Tatsache, dass auch das Wort "Smirting" zu den natürlichsprachlichen Verbalformen zählt, mit welchen der Algorithmus arbeitet, haben zu dieser Namensgebung geführt. Zur Behandlung von Neologismen siehe auch 5.

3.1.3 Zweiter Schritt des synthetisch-generativen Teilbereichs: Erzeugung der Schlüsselformen des Suchlexems

In diesem Forschungsbereich sollen zukünftig die morphologischen Eigenschaften (d.h. hier das Bildeprinzip der Schlüsselformen) der durch den Algorithmus ermittelten Regellexeme auf das Suchlexem übertragen werden. Im Falle des deutschen Suchlexems *schreiben* liefert der Algorithmus aus 3.1.2 das Regellexem *reiben* mit dessen Schlüsselformen *rieb* und *gerieben* als Ergebnismenge. Eine analogische Übertragung auf das Suchlexem *schreiben* liefert die Wortformen *schrieb* und *geschrieben*.

3.2 Beschreibung des analytischen Verarbeitungsschritts

Vor der Entwicklung eines Algorithmus ist es erforderlich, den Aufbau des Inputs zu analysieren, d.h. sich den vorgelagerten analytischen Verarbeitungsprozess sowie dessen Teilergebnisse näher anzuschauen. Hierzu wird die Struktur der Lexemregister analysiert (3.2.1) und das Phasenmodell des Data-Mining-Prozesses vorgestellt (3.2.2).

3.2.1 Struktur der Lexemregister

Im analytischen Verarbeitungsschritt bilden aufbereitete Daten eines Verbalsystems den Input der Datenanalyse. Der Output besteht aus Registern morphologischer Cluster, welche die nachfolgende Tabelle zusammenfassend illustriert.

Homogene Cluster - homC	**Inhomogene Cluster (basic cluster - basC)**
1) homC aus verschiedenen Verben (Simplicia), von denen nur **ein** Repräsentant in der Kommentarspalte des Lexemregisters genannt ist (keine Detaillierung): Dies ist der Fall bei regelmäßig flektierenden Lexemen mit regelhaften Besonderheiten. Beispiel: homC *~sh, ~shed, ~shed* *(fish, fished, fished)* wegen 3. Person Singular auf *-es*.	basC aus verschiedenen Verben (Simplicia), von denen nur ein Repräsentant in der Kommentarspalte des Lexemregisters genannt ist (keine Detaillierung): Es gibt Ausnahmen, aber die Mehrheit sind regelmäßig flektierende Lexeme. Beispiel: *basC ~C, ~Ced, ~Ced* *(look, looked, looked)* Ausnahme: z.B. *send, sent, sent*
2) homC aus verschiedenen Verben (Simplicia), von denen **alle** Repräsentanten in der Kommentarspalte des Lexemregisters genannt sind (Detaillierung): Dies ist der Fall bei unregelmäßig flektierenden Lexemen. Beispiel: homC *~ling, ~lung, ~lung* *(cling, sling, fling)*	basC2: Ein Cluster, das im Bereich eines basic clusters enthalten ist, aber anders flektiert als das übergeordnete basic cluster. Von den verschiedenen Verben (Simplicia) ist nur ein Repräsentant in der Kommentarspalte des Lexemregisters genannt (keine Detaillierung): Es bestehen Ausnahmen, aber die Mehrheit sind regelmäßig flektierende Lexeme mit regelhaften Besonderheiten. Beispiel: basC2 *#CCVd, #CCVdded, #CCVdded* *(kid, kidded, kidded)* # markiert die Wortgrenze Ausnahme: z.B. *shed, shed, shed*
3) homC aus einem Verb (Simplex): Einelementiges Cluster, welches nicht als solches markiert wird, da trivial (keine Detaillierung): Beispiel: *send, sent, sent*	Weitere basC(3,…,n)-Cluster, welche sich im Bereich eines übergeordneten Clusters befinden, sind möglich.

Tabelle 3.01: Typisierung der homogenen und inhomogenen Cluster (vgl. Holl, 2007, 108-109,115)

Diese Art der Clusterbildung erfolgt im Allgemeinen nach einer hierarchischen Struktur, welche der nachfolgende selbstähnliche Strukturbaum illustriert.

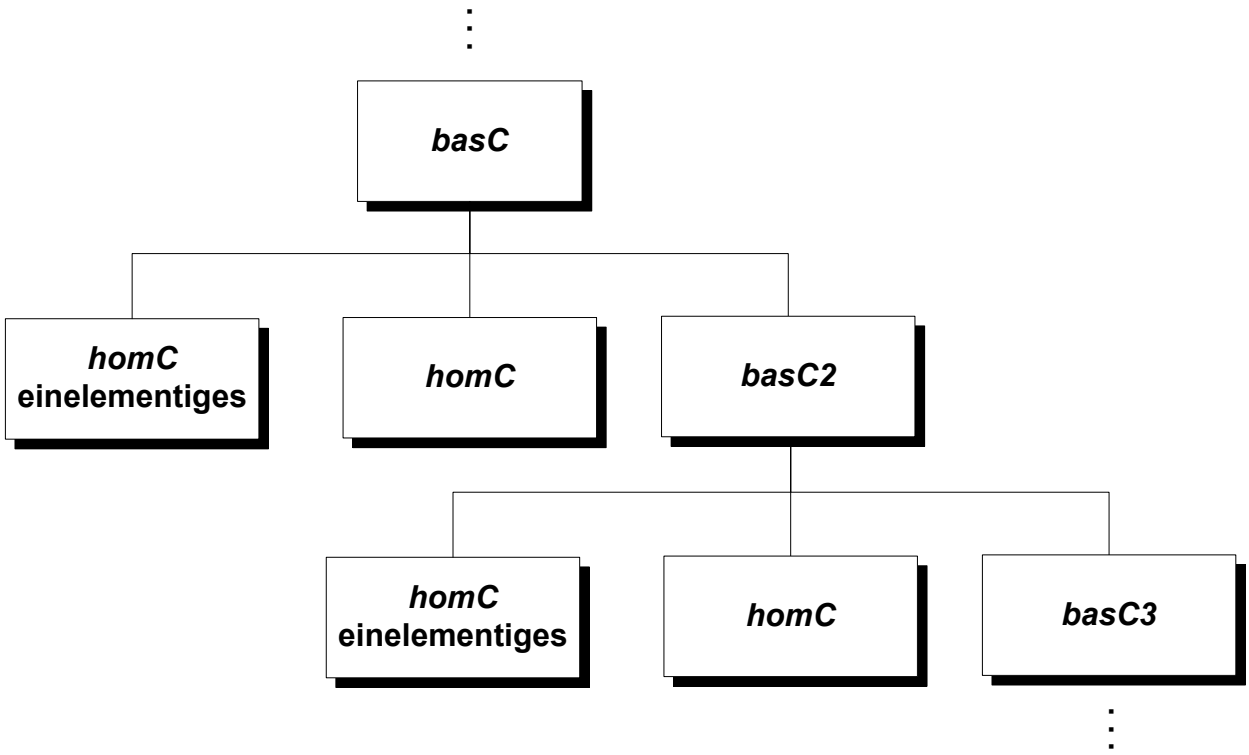


Abbildung 3.02: Allgemeiner Strukturbaum der Clusterbildung

Der in Abb. 3.02 beschriebene Baum findet sich beispielsweise in einem Ausschnitt der Struktur des englischen Verbalsystems wieder.

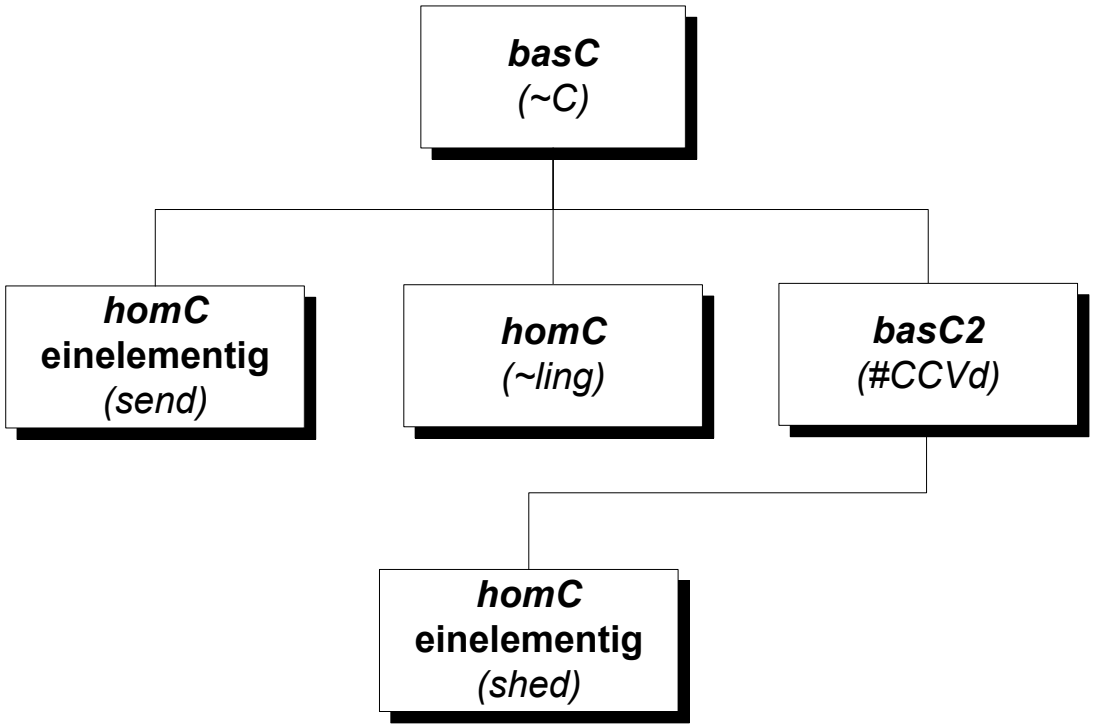


Abbildung 3.03: Beispiel eines Strukturbaums für englische Verben

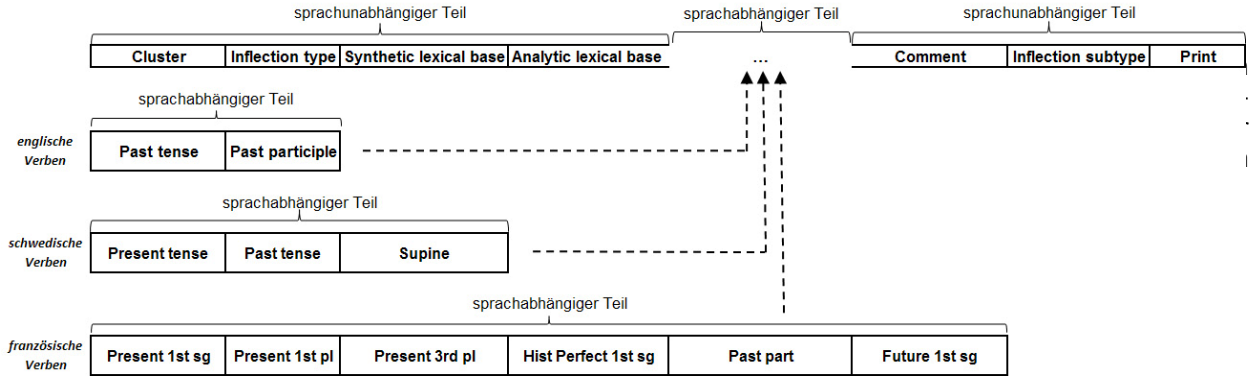


Abbildung 3.04: Datenstrukturen für verschiedene Verbalsysteme

Definition der Registerspalten:

- Sprachunabhängige Spalten
  - Cluster
    In dieser Spalte erfolgt die Kennzeichnung eines Clusters (siehe Tabelle 3.01). Hierbei kann es sich um homC oder basC handeln.
  - Inflection type
    Der Flexionstyp beschreibt die Beugung eines Lexems, im Fall von Verben den Konjugationstyp und im Fall von Substantiven den Deklinationstyp.
  - Synthetic lexical base (ähnlich dem Lexikoneintrag)
    Die lexikalische Grundform ist bei Substantiven der Nominativ Singular und bei Verben meist der Infinitiv Präsens Aktiv.
    Die synthetische lexikalische Grundform dient dem synthetisch-generativen Algorithmus als Vergleichsform mit dem Suchlexem. Sie entspricht den Einträgen der Spalte "analytic lexical base" reduziert um deren Kennzeichnungen (außer dem #-Zeichen für die Wortgrenze), da vom Benutzer eine derartige Kennzeichnung des Suchlexems nicht erwartet werden kann.
  - Analytic lexical base (aufbereiteter Lexikoneintrag)
    Die Einträge in konventionellen Lexika sind ggf. um folgende Kennzeichnungen ergänzt:
    - Unterstrich zur Kennzeichnung von Präfixen
      Beispiel (engl.): *de_pend, depended, depended*
    - Ziffern zur Kennzeichnung morphologischer Varianten mit Bedeutungsunterschied
      Beispiel (engl.):
      1shed, shedded, shedded ('ein Fahrzeug in einem Depot parken')
      2shed, shed, shed ('Blätter oder Früchte fallen zu Boden')
    - Griechische Buchstaben zur Kennzeichnung morphologischer Varianten ohne Bedeutungsunterschied
      Beispiel (engl.): *α+learn, learned, learned* ('lernen')
      β+learn, learnt, learnt ('lernen')
  - Comment
    In der Kommentarspalte werden die Lexeme eines Clusters mit ihren lexikalischen Grundformen in Abhängigkeit vom Clustertyp (wie in Tabelle 3.01 dargestellt) aufgezählt oder die Bedeutung von Flexionsvarianten angegeben.
  - Inflection subtype
    Beugungsfeinklassifizierung, die nicht zur Clusterbildung herangezogen wird.
  - Print
    In dieser Spalte werden alle ausgaberelevanten Zeilen mit "P" gekennzeichnet. Der synthetisch-generative Algorithmus wird später nur diese Zeilen zur Verarbeitung heranziehen.

- Sprachabhängige Spalten
  - Die Spalten in diesem Teil sind bestimmt durch die Schlüsselformen der jeweiligen Sprache-Wortart-Kombination (siehe Abbildung 3.04).

3.2.2 Die drei Phasen des Data-Mining-Prozesses

Pre-processing oder auch Datenvorverarbeitung ist die manuelle Aufbereitung und Strukturierung der Ausgangsdaten. Hierbei ist zu beachten, dass die Aufbereitung nicht zu einem Verlust informationsrelevanter Parameter führen darf. Die Datenvorverarbeitung ist der Schlüssel zu einer effizienten Muster- / Clustererkennung. In dieser Phase werden von einem Linguisten die Lexeme der zu untersuchenden Sprache-Wortart-Kombination in einer Excel-Tabelle erfasst und kategorisiert (vgl. Abbildung. 3.05).

	sprachunabhängiger Teil			sprachabhängiger Teil		sprachunabhängiger Teil		
Cluster	**Inflection type**	**Synthetic lexical base**	**Analytic lexical base**	**Past tense**	**Past participle**	**Comment**	**Inflection subtype**	**Print**
	i-u-u / 0	cling	cling	clung	clung			
	i-u-u / 0	fling	fling	flung	flung			
	i-u-u / 0	sling	sling	slung	slung			

Abbildung 3.05: Beispielergebnis einer Pre-processing-Phase (vgl. Holl, Maroldo, Urban, 2007, 77)

Processing
Der Data-Mining-Algorithmus analysiert die im Pre-processing aufbereiteten Daten auf der Suche nach bestimmten Mustern / Clustern, d.h. es werden morphologisch analoge Cluster ermittelt und in das jeweilige Register eingetragen. Bezogen auf das Beispiel wird das homogene Cluster *~ling* mit den lexikalischen Grundformen (lexical bases) *cling*, *fling* und *sling* gefunden.

	sprachunabhängiger Teil			sprachabhängiger Teil		sprachunabhängiger Teil		
Cluster	**Inflection type**	**Synthetic lexical base**	**Analytic lexical base**	**Past tense**	**Past participle**	**Comment**	**Inflection subtype**	**Print**
ling	i-u-u / 0	cling	cling	clung	clung			
ling	i-u-u / 0	fling	fling	flung	flung			
ling	i-u-u / 0	sling	sling	slung	slung			

Abbildung 3.06: Beispielergebnis einer Processing-Phase (vgl. Holl, Maroldo, Urban, 2007, 115)

Post-processing meint die manuelle Nachbearbeitung und Aufbereitung der durch Data Mining gewonnenen Register. Für das untersuchte Verbalsystem bedeutet dies:

<u>Erster Schritt der manuellen Nachbearbeitung:</u>

Erzeugung einer übersichtlichen Druckaufbereitung

- Einführung einer neuen Zeile für das Cluster mit entsprechender Benennung (im Beispiel homC)
- Zusammenfassung der sich im Cluster befindlichen lexikalischen Grundformen (lexical bases) im Kommentarfeld (Comment):
  vollständige Aufzählung bei unregelmäßigen Lexemen, Beispiele bei regelmäßigen Lexemen (ggf. mit regelhaften Besonderheiten)
- Kennzeichnung von unter verschiedenen linguistischen Perspektiven relevanten Regellexemen mit “P“ in der Printspalte

	sprachunabhängiger Teil			sprachabhängiger Teil		sprachunabhängiger Teil		
Cluster	**Inflection type**	**Synthetic lexical base**	**Analytic lexical base**	**Past tense**	**Past participle**	**Comment**	**Inflection subtype**	**Print**
homC	i-u-u / 0	ling	~ling	~lung	~lung	cling,sling,fling		P
ling	i-u-u / 0	cling	cling	clung	clung			
ling	i-u-u / 0	fling	fling	flung	flung			
ling	i-u-u / 0	sling	sling	slung	slung			

Abbildung 3.07: Beispielergebnis einer Post-processing-Phase, erster Schritt (vgl. Holl, Maroldo, Urban, 2007, 106)

Zweiter Schritt der manuellen Nachbearbeitung:

Aufbereitung für den synthetisch-generativen Algorithmus

- Elimination nicht mit “P“ gekennzeichneter Regellexeme

	sprachunabhängiger Teil			sprachabhängiger Teil		sprachunabhängiger Teil		
Cluster	**Inflection type**	**Synthetic lexical base**	**Analytic lexical base**	**Past tense**	**Past participle**	**Comment**	**Inflection subtype**	**Print**
homC	i-u-u / 0	ling	~ling	~lung	~lung	cling,sling,fling		P

Abbildung 3.08: Beispielergebnis einer Post-processing-Phase, zweiter Schritt (vgl. Holl, Maroldo, Urban, 2007, 106)

Das Ergebnis des Post-processing bildet einen Teil des Inputs für den ersten Schritt der Synthese. Einen weiteren Teil bildet ein zufällig gewähltes Suchlexem.

3.3 Konzept des synthetisch-generativen Algorithmus

3.3.1 erklärt die Grundidee des Algorithmus anhand einer Suchbaumhierarchie mit verschiedenen Clustertypen. 3.3.2 spezifiziert die Eingangsgrößen (Input) für den Algorithmus. 3.3.3 spezifiziert das Suchlexem. 3.3.4 erläutert die Funktionsweise des Algorithmus. 3.3.5, 3.3.6 und 3.3.7 erklären die Funktionen “total length compare“, “prefix cut“ und “longest match“.

3.3.1 Grundidee des Algorithmus

Der Algorithmus soll zu einem beliebigen Suchlexem das ihm am besten entsprechende Regellexem (oder mehrere solche) aus dem Register liefern unter Beachtung von Flexionstyp und Lexemausgang. Dazu stellt 3.3.1.1 den verwendeten Suchbaum vor. 3.3.1.2 und 3.3.1.3 erklären die Bedeutung von Clustern mit und ohne Wortbegrenzung bzw. mit und ohne Jokerzeichen. 3.3.1.4 kommentiert den in 3.3.1.1 beschriebenen Suchbaum. 3.3.1.5 illustriert Testfälle des Algorithmus.

3.3.1.1 Suchbaum

Der Start des Algorithmus wird durch die Eingabe eines zufälligen Suchlexems initiiert. Der Endpunkt des Algorithmus ist erreicht, wenn zum eingegebenen Suchlexem einelementige bzw. mehrelementige Cluster mit rückläufig gleichem Ausgang gefunden sind oder die Suche erfolglos bleibt und die “leere Menge“ zurückgegeben wird.
Die Suchlexeme werden ausschließlich mit den Einträgen in der Tabellenspalte “Synthetic lexical base“ verglichen. Da dort keine Präfixmarkierungen vorhanden sind, muss der Algorithmus, um die Korrektheit des Ergebnisses sicherzustellen, befähigt sein, ein morphologisch präfigiertes Suchlexem mit Präfixmarkierung um dieses Präfix zu reduzieren (3.3.6 prefix cut). Ziel des Algorithmus ist es, die Cluster zu finden, welche mit dem Suchlexem die längste rückläufige Übereinstimmung besitzen (3.3.7 longest matching von rechts) (vgl. Holl, 2006, 51). Dies kann, wie die nachfolgende Abbildung deutlich macht, in unterschiedlichen Fällen erreicht werden.[7]

[7] Hier und im Folgenden wird die Zuordnung eines Suchlexems zu Regellexemen mit einem einheitlichen Muster beschrieben:

Suchlexem → Spalte “Synthetic lexical base“ : Spalte “Analytic lexical base“, Schlüsselformen

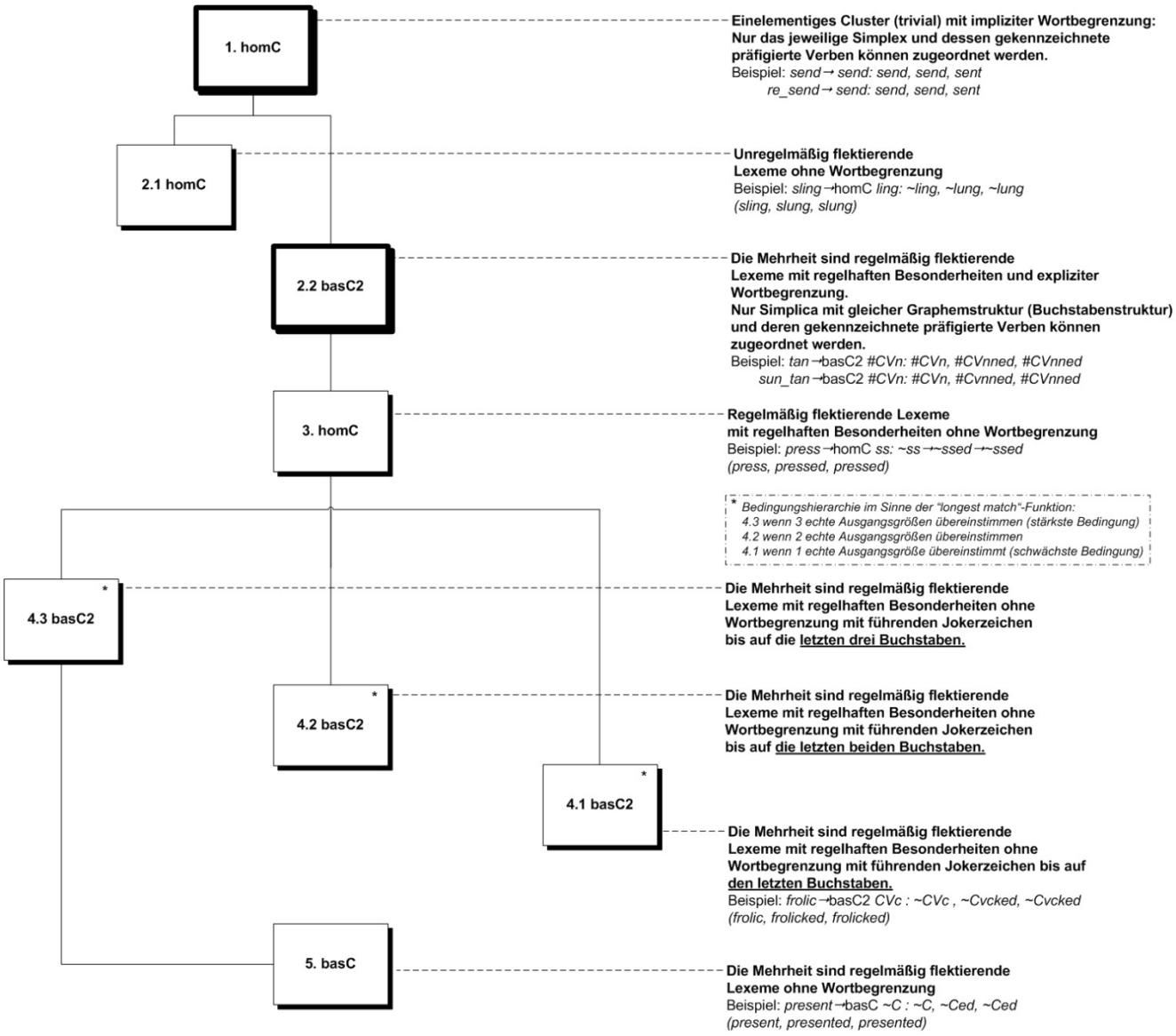


Abbildung 3.09: Suchbaumhierarchie für den Algorithmus

Ausgangspunkt für die Entwicklung einer Regel- / Suchbaumhierarchie ist der Strukturbaum der Clusterbildung (Abb. 3.03). Der Strukturbaum wird in umgekehrter Reihenfolge durchlaufen, d.h. während sich der Strukturbaum von allgemeinen zu speziellen Clustern orientiert, orientiert sich die Suchbaumhierarchie von speziellen zu allgemeinen Clustern, um die längste rückläufige Übereinstimmung zwischen Suchlexem und Regellexemen zu finden. Die Eigenschaften der verschiedenen Cluster der Suchbaumhierarchie werden in Tabelle 3.02 tabellarisch dargestellt.

Fälle nach Abb. 3.09	Wortbegrenzung (3.3.1.2)	Jokerzeichen (3.3.1.3)	Elemente im Cluster	Vollstring (ohne ~) oder Teilstring (mit ~)
1 homC	implizit	-	ein Simplex	Voll
2.1 homC	-	-	>1	Teil
2.2 basC2	explizit	ja	>1	Voll
3 homC	-	-	>1	Teil
4.3 basC2	-	ja	>1	Teil
4.2 basC2	-	ja	>1	Teil
4.1 basC2	-	ja	>1	Teil
5 basC	-	ja	>1	Teil

Tabelle 3.02: Suchbaumhierarchie für den Algorithmus

3.3.1.2 Erste Fallunterscheidung: Cluster mit und ohne Wortbegrenzung

Hierbei sind zwei grundsätzliche Fälle zu unterscheiden: Cluster mit Wortbegrenzung und Cluster ohne Wortbegrenzung. Eine Wortbegrenzung entspricht der exakten zulässigen Länge (Anzahl Buchstaben), die ein Wort besitzen muss (kürzere Wortlängen sind nicht zulässig). Die Nummerierungen in den Beispielen beziehen sich auf Tab. 3.02.
Cluster mit einer Wortbegrenzung sind entweder durch ein voranstehendes #-Zeichen oder gar nicht gekennzeichnet.

Beispiel zu 1 (implizite Wortbegrenzung):
send→send: send, send, sent
(Wortlänge des Simplex: 4 Buchstaben)
re_send→send: send, send, sent
(Wortlänge des Simplex: 4 Buchstaben)
also *re_send, re_sent, re_sent*
Derartige einelementige Cluster (bestehend aus einem Simplex und seinen präfigierten Verben) werden in den Registern nicht eigens als “homC“ markiert.

Beispiel zu 2.2 (explizite Wortbegrenzung):
tan→basC2 *#CVn: #CVn, #CVnned, #CVnned*
(Wortlänge des Simplex: 3 Buchstaben)
also *tan, tanned, tanned*
sun_tan→basC2 *#CVn: #CVn, #CVnned, #CVnned*
(Wortlänge des Simplex: 3 Buchstaben)
also *sun_tan, sun_tanned, sun_tanned*

Cluster ohne Wortbegrenzung sind mit einer voranstehenden Tilde (~) gekennzeichnet.

Beispiel zu 3:
press→homC *ss: ~ss, ~ssed, ~ssed*
also *press, pressed, pressed*
un_dress / *undress*→homC *ss: ~ss, ~ssed, ~ssed*
also *undress, undressed, undressed*

Beispiel zu 5:
present→basC *C: ~C, ~Ced, ~Ced*
also *present, presented, presented*
look→basC *C: ~C, ~Ced, ~Ced*
also *look, looked, looked*

3.3.1.3 Zweite Fallunterscheidung: Cluster mit und ohne Jokerzeichen

Um den Vergleich eines Suchlexems mit “Synthetic lexical base“-Einträgen zu ermöglichen, die die Jokerzeichen “C“ und “V“ enthalten, muss die Buchstabenfolge des Suchlexems teilweise in Jokerzeichen umgesetzt werden. Das Jokerzeichen “C“ repräsentiert die Konsonanten, “V“ die Vokale der jeweiligen Sprache.
Hierzu wird das Suchlexem auf dessen Buchstabenzusammensetzung in Leserichtung (europäisch) folgendermaßen überprüft. Handelt es sich bei einem Buchstaben um einen Vokal, wird in den Suchalternativen stellvertretend ein “V“ eingetragen, und wenn es sich um einen Konsonanten handelt, ein “C“. Diese Suchalternativen werden in Abhängigkeit von einer Alphabet-Tabelle des untersuchten Verbalsystems gebildet.

Das ergibt die nachfolgenden unterschiedlichen Suchalternativen (Nummerierung erfolgt nach Abb. 3.09).

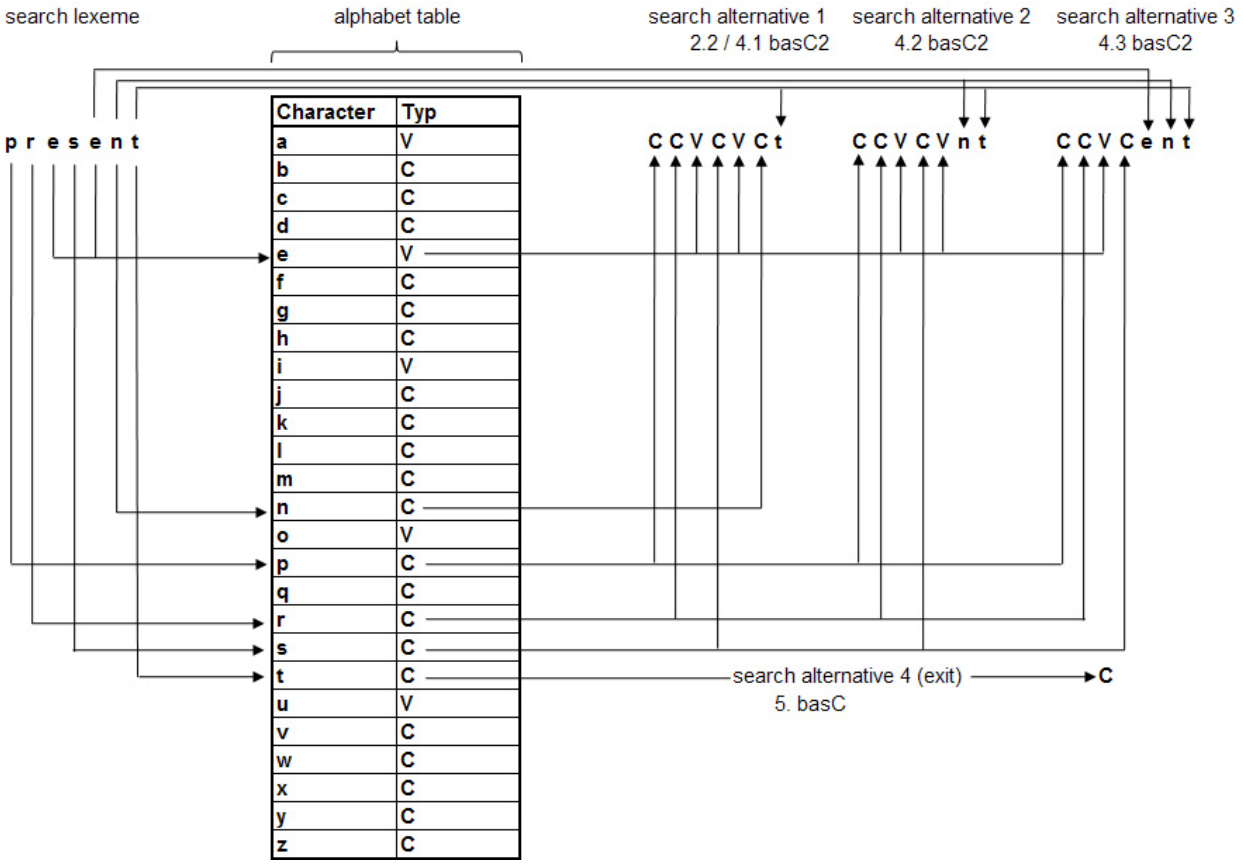


Abbildung 3.10: Generierung der Suchalternativen

Die Suchalterativen in Abb. 3.10 werden nur beim Aufruf der “longest match“-Funktion generiert. In Abhängigkeit von der Länge des an die Funktion übergebenen Suchlexems werden drei Fälle unterschieden.

1. D
   ie Suchlexemlänge ist größer eins.
   →Generierung der Suchalternative 1

2. D
   ie Suchlexemlänge ist größer zwei.
   →Generierung der Suchalternativen 1 und 2

3. D
   ie Suchlexemlänge ist größer drei.
   →Generierung der Suchalternativen 1, 2 und 3

Hierbei wird in Leserichtung (europäisch) Buchstabe um Buchstabe mittels der Alphabet-Tabelle durch das jeweilige Jokerzeichen ersetzt. Die oben genannten Fallunterscheidungen stellen sicher, dass bei der Suchalternative 1 alle Buchstaben bis auf den letzten, bei Suchalternative 2 bis auf die letzten beiden und bei Suchalternative 3 bis auf die letzten drei Buchstaben ersetzt werden. Im Anschluss werden die Suchalternativen durch die echten Buchstaben des Suchlexems nach rechts aufgefüllt.

Die “longest match“-Funktion benötigt jedoch für den weiteren Funktionsverlauf zwei weitere Alternativen. Die Suchalternative (# search alternative 1) entspricht exakt der Suchalternative 1 mit vorangestelltem Hashzeichen, während die Suchalternative (search alternative 4 exit) der Jokerzeichendarstellung des letzten Buchstabens des Suchlexems entspricht.

3.3.1.4 Kommentar zum Suchbaum

1. Der Buchstabenfolge des Suchlexems entspricht ein genau gleich langes Cluster (total length compare) mit impliziter Wortbegrenzung.
   Beispiel: *send→send: send, send, sent*

2. – 5. Es gibt ein Cluster, das einem Ausgang des Suchlexems entspricht (longest match). Die Generierung der Suchalternativen der Fälle 2.2, 4.1, 4.2, 4.3 und 5 erfolgt, wie in Abbildung 3.10 illustriert.

2.1 Ein Ausgang des Suchlexems wird von einem unregelmäßig flektierenden homC repräsentiert. Dieses Cluster besitzt keine Wortbegrenzung. Beispiel: *sling→*homC *ling: ~ling, ~lung, ~lung*

2.2 Das Suchlexem gehört zu einem Cluster, in welchem alle Buchstaben bis auf den letzten durch Jokerzeichen (C / V) ersetzt wurden. Eine Wortbegrenzung ist hier durch das voranstehende #-Zeichen gekennzeichnet (# search alternative 1).
Beispiel: *tan→*basC2 *#CVn: #CVn, #CVnned, #CVnned*

3. Ein Ausgang des Suchlexems entspricht einem Cluster ohne Wortbegrenzung und ohne Jokerzeichen.
Beispiel: *press→*homC *ss: ~ss, ~ssed, ~ssed*

4.3 Ein Ausgang des Suchlexems, in welchem alle Buchstaben bis auf die letzten drei durch Jokerzeichen (C / V) ersetzt wurden, entspricht einem Cluster ohne Wortbegrenzung (search alternative 3).

4.2 Ein Ausgang des Suchlexems, in welchem alle Buchstaben bis auf die letzten beiden durch Jokerzeichen (C / V) ersetzt wurden, entspricht einem Cluster ohne Wortbegrenzung (search alternative 2).
Die Suchalternativen 2 und 3 werden nur sicherheitshalber eingeführt, obwohl dazu noch keine Beispiele vorliegen.

4.1 Ein Ausgang des Suchlexems, in welchem alle Buchstaben bis auf den letzten durch Jokerzeichen C / V) ersetzt wurden, entspricht einem Cluster ohne Wortbegrenzung (search alternative 1).
Beispiel: *frolic→*basC2 *CVc : ~CVc , ~CVcked, ~CVcked*

5. Wenn das Suchlexem weder einem homC noch einem basC2 entspricht, wird eine weitere Alternative generiert. Hierbei wird der letzte Buchstabe des Suchlexems durch ein Jokerzeichen (C / V) ersetzt. Dieser allgemeine Ausgang entspricht im Fall der englischen Verben einem basC ohne Wortbegrenzung (search alternative 4).
Beispiel: *present→*basC *C : ~C, ~Ced, ~Ced*

Diese Art der Suchlexem-Konvertierung in Jokerzeichen-Kombinationen ist derzeitig lediglich für die englischen und deutschen Verbregister nötig. Es besteht jedoch die Möglichkeit, diese Form der Cluster-Benennung auch bei anderen Verbalsystemen einzusetzen. Daher müssen auch diese Fälle in vollem Umfang durch den Algorithmus abgebildet werden. Die nachfolgenden Testfälle sollen dies verdeutlichen und einen kurzen Überblick über die Vielzahl möglicher Suchresultate illustrieren.

3.3.1.5 Testfälle

		1	2.1 / 3	2.2	4.1-4.3 / 5
	Clustertypen	homC	homC / basC / basC2	basC / basC2	homC / basC / basC2
	Regellexeme / Suchlexeme	Vollstring	Teil- / Vollstring	Vollstring	Teilstring
		einelementige Cluster	mehrelementige Cluster	mehrelementige Cluster	mehrelementige Cluster
		ohne Jokerzeichen	ohne Jokerzeichen	mit Jokerzeichen	mit Jokerzeichen
		Wortbegrenzung	ohne Wortbegrenzung	Wortbegrenzung	ohne Wortbegrenzung
1	*send* (senden)	*send*	-	-	-
	resend (wieder senden)	-	-	-	*~C* (falsch)*
	re_send (wieder senden)	*send*	-	-	-
	grave (gravieren)	*1grave / 2grave*	-	-	-
	engrave (eingravieren)	*en_2grave*	-	-	-
	en_grave (eingravieren)	*en_2grave*	-	-	-
2.1					
	sling (schleudern)	-	*~ling*	-	-
2.2					
	tan (bräunen)	-	-	*#CVn*	-
	suntan (sonnenbräunen)	-	-	-	*~C* (falsch)*
	sun_tan (sonnenbräunen)	-	-	*#CVn*	-
3					
	press (pressen)	-	*~ss*	-	-
	compress (komprimieren)	-	*~ss*	-	-
	com_press (komprimieren)	-	*~ss*	-	-
4.1					
	frolic (scherzen)	-	-	-	*~CVc*
5					
	present (präsentieren)	-	-	-	*~C*
	represent (repräsentieren)	-	-	-	*~C*
	re_present (repräsentieren)	-	-	-	*~C*

*** nicht vom Algorithmus her falsch, sondern linguistisch (siehe 3.3.4)**

Tabelle 3.03: Testfälle für den ersten Schritt der Synthese (Ausgabe: Regellexeme)

In der linken Spalte stehen die Suchlexeme, rechts davon die gefundenen Regellexeme, klassifiziert nach Tabelle 3.02.

Varianten:
Die Ausgabe von mehreren Regellexemen als Suchresultat ist möglich, wie der Fall *grave* zeigt. Wie in 3.2.1 beschrieben, gibt es Lexeme unterschiedlicher Flexion mit gleicher lexikalischer Grundform sowohl mit als auch ohne Bedeutungsunterschied. Diese Varianten sind jeweils als eigenständige Regellexeme in den Registern vorhanden (unterschieden durch Ziffern oder kleine griechische Buchstaben) und werden daher in ihrer Gesamtheit als Resultat angezeigt, so dass der Benutzer die Möglichkeit besitzt, diese ggf. in ihrer Semantik zu unterscheiden.
Desweiteren können sich präfigierte Verben als Regellexeme in den Registern finden. Bei präfigierten Suchlexemen ist die Eingabe mit Präfixmarker empfohlen, da es sonst zu falschen Suchresultaten, wie in den Fällen *resend* und *suntan*, kommen kann. Die Präfixkennzeichnung für den Algorithmus wird in Form eines Unterstrichs nach dem Präfix realisiert.

Präfigierte Lexeme:
Erster Fall: als Regellexeme im Register nicht vorhanden.
In den beiden Beispielen (*resend / suntan)* befindet sich das präfigierte Verb selbst nicht als Regellexem im Register, sondern lediglich das Simplex in orthographischer Form für *send* bzw. Jokerzeichenform für *tan,* welches hier das richtige Suchresultat wäre. Ohne eine Kennzeichnung der Präfixe findet der Algorithmus kein homC mit Wortbegrenzung und wird daher versuchen, die Suchlexeme soweit zu reduzieren, bis er ein homC, basC2 oder basC mit rückläufiger Übereinstimmung findet. In den beiden Beispielen findet der Algorithmus fälschlicherweise das basC *~C* für regelmäßig flektierende Lexeme. Erfolgt eine Kennzeichnung des Präfixes durch einen nachgestellten Unterstrich, dann findet der Algorithmus (wie in Tabelle 3.03 ersichtlich) richtigerweise jeweils das dazugehörige homogene Cluster.

Präfigierte Lexeme:
Zweiter Fall: als Regellexeme im Register vorhanden.
Solche Lexeme werden im Register immer richtig gefunden, egal ob das Suchlexem mit oder ohne Präfixmarkierung eingegeben wird.

Die nachfolgende Abbildung ist eine kompakte Darstellung der wesentlichen Testfälle aus Tabelle 3.03 und kennzeichnet die Bereiche, welche bei morphologisch präfigierten Lexemen ohne Präfixmarkierung zu Falschresultaten führen können.

Clustertypen	homC	homC / basC / basC2	basC / basC2	homC / basC / basC2
Regellexeme / Suchlexeme	Vollstring	Teil- / Vollstring	Vollstring	Teilstring
	einelementige Cluster	mehrelementige Cluster	mehrelementige Cluster	mehrelementige Cluster
	ohne Jokerzeichen	ohne Jokerzeichen	mit Jokerzeichen	mit Jokerzeichen
	Wortbegrenzung	ohne Wortbegrenzung	Wortbegrenzung	ohne Wortbegrenzung
Suchwort morphologisch nicht präfigiert	*send*	*press*	*tan*	*present*
Suchwort morphologisch präfigiert, formal nicht präfigiert (ohne Präfixmarker)	*engrave*	*compress*	-	*represent resend, suntan /* "mögliche Falschresultate"
Suchwort morphologisch präfigiert, formal präfigiert (mit Präfixmarker)	*re_send en_grave*	*com_press*	*sun_tan*	*re_present*

Tabelle 3.04: Kompakte Darstellung der Testfälle (Eingabe: Suchlexeme)

Diese Testfälle motivieren die nachfolgende formale Darstellung (Tabelle 3.05). In dieser Tabelle werden sämtliche Suchlexemvarianten den möglichen Resultaten bestehend aus Regellexemen gegenübergestellt. Dabei können mehrere Suchlexeme nicht gleichzeitig, sondern nur nacheinander eingegeben werden. Desweiteren sind (wie beschrieben) Falschresultate möglich.
Die Suchbaumhierarchie (Abb. 3.09) und diese formale Darstellung sind der Ausgangspunkt zur Entwicklung des Algorithmus und dienen später zu dessen Korrektheitsprüfung.

Clustertypen	homC	homC / basC / basC2	basC / basC2	homC / basC / basC2
Regellexeme / Suchlexeme	Vollstring	Teil- / Vollstring	Vollstring	Teilstring
	einelementige Cluster	mehrelementige Cluster	mehrelementige Cluster	mehrelementige Cluster
	ohne Jokerzeichen	ohne Jokerzeichen	mit Jokerzeichen	mit Jokerzeichen
	Wortbegrenzung	ohne Wortbegrenzung	Wortbegrenzung	ohne Wortbegrenzung
Suchwort morphologisch nicht präfigiert	☑	☑	☑	☑
Suchwort morphologisch präfigiert, formal nicht präfigiert (ohne Präfixmarker)	☑	☑	-	☑
Suchwort morphologisch präfigiert, formal präfigiert (mit Präfixmarker)	☑	☑	☑	☑

Tabelle 3.05: Formale Darstellung der vorkommenden Testfälle ohne Bewertung des Resultats

3.3.2 Definition des Inputs

Der Input des Algorithmus setzt sich (siehe Abb. 3.01) aus zwei Teilen zusammen, dem Register und einem zufällig gewählten Suchlexem (und ggf. einer Alphabettabelle).

3.3.3 Definition des Suchlexems

Für das Suchlexem[8] sind nachfolgende Varianten zulässig. Suchlexeme sind mit und ohne Präfixe erlaubt. Präfixe müssen, wie in 3.3.1.5 beschrieben, bei der Eingabe durch einen Unterstrich nach dem Präfix gekennzeichnet werden. Die Markierung von mehreren Präfixen bei einem Suchlexem ist möglich (zur Behandlung von präfigierten Suchlexemen siehe 3.3.6 und 3.3.7).

3.3.4 Funktion des Algorithmus

Das akzeptierte Suchlexem wird an eine Do-While-Schleife übergeben. Diese übergibt im ersten Schritt das komplette Suchlexem, beim wiederholten Aufruf das veränderte (um ein Präfix reduzierte) Suchlexem, an die "total length compare"-Funktion (3.3.5). Diese überprüft, ob in der Spalte "Synthetic lexical base" des Registers ein dem Suchlexem identisches Regellexem existiert. Wenn dies der Fall ist, terminiert die Schleife. Andernfalls wird überprüft, ob das Suchlexem einen weiteren Präfixmarker besitzt.
Besitzt das Suchlexem keinen Präfixmarker, terminiert die Schleife ebenfalls. Andernfalls wird das Suchlexem an die Funktion "prefix cut" (3.3.6) übergeben, welche das Suchlexem um ihr erstes Präfix reduziert, und die Schleife beginnt von vorn.

Wenn die Schleife terminiert, bedeutet dies entweder:

- Es befindet sich das ganze oder "reduzierte" Suchlexem im Register
- Das "reduzierte" Suchlexem befindet sich nicht im Register und besitzt auch keinen Präfixmarker.

Wenn sich das Suchlexem im Register befindet, werden die zugehörigen Regellexeme ausgegeben, andernfalls wird das "reduzierte" Suchlexem an die "longest match"-Funktion (3.3.7) übergeben.
Diese generiert die Suchalternativen aus 3.3.1.3 und überprüft, ob sich eine davon im Register findet. Die Suchalternativen und das Suchlexem werden von links buchstabenweise verkürzt, bis eine Übereinstimmung gefunden wird.

Werden beispielsweise die Regellexeme für das Verb *reengrave* gesucht, so sind vier Eingabevarianten zulässig und liefern folgende Regellexeme mit ihren Schlüsselformen als Ergebnismenge:

- *reengrave* → basC2 *Ce* : *~Ce, ~Ced, ~Ced*
- *re_engrave → engrave* : *en_2grave, engraved, engraved*
- *re_en_grave → engrave* : *en_2grave, engraved, engraved*
- *reen_grave → grave* : *1grave, graved, graved*
  α +2grave, graved, graved
  β +2grave, graved, graven

Nur das zweite und dritte Resultat ist linguistisch richtig. Das erste Ergebnis wird von "longest match" gefunden, die restlichen von "total length compare" nach "prefix cut".

[8] In dieser Untersuchung steht das Suchlexem (search lexeme) entweder für ein Eingabewort in seiner ursprünglichen Form, d.h. wie es durch einen Benutzer eingegeben wird, oder für ein Eingabewort, welches durch den Algorithmus um sein Präfix reduziert wurde.

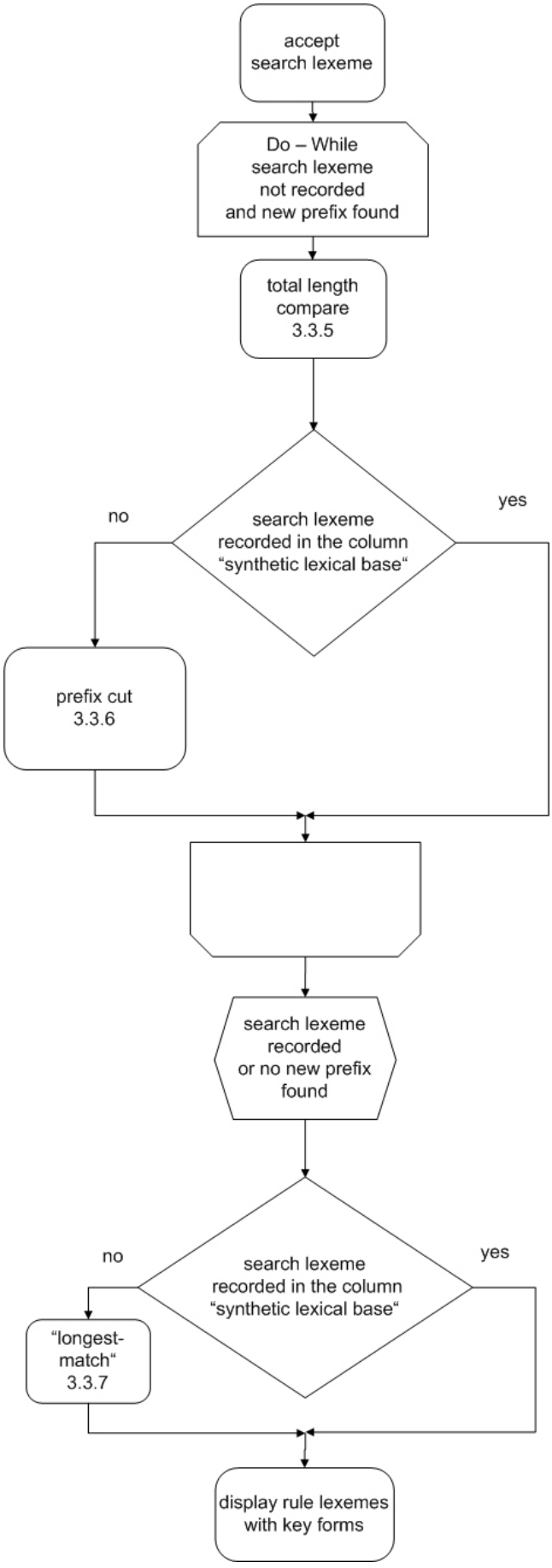


Abbildung 3.11: Ablaufplan des SMIRT-Algorithmus

3.3.5 Funktion “total length compare“

In der Vergleichsspalte “Synthetic lexical base“ existieren keine Präfixmarkierungen. Daher kann ein formal präfigiertes Suchlexem nur dann in voller Länge im Register gefunden werden, wenn im Suchlexem die Präfixmarkierungen entfernt werden. Das geschieht in dieser Funktion.
Das (ggf. um Präfixe reduzierte) Suchlexem ohne Präfixmarkierung bildet den Input der “total length compare“-Funktion. Diese Funktion sucht im Register in der Tabellenspalte “Synthetic lexical base“, ob sich das Suchlexem dort in voller Länge findet (auch mehrere gefundene Regellexeme sind möglich).

Beispiel für mehrere gefundene Regellexeme:

Wie in 3.2.1 beschrieben, handelt es sich im Fall *grave* um morphologische Varianten mit Bedeutungsunterschied.

grave → grave: 1grave, graved, graved (‘clean a ship's bottom‘)
α +2grave, graved, graved
β +2grave, graved, graven (‘engrave an inscription on a surface‘)

Wird ein Regellexem gefunden, ist es gleich dem (ggf. um Präfixe reduzierten) Suchlexem. Es handelt sich um den Fall 1 in Abb. 3.09 und Tab. 3.03:

- **Vollstring, einelementiges Cluster, ohne Jokerzeichen, Wortbegrenzung**
  Beispiel *send* mit seinen Schlüsselformen *send, sent, sent*

Andernfalls passiert nichts. In beiden Fällen bleibt das Suchlexem unverändert.

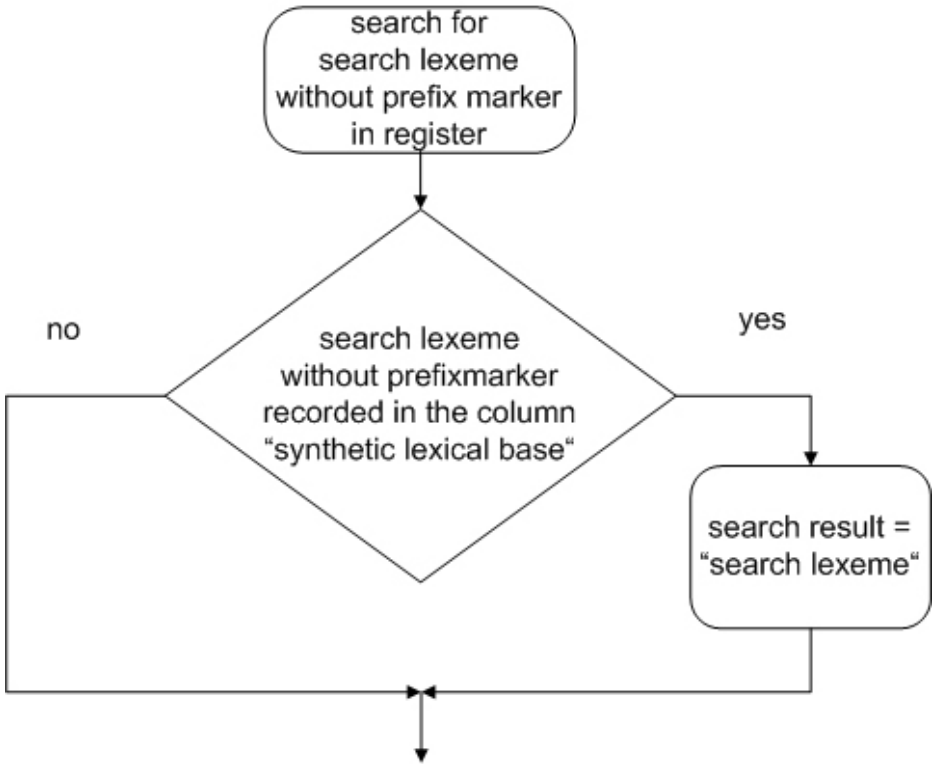


Abbildung 3.12: Ablaufplan der “total length compare“-Funktion

3.3.6 Funktion “prefix cut“

In der Funktion “prefix cut“ bildet beim ersten Aufruf das akzeptierte Suchlexem den Input, bei einem wiederholten Aufruf ein um ein Präfix reduziertes Suchlexem. Der Input wird an eine Do-While-Schleife übergeben. Diese übergibt den Input an eine Suchfunktion, welche Buchstabe für Buchstabe in Leserichtung (europäisch) nach einem Unterstrich im Suchlexem sucht. Wird ein Unterstrich gefunden, wird ein Flag gesetzt, das in der Bedingung “search lexeme not recorded and new prefix found“ (Abb. 3.11) abgefragt wird. Alle sich rechts vom Unter-strich befindlichen Buchstaben bilden das neue Suchlexem, und die Suche wird fortgesetzt.
Durch diese Vorgehensweise werden auch Mehrfach-Präfixe erkannt und reduziert, wie das nachfolgende Beispiel verdeutlicht:

- *re_send* reduziert auf *send* bildet das neue Suchlexem
- *re_en_grave*
  - nach dem ersten Durchlauf:
    reduziert auf *en_grave* bildet das neue Suchlexem
  - *en_grave* nach dem zweiten Durchlauf:
    reduziert auf *grave* bildet das neue Suchlexem

Solange die Suchfunktion keinen Unterstrich findet, überprüft sie Buchstabe für Buchstabe, bis sie ans Wortende gelangt. Dann terminiert die Schleife.

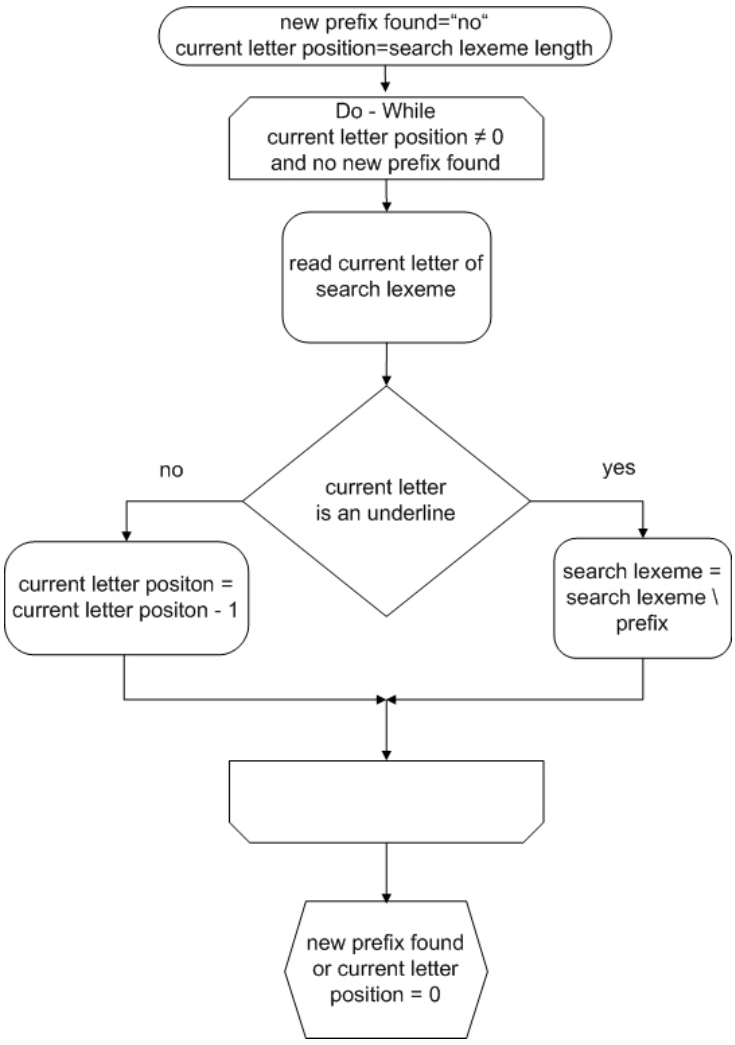


Abbildung 3.13: Ablaufplan der “prefix cut“-Funktion

3.3.7 Funktion “longest match“

Den Input der “longest match“-Funktion bildet das akzeptierte oder reduzierte Suchlexem.

3.3.7.1 Generierung der Suchalternativen (Abb. 3.14)

Im ersten Schritt werden zusätzlich zum Suchlexem weitere Suchalternativen generiert zum Vergleich mit den in den Registern vorzufindenden Vokal-Konsonant-Konstellationen, welche in 3.3.1.3 beschrieben wurden. Hierzu wird das Suchlexem auf dessen Buchstabenzusammensetzung folgendermaßen überprüft. Handelt es sich bei einem Buchstaben um einen Vokal, wird in der Suchalternative stellvertretend ein “V“ eingetragen, und wenn es sich um einen Konsonanten handelt, ein “C“. Diese Suchalternativen werden in Abhängigkeit von einer Alphabet-Tabelle des untersuchten Verbalsystems gebildet, welche dem Algorithmus zu Beginn zur Verfügung gestellt (importiert) werden muss (siehe Abbildung 3.10). Als Erstes wird die Länge des Suchlexems ermittelt. Diese legt die maximale Anzahl der Schleifendurchläufe fest und bestimmt die Anzahl der zu generierenden Suchalternativen. Suchalternativen werden erst ab einer Suchlexemlänge größer eins generiert. Andernfalls ist eine Suchalternative nicht nötig, da einbuchstabige Suchlexeme entweder durch das reduzierte Suchlexem selbst oder durch eine Bedingung am Ende der “longest match“-Funktion abgefangen werden (Suchalternative 4 exit). Ab einer Länge des Suchlexems von größer eins wird nur die Suchalternative 1, ab einer Länge größer zwei die Suchalternativen 1-2 und ab einer Länge von größer drei die Suchalternativen 1-3 generiert (vgl. Abb. 3.10).

Beispiel:
Suchlexem: *tan*
#-Suchalternative 1: #*CVn*
Suchalternative 1: *CVn*
Suchalternative 2: *Can*
Suchalternative 3: wird nicht generiert, da sie gleich dem Suchlexem wäre (=*tan*)
Suchalternative 4 (exit): *C*

Bei jedem Schleifendurchlauf wird in Leserichtung Buchstabe für Buchstabe durch ein Jokerzeichen ersetzt und die Buchstabenposition (beginnend mit der Suchlexemlänge) um eins dekrementiert. Fällt der Wert der Buchstabenposition unter vier, endet die Generierung der Alternative 3, ab einem Wert von kleiner drei die Generierung der Alternative 2, unter einem Wert von zwei, d.h. Buchstabenposition = 1, die Generierung der Alternative 1, und die Schleife terminiert. Danach werden die Suchalternativen mit den letzten Buchstaben des Suchlexems aufgefüllt.

3.3.7.2 Überprüfung der Suchalternativen (Abb. 3.15)

Im ersten Match-Schritt erfolgt eine Überprüfung, ob sich die #-Suchalternative 1 im Register befindet. Ist dies der Fall, wird die Funktion beendet, und die dazugehörigen Regellexeme werden ausgegeben.
Ansonsten werden das reduzierte Suchlexem und die Suchalternativen 4.3, 4.2 und 4.1 (sofern diese erzeugt wurden, andernfalls eine leere Alternative) an eine Do-While-Schleife[9] übergeben. Diese reduziert in Leserichtung (europäisch) sowohl das Suchlexem als auch dessen Suchalternativen Buchstabe um Buchstabe und überprüft in jedem Schleifendurchlauf, ob sich die verbleibende Buchstabenfolge in der Spalte “Synthetic lexical base“ im Register findet. Diese Überprüfung erfolgt nach der Reihenfolge der Bedingungshierarchie, d.h. reduziertes Suchlexem, reduzierte Alternative 4.3, dann 4.2 und endlich 4.1. Ergibt sich eine Suchüberein-

[9] Im ersten Durchlauf der Schleife wird das Suchlexem (hier noch in voller Länge) nicht überprüft, da dies schon die Funktion “total length compare“ getan hat (length of search lexeme ≠ original length of search lexeme).

stimmung, wird nur jenes Regellexem, welches eine Kennzeichnung (homC, basC2, basC u.a.) in der Clusterspalte besitzt, ausgegeben, und der Algorithmus terminiert.

Folgende Strukturen sind hierbei möglich:

#-Suchalternative 1 (2.2 aus Abb. 3.09 und Tab. 3.03)

- **Vollstring, mehrelementiges Cluster, mit Jokerzeichen, Wortbegrenzung**
  Beispiel *tan* → basC2 *#CVn: #CVn, #CVnned, #CVnned*
  also *tan, tanned, tanned*

Reduziertes Suchlexem (3)

- **Teil- / Vollstring, mehrelementiges Cluster, ohne Jokerzeichen, ohne Wortbegrenzung**
  Beispiel *search* → homC *ch: ~ch, ~ched, ~ched*
  also *search, searched, searched*

Reduzierte Suchalternative 3 (4.3)

- **Teilstring, mehrelementiges Cluster, mit Jokerzeichen, ohne Wortbegrenzung**
  Im Fall des Suchlexems *present* ergibt sich im ersten Schleifendurchlauf die Suchvariante *CCVCent*, welche auch durch Reduktion zu keiner Übereinstimmung mit Regellexemen führt.

Reduzierte Suchalternative 2 (4.2)

- **Teilstring, mehrelementiges Cluster, mit Jokerzeichen, ohne Wortbegrenzung**
  Im Fall des Suchlexems *present* ergibt sich im ersten Schleifendurchlauf die Suchvariante *CCVCVnt*, welche auch durch Reduktion zu keiner Übereinstimmung mit Regellexemen führt.

Reduzierte Suchalternative 1 (4.1)

- **Teilstring, mehrelementiges Cluster, mit Jokerzeichen, ohne Wortbegrenzung**
  Beispiel *frolic* → basC2 *CVc: ~CVc, ~CVcked, ~CVcked*

Wurde auch nach Überprüfen der dritten “Suchalternative“ keine Übereinstimmung in den Registern gefunden, so werden das Suchlexem wie auch die Suchalternativen in Leserichtung (europäisch) um einen Buchstaben verkürzt, die Suchlexemlänge um eins dekrementiert und die Schleife beginnt von vorn.
Bei erfolgloser Suche, d.h. wenn die Suchlexemlänge den Wert 0 erreicht hat, terminiert die Schleife endgültig. Nach dem Schleifenende wird eine letzte Suche initiiert, bestehend aus dem durch Jokerzeichen ersetzten letzten Buchstaben des Suchlexems (Suchalternative 4 exit).

Suchalternative 4 exit (5)

- **Teilstring, Cluster, mit Jokerzeichen, ohne Wortbegrenzung**
  Beispiel *present* → basC *C: ~C, ~Ced, ~Ced*

Führt auch diese Überprüfung zu keinem Resultat, d.h. befindet sich weder das Suchlexem noch eine rückläufige Übereinstimmung als Regellexem im Register des Verbalsystems, terminiert der Algorithmus, und es erscheint die Meldung, dass sich kein passender Eintrag zum eingegebenen Suchlexem im Register dieser Sprache findet.

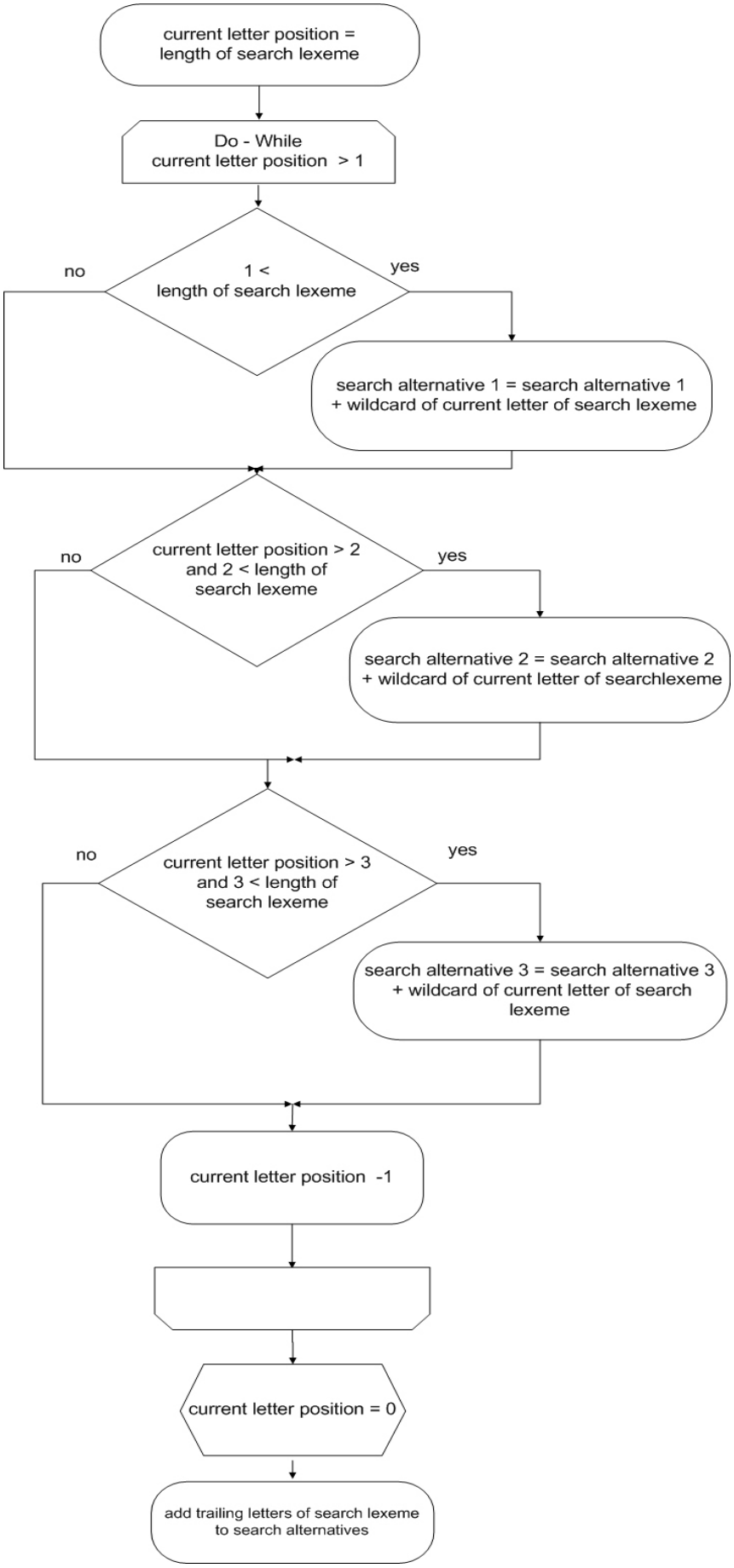


Abbildung 3.14: Ablaufplan der “longest match“-Funktion – Generierung der Suchalternativen

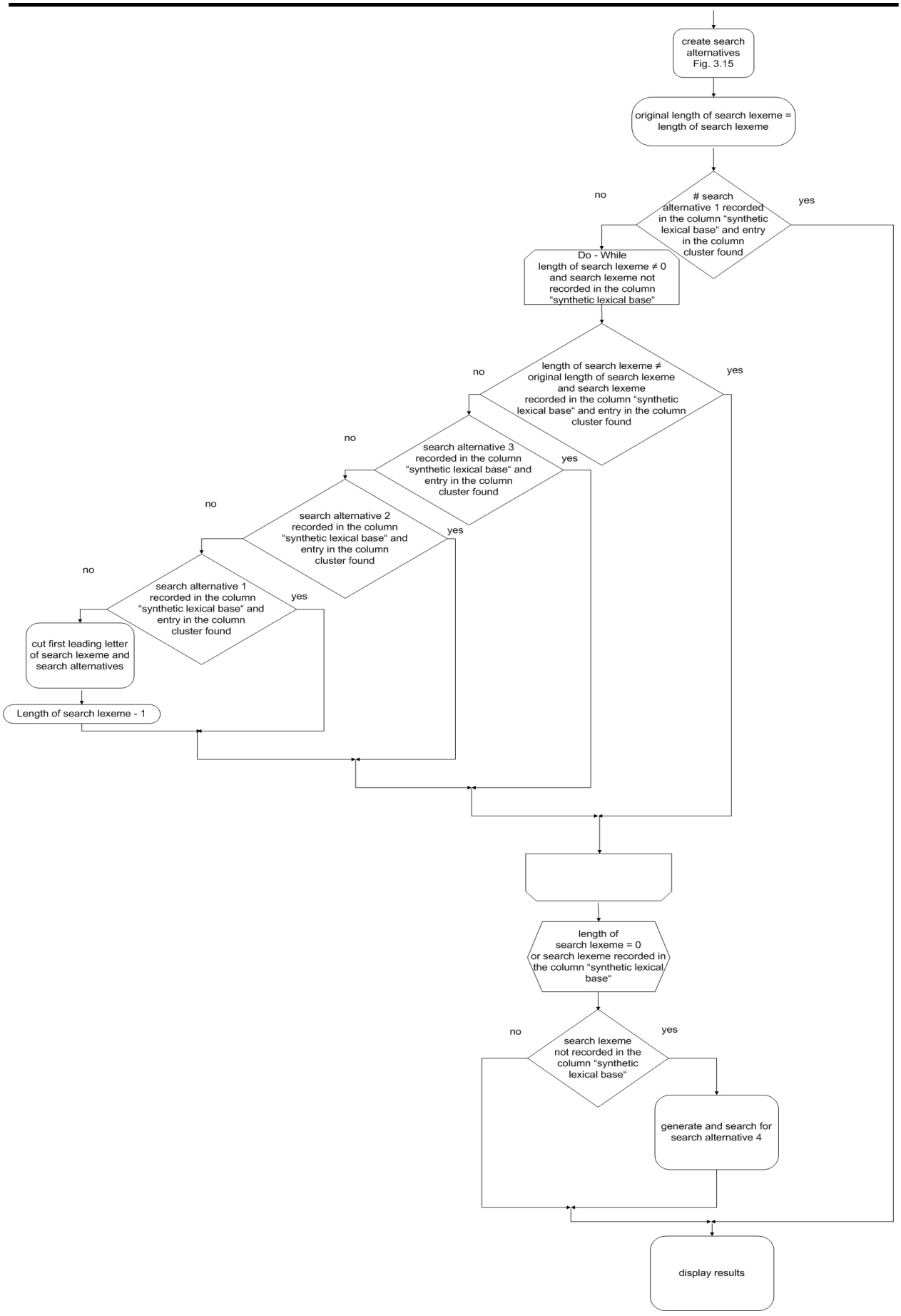


Abbildung 3.15: Ablaufplan der "longest match"-Funktion – Überprüfung der Suchalternativen

3.4 Konsequenzen für die Gestaltung der Register

Der Input des Algorithmus besteht aus drei Teilen, einem zufälligen Suchlexem, einer sprachabhängigen Alphabettabelle und dem entsprechenden Register aus dem Data-Mining-Prozess. Sowohl der erste als auch der zweite Schritt des synthetisch-generativen Teilbereichs aus 3.1.2 und 3.1.3 verdeutlichen, dass im Register einige Konventionen erfüllt sein müssen, um die korrekte Funktionsweise des SMIRT-Algorithmus gewährleisten zu können. In der nachfolgenden Tabelle 3.06 sind diese Konventionen aufgelistet und um einige weitere ergänzt, welche für die spätere technische Umsetzung ebenfalls von Bedeutung sein werden. Durch die Festlegung der Konventionen sowie eine exakte Definition der Eingangsgrößen ist es gelungen, den Algorithmus so zu formalisieren, dass dieser für die Register unterschiedlicher Verbalsysteme einwandfrei funktioniert und zu einem zufälligen Suchlexem die dazugehörigen Regellexeme findet.

Nr.	Konvention	Beschreibung
01	Jokerzeichen für Vokale und Konsonanten	Damit für ein zu untersuchendes Verbalsystem die entsprechenden Kombinationen aus Konsonant- / Vokal-Jokerzeichen durch den Algorithmus identifiziert werden können, benötigt dieser eine Zuordnungstabelle. In dieser sprachabhängigen Alphabet-Tabelle ist jedem Buchstaben des Alphabets ein "V" für Vokal oder ein "C" für Konsonant zugeordnet.
02	Anzahl sprachabhängiger Spalten	Der sprachabhängige Teil eines Registers kann in der Anzahl der Spalten variieren. Zur Vereinheitlichung und Übersichtlichkeit des Outputs wird der Import eines Registers auf die ersten 17 Spalten (inkl. sprachabhängigem Teil) begrenzt.
03	Bereinigung der Synthetic Lexical Base	Die "Synthetic Lexical Base"-Spalte ist die vom Algorithmus zu untersuchende Spalte. Suchlexem bzw. Suchalternativen werden mit den Registereinträgen dieser Spalte verglichen. Daher muss diese Spalte von numerischen, Sonderzeichen (ausgenommen dem #-Zeichen) und Blanks bereinigt werden.
04	Sprachabhängige Benennung der Registerspalten berücksichtigen	Die Benennung der Registerspalten erfolgt entsprechend den Schlüsselformen des zu untersuchenden Verbalsystems. Die Ausgabe der Regellexeme des jeweiligen Verbalsystems soll dies mitberücksichtigen und die Ausgabespalten entsprechend benennen.
05	Reihenfolge der Registerspalten	Die Suche erfolgt in der Spalte "Synthetic Lexical Base", während die Cluster-Spalte zur Verifizierung in der "longest match"-Funktion benötigt wird. Es ist erforderlich, dass diese beiden Spalten immer an derselben Position im Register zu finden sind. Die Einträge für die Cluster werden in der ersten, der Flexionstyp in der zweiten und die Einträge für Synthetic Lexical Base in der dritten Spalte des Registers erwartet.
06	Anzahl Jokerzeichen	Die Suchalternativen setzen sich aus Konsonant- / Vokalkombinationen plus den letzten Buchstaben des Suchwortes zusammen. Die Anzahl dieser letzten Buchstaben wird auf maximal drei begrenzt. Einträge in "Synthetic Lexical Base" müssen diese Struktur berücksichtigen.
07	Flexionsvarianten	Wie am Beispiel *grave* deutlich wird, sind Mehrfachresultate zulässig, was bei einem doppelten Eintrag der Regellexeme in den Regelsätzen vorkommt.
08	Klammerung	Keine Klammerung in der Spalte "Synthetic Lexical Base". Jeder Eintrag muss in einer eigenen Zeile eingetragen werden. Beispiel: *#C(C)Vn → #CCCVn, #CCVn und #CVn*

Tabelle 3.06: Konventionen für den Input

4. Implementierung des synthetisch-generativen Algorithmus

4.1 stellt das verwendete Implementierungskonzept vor. 4.2 erklärt die technische Umsetzung der in 3.3 entworfenen Funktionen.

4.1 Art der Implementierung

4.1.1 erklärt allgemein das verwendete Implementierungskonzept, dessen Aufbau und Struktur und den sich daraus ergebenden Nutzen und die Vorteile. 4.1.2 erläutert die technische Umsetzung dieses Konzepts in einem ersten Prototyp, die ausgewählten Komponenten und verwendeten Programmiersprachen.

4.1.1 Programmaufbau

Die Umsetzung des SMIRT-Algorithmus in einem ersten Prototyp (SMIRT Ver. 1.0) erfolgt auf Basis des MVC-Konzeptes (Model View Controller)[10]

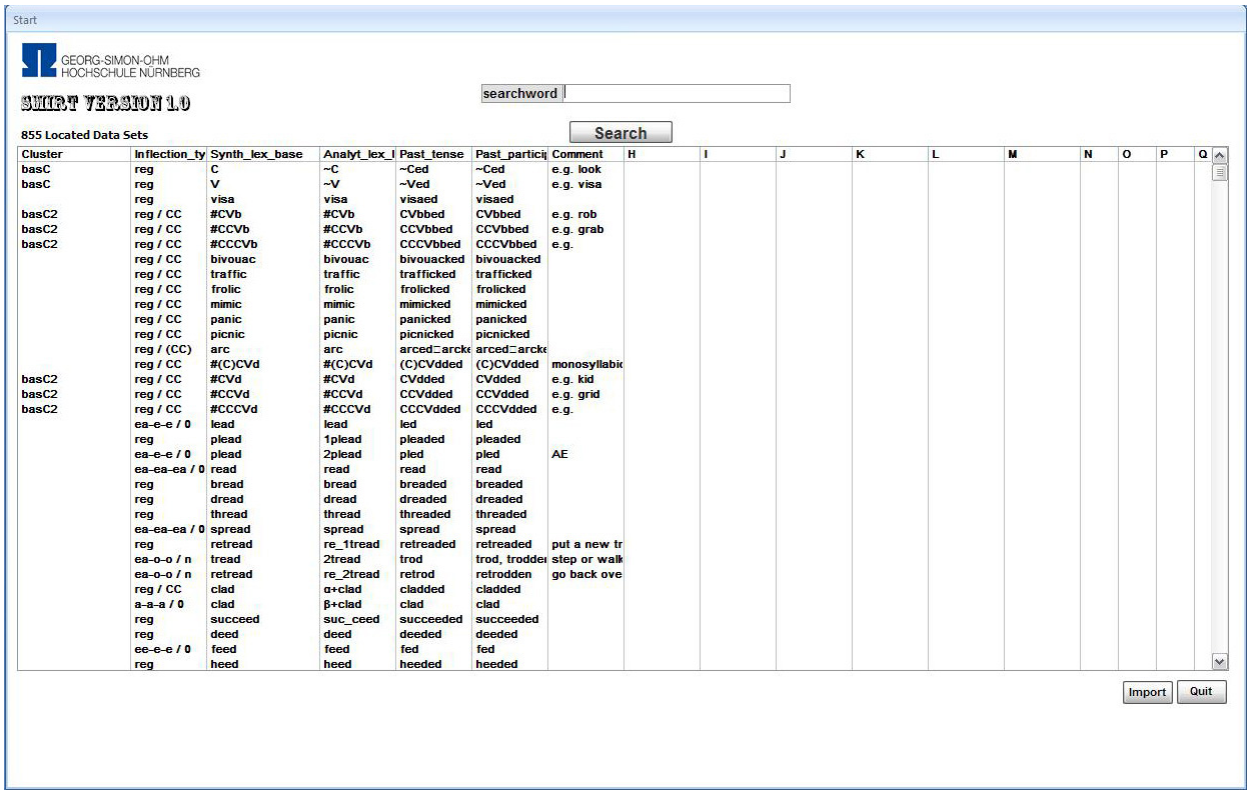


Abbildung 4.01: Benutzeroberfläche (View) Ver.1.0

[10] MVC-Konzept (vgl. Middendorf, Singer, Heid 2002)

Das Model-View-Controller-Konzept wird in vielen Bereichen moderner Softwareentwicklung eingesetzt und bedeutet die strikte Aufgabenverteilung bei einer Anwendung.
So wird als **Model** die Datenquelle bezeichnet, die die Daten unabhängig vom Design des Erscheinungsbildes der Benutzeroberfläche liefert (also beispielsweise aus einer relationalen Datenbank).
Das **View** zeigt diese Daten dann in passender Art und Weise an – bestimmt dadurch den »Look«. Wie das View die Daten anzeigt, wird nicht vom Model beeinflusst. Der **Controller** kümmert sich um die Interaktion mit dem Benutzer. Der Controller ist also die Logik der Anwendung. Der Vorteil dieser Aufgabenverteilung ist einerseits die Möglichkeit der Aufteilung in logische, unabhängige Bereiche, andererseits die Möglichkeit, jeden der drei Teile jederzeit auszutauschen.

4.1.2 Einsatz und Verwendung von Programmiersprachen

Die technische Umsetzung dieses Konzeptes erfolgt durch Microsoft Access. Für die Darstellung (Look) der Benutzeroberfläche (**View**) wird die integrierte Designumgebung von Access verwendet. Die Funktionalität der Buttons und Auswahlmenüs der Oberfläche ist durch die Programmiersprache VBA (Visual Basic for Application) realisiert. Die Kommunikation und Manipulation (**Controller**) der Daten erfährt ihre Umsetzung durch VBA als auch durch SQL (Structured Query Language). Die Tabellen der Datenbank (**Model**) stehen in keinerlei Relation zu einander, sie dienen in diesem Prototyp lediglich als Datenspeicher (Speicherung der Regellexeme, des Sprachalphabets und Speicherung der spachabhängigen Spaltenüberschriften). Das Suchlexem sowie dessen Suchalternativen (wie in 3.3.1 beschrieben) wird in VBA in String-Variablen gespeichert und den entsprechenden Funktionen übergeben. Die zur Realisierung verwendeten Do-While-Schleifen von Microsoft Access sind kopfgesteuert und somit abweisende Schleifen.

4.2 Realisierung der Funktionen

4.2.1 erläutert das Hauptprogramm. 4.2.2, 4.2.3 und 4.2.4 gehen auf die Umsetzung der Funktionen “total length compare“, “prefix cut“ und “longest match“ ein. 4.2.5 erklärt in kurzen Zügen die Zusatzfunktion für den Datenimport.

Die Realisierung der Funktionen erfolgt in Form einer eigenständigen Kapselung, d.h. sie sind in sich geschlossen und vom aufrufenden Programmteil getrennt. Dies ermöglicht einen gezielten Aufruf (Instanziierung), eine modulare Anpassung sowie die Wiederverwendbarkeit. In den Funktionen werden sowohl lokale als auch globale Variablen verwendet. Die lokalen Variablen sind nur innerhalb der Funktionen existent, während die globalen Variablen funktionsübergreifend verwendet werden.

4.2.1 Hauptprogramm

Name	Typ	Ort	Verwendung
F1 bis F17	String	global	In diesen Variablen werden während des Programmablaufs die Spaltenüberschriften des jeweiligen Verbregisters gespeichert, welche auch bei der Ausgabe der Regellexeme als Spaltenüberschriften verwendet und durch die Einträge aus der Tabelle “Head“ befüllt werden (s. u. in dieser Tabelle).
Flag_Match_Found	Integer	global	Zeigt an, ob eine Übereinstimmung des (reduzierten) Suchlexems mit einem Regellexem gefunden wurde. 0: keine Übereinstimmung gefunden 1: Übereinstimmung gefunden
Flag_Prefix_Found	Integer	global	Zeigt an, ob im Suchlexem ein Präfix gefunden wurde. 0: kein Präfix gefunden 1: Präfix gefunden
Searchlexeme	String	global	Speichert das durch den Benutzer eingegebene und ggf. im Programmverlauf bei einer Präfixmarkierung von “prefix cut“ verkürzte Suchlexem.
Que_Rules	SQL Abfrage	global	Que_Rules ist eine SQL-Text-Variable in ACCESS. Diese Variable wird im Laufe des Programms so lange verändert, bis sie den Text (SQL-String) der endgültigen, für die Ausgabe des Suchresultats relevanten SQL-Abfrage enthält. Diese SQL-Abfrage wird im Laufe des Programms nur an einer einzigen Stelle ausgeführt, nämlich am Programmende durch die Zuweisung (Me.Rule_Lexemes.RowSource="Que_Rules") an das Listenfeld. Diese Abfrage ist verantwortlich für die Ausgabe der Regellexeme und ist eine Selektion über alle 17 Spalten der Tabelle “Rules“.
User_Input_Of_ Searchlexeme	Eingabefeld	global	Name für das Eingabefeld des Suchlexems, definiert in der Designumgebung von ACCESS.
Rule_Lexemes	Listenfeld	global	Name für das Ausgabelistenfeld der Regellexeme, definiert in der Designumgebung von ACCESS.
Alphabet	Tabelle	global	Verwaltet das Alphabet des untersuchten Verbalsystems mit der Kennzeichnung, ob es sich bei einem Buchstaben um einen Vokal oder einen Konsonanten handelt.
Head	Tabelle	global	Die Spaltenüberschriften des untersuchten Verbalsystems sind hier hinterlegt.
Rules	Tabelle	global	Verwaltet die Regellexeme des untersuchten Verbalsystems, welche mittels Datenimport hinterlegt wurden; siehe 4.2.5.

Tabelle 4.01: Datenlexikon des SMIRT-Algorithmus

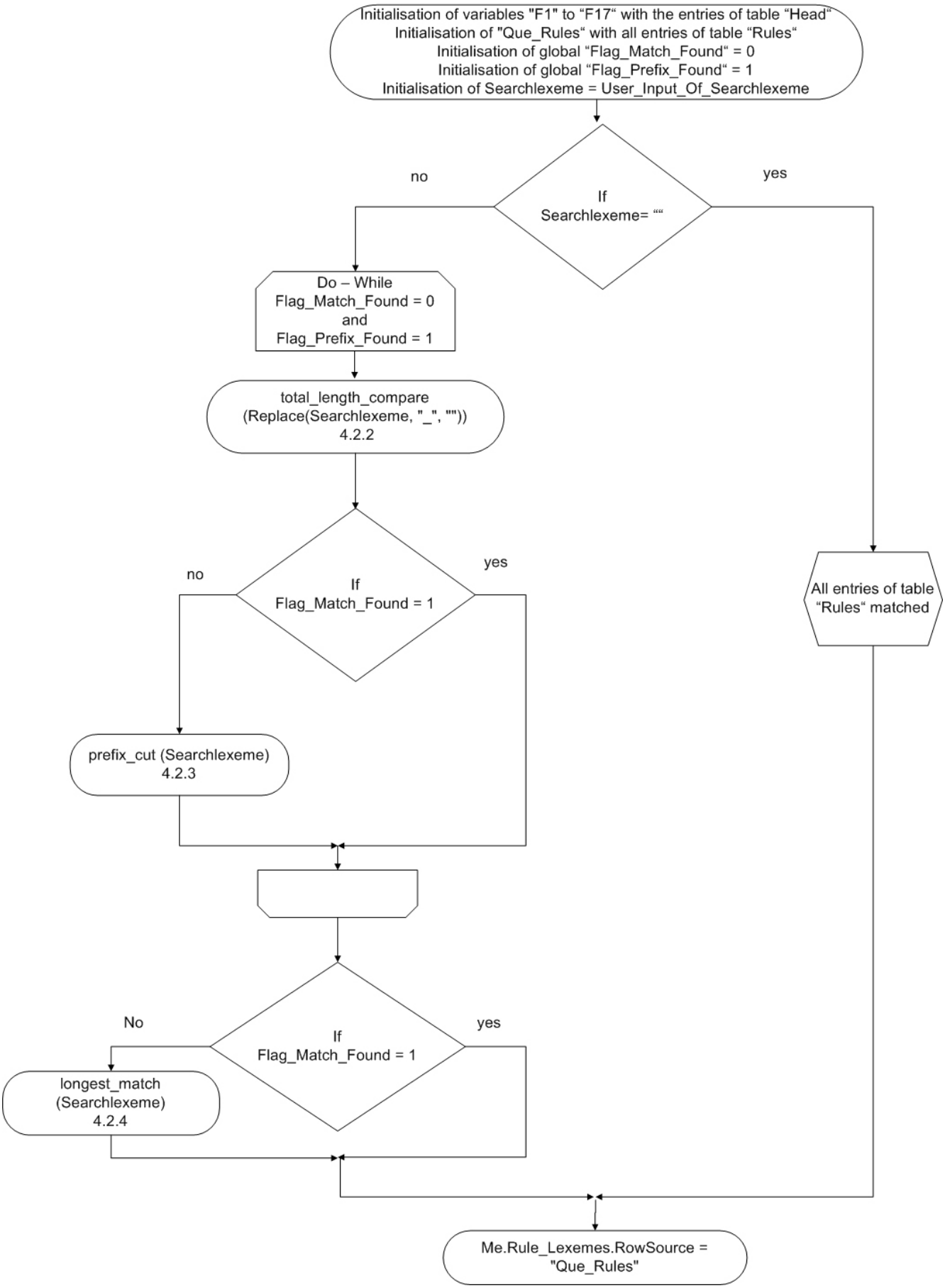


Abbildung 4.02: Ablaufplan der Implementierung des SMIRT-Algorithmus

Die Main-Funktion, welche durch das Betätigen des “Search-Buttons“ der Benutzeroberfläche aufgerufen wird, beginnt mit der Initialisierung der Variablen “F1“ bis “F17“ und weist diesen die entsprechenden Einträge aus der Tabelle “Head“ zu. Beim Programmstart sowie bei einer Leereingabe durch den Benutzer sollen alle sich in der Tabelle “Rules“ befindlichen Regellexeme als Resultat ausgegeben werden; daher wird die SQL-Abfrage “Que_Rules“ mit der Abfrage initialisiert, die alle Einträge der Tabelle “Rules“ liefert. Dabei werden die zuvor belegten Spaltenüberschriften in den Variablen “F1“ bis “F17“ verwendet, welche als Aliasnamen der Ausgabespalten benutzt werden. Die Variable “Flag_Match_Found“ erhält eine Vorbelegung mit dem Wert “0“ und “Flag_Prefix_Found“ mit “1“, die Variable “Searchlexeme“ wird mit dem durch den Benutzer eingegebenen Suchlexem initialisiert.

Eine If-Anweisung überprüft, ob die Variable “Searchlexeme“ einen Wert erhalten hat oder ob der “Search-Button“ ohne Eingabe ausgeführt wurde. Ist dies der Fall, werden (wie beschreiben) alle Regellexeme ausgegeben (dies entspricht der vorangegangenen Initialisierung der “Que_Rules“).

Besitzt die Variable “Searchlexeme“ einen Wert, wird eine Do-While-Schleife ausgeführt, die als Erstes die “total_length_compare“-Funktion aufruft und dieser das um sämtliche Unterstriche bereinigte Suchlexem übergibt[11]. Wird ein entsprechendes Regellexem gefunden, so erhält die Variable “Flag_Match_Found“ den Wert “1“. Andernfalls bleibt der Wert dieser Variablen unverändert.

Im weiteren Verlauf der Schleife überprüft eine If-Anweisung, ob die Variable “Flag_Match_Found“ den Wert “1“ erhalten hat. Trifft dies zu, terminiert die Schleife.
Wenn diese Bedingung nicht erfüllt ist, wird die “prefix_cut“-Funktion aufgerufen, welcher ebenfalls der Wert der Variablen “Searchlexeme“ übergeben wird. Diese überprüft, ob sich ein Unterstrich in dem übergebenen Suchlexem befindet. In diesem Fall wird das “Flag_Prefix_Found“ auf den Wert “1“ gesetzt, und dem “Searchlexeme“ wird der präfixreduzierte Wert (alle Buchstaben rechts vom Unterstrich) zugewiesen. Bleibt die Suche nach einem Unterstrich erfolglos, wird der Wert von “Processing_Lexeme“ nicht verändert, und das “Flag_Prefix_Found“ wird auf “0“ gesetzt, und die Schleife terminiert.

Nach der Schleife überprüft eine If-Anweisung, ob die Variable “Flag_Match_Found“ den Wert “1“ besitzt. Ist dies zutreffend, wird das Listenfeld der Ausgabe (Rule_Lexemes) mit der durch die “total_length_compare“-Funktion veränderten SQL-Abfrage “Que_Rules“ belegt. Ansonsten wird die “longest_match“-Funktion aufgerufen, die ihrerseits die SQL-Abfrage “Que_Rules“ verändert. Die Belegung des Listenfelds erfolgt erst nach deren Abarbeitung.

Hinweis:
In den Funktionen “total length compare“ und “longest match“ wird die Suche nach dem ggf. durch die “prefix cut“-Funktion veränderten Suchlexem durch eine Selektion über die ersten drei Spalten (F1-F3) der Tabelle “Rules“ durchgeführt. Das Ergebnis wird einer Variable vom Typ DAO.recordset übergeben, um mit Hilfe der VBA-Funktion “RecordCount“ festzustellen, ob zu diesem Suchlexem Regellexeme vorhanden sind. Die Ausgabe der tatsächlich gefundenen Regellexeme erfolgt durch eine andere Selektionsanweisung über alle 17 Spalten der Tabelle “Rules“, welche einer SQL-Text-Variablen namens “Que_Rules“ übergeben wird. Das Ergebnis der Abfrage dient als Datenbasis für die Ausgabe des Listenfeldes.

[11] Dies wird über die accessinterne Replace-Funktion erreicht, welche eine ausgewählte Zeichenfolge sucht und durch eine andere ersetzt, in diesem Fall den Unterstrich durch eine leere Zeichenfolge.

```
Private Sub Search_Button_Click()

F1 = DLookup("[F1]", "[Head]", "")
F2 = DLookup("[F2]", "[Head]", "")
F3 = DLookup("[F3]", "[Head]", "")
F4 = DLookup("[F4]", "[Head]", "")
F5 = DLookup("[F5]", "[Head]", "")
F6 = DLookup("[F6]", "[Head]", "")
F7 = DLookup("[F7]", "[Head]", "")
F8 = DLookup("[F8]", "[Head]", "")
F9 = DLookup("[F9]", "[Head]", "")
F10 = DLookup("[F10]", "[Head]", "")
F11 = DLookup("[F11]", "[Head]", "")
F12 = DLookup("[F12]", "[Head]", "")
F13 = DLookup("[F13]", "[Head]", "")
F14 = DLookup("[F14]", "[Head]", "")
F15 = DLookup("[F15]", "[Head]", "")
F16 = DLookup("[F16]", "[Head]", "")
F17 = DLookup("[F17]", "[Head]", "")

CurrentDb.QueryDefs("Que_Rules").sql = " SELECT Rules.F1 as [" & F1 & "], Rules.[F2] as [" & F2 & "], Rules.F3 as
[" & F3 & "], Rules.F4 as [" & F4 & "], Rules.F5 as [" & F5 & "], Rules.F6 as [" & F6 & "], Rules.F7 as [" & F7 & "], Ru-
les.F8 as [" & F8 & "], Rules.F9 as [" & F9 & "], Rules.F10 as [" & F10 & "], Rules.F11 as [" & F11 & "], Rules.F12 as
[" & F12 & "], Rules.F13 as [" & F13 & "], Rules.F14 as [" & F14 & "], Rules.F15 as [" & F15 & "],Rules.F16 as [" &
F16 & "], Rules.F17 as [" & F17 & "]" & _
" FROM Rules "
Flag_Match_Found = 0
Flag_Prefix_Found = 1
Searchlexeme = Nz(Me.User_Input_Of_Searchlexeme.Value)

If (Searchlexeme) = "" Then

Else

    Do While (Flag_Match_Found = 0 and Flag_Prefix_Found = 1)

      total_length_compare (Replace(Searchlexeme, "_", ""))

      If (Flag_Match_Found = 1) Then

         Else
           prefix_cut (Searchlexeme)
      End If

    Loop

    If Flag_Match_Found = 1 Then

    Else
        longest_match (Searchlexeme)

    End If
End If

Me.Rule_Lexemes.RowSource = "Que_Rules"

End Sub
```

Quellcode 4.01: Main-Funktion (Search_Button_Click)

4.2.2 Realisierung der Funktion “total length compare“

Name	Typ	Ort	Verwendung
rs	DAO.recordset	lokal	Speichert das Ergebnis einer SQL-SELECT-Anweisung. Mithilfe der in VBA für DAO.recordset-Variablen bereitgestellten Funktion “Recordcount“ lässt sich einfach feststellen, wie viele Einträge in der Variable gespeichert sind, d.h. ob der an die Variable übergebene SQL-String mögliche Ausgaberesultate erzielte.
Flag_Match_Found	Integer	global	Zeigt an, ob eine Übereinstimmung des (reduzierten) Suchlexems mit einem Regellexem gefunden wurde. 0: keine Übereinstimmung gefunden 1: Übereinstimmung gefunden
Processing_Lexeme	String	lokal	Speichert das an die Funktion übergebene Suchlexem.
Que_Rules	SQL	global	Que_Rules ist eine SQL-Text-Variable in ACCESS. Diese Variable wird im Laufe des Programms so lange verändert, bis sie den Text (SQL-String) der endgültigen, für die Ausgabe des Suchresultats relevanten SQL-Abfrage enthält. Diese SQL-Abfrage wird im Laufe des Programms nur an einer einzigen Stelle ausgeführt, nämlich am Programmende durch die Zuweisung (Me.Rule_Lexemes.RowSource="Que_Rules") an das Listenfeld. Diese Abfrage ist verantwortlich für die Ausgabe der Regellexeme und ist eine Selektion über alle 17 Spalten der Tabelle “Rules“.

Tabelle 4.02: Datenlexikon der “total length compare“-Funktion

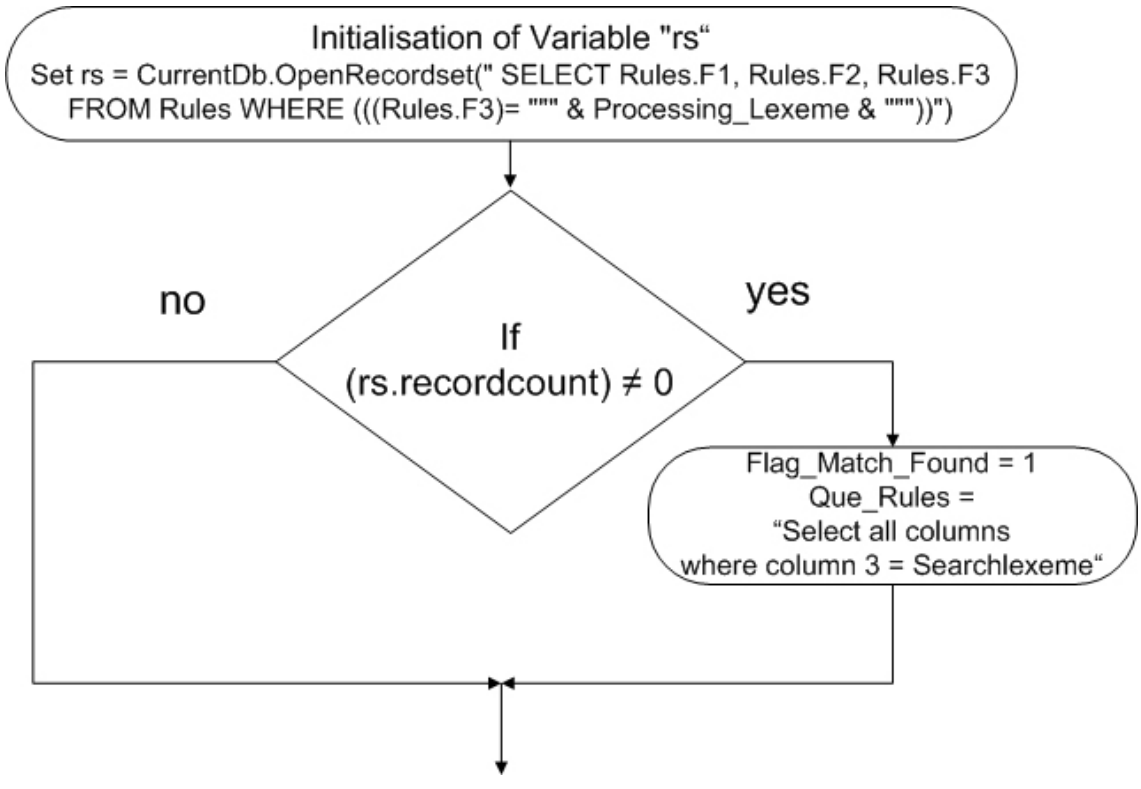


Abbildung 4.02: Ablaufplan der Implementierung der “total length compare“-Funktion

Durch den Aufruf durch die “Main“-Funktion wird der “total length compare“-Funktion eine Variable vom Typ String übergeben. Diese entspricht dem in 3.3.4 beschriebenen akzeptierten Suchlexem ohne Präfixmarker (Unterstriche).
Es wird eine lokale Variable “rs“ vom Typ Recordset[12] deklariert. Nach deren Instanziierung wird diesem Objekt das Ergebnis einer SQL-SELECT-Anweisung übergeben. Die SELECT-Anweisung selektiert die Tabellenspalten von F1 bis F3 aus der Tabelle Rules und ermittelt jene, bei denen sich in der Spalte F3 (Synthetic lexical base) das Suchlexem befindet.

Eine If-Anweisung überprüft, ob der Recordcount[13] des Objektes “rs“ ungleich “0“ ist. Wenn dies der Fall ist, existiert zu dem eingegebenen Suchlexem mindestens ein Eintrag in der Tabelle der Regellexeme, und die Variable “Flag_Match_Found“ wird auf “1“ gesetzt.
Desweiteren wird die im DBMS von ACCESS hinterlegte SQL-Abfrage namens “Que_Rules“ mit einer aktualisierten SQL-Anweisung überschrieben. Diese Anweisung selektiert alle Tabellenspalten von “F1“ bis “F17“ aus der Tabelle “Rules“ und weist ihnen als Alias-Namen die Inhalte der globalen Variablen “F1“ bis “F17“ zu. Es werden nur jene Zeilen ausgewählt, bei denen der Wert der Spalte F3 gleich der an die Funktion übergebenen Variable “Searchlexeme“ (entspricht dem Wert der Variable “Processing_Lexeme“) ist.

Danach endet diese Funktion. Das geschieht ebenfalls, wenn “RecordCount“ das Ergebnis “0“ enthält und der leere Else-Zweig der If-Anweisung ausgeführt wird.

```
Function total_length_compare(Processing_Lexeme As String) As String

Dim rs As DAO.Recordset

Set rs = CurrentDb.OpenRecordset(" SELECT Rules.F1, Rules.F2, Rules.F3 FROM Rules WHERE (((Rules.F3)= """ & Processing_Lexeme & """))")

If rs.RecordCount <> 0 Then

Flag_Match_Found = 1

CurrentDb.QueryDefs("Que_Rules").sql = " SELECT Rules.F1 as [" & F1 & "], Rules.[F2] as [" & F2 & "], Rules.F3 as [" & F3 & "], Rules.F4 as [" & F4 & "], Rules.F5 as [" & F5 & "], Rules.F6 as [" & F6 & "], Rules.F7 as [" & F7 & "], Rules.F8 as [" & F8 & "], Rules.F9 as [" & F9 & "], Rules.F10 as [" & F10 & "], Rules.F11 as [" & F11 & "], Rules.F12 as [" & F12 & "], Rules.F13 as [" & F13 & "], Rules.F14 as [" & F14 & "], Rules.F15 as [" & F15 & "],Rules.F16 as [" & F16 & "], Rules.F17 as [" & F17 & "]" & _
" FROM Rules WHERE (((Rules.F3) = """ & Processing_Lexeme & """))"
Else

End If

End Function
```

Quellcode 4.02: “total length compare“-Funktion

[12] Durch die Konzeption des Recordsets als Objekt mit Eigenschaften und Methoden hat man einen einfachen und komfortablen Zugriff auf die Daten einer Datenbank. Dieser bietet jedem Recordset-Objekt die Möglichkeit, nach bestimmten Werten zu suchen, Werte zu sortieren, Datensätze hinzuzufügen, wieder zu löschen oder zu zählen (vgl. Zimmer, 2006).

[13] Die RecordCount-Eigenschaft ist eine DAO (Data Access Object)-Eigenschaft, die die Anzahl der Datensätze in einer Tabelle einer Access-Datenbank anzeigt (vgl. Microsoft Hilfe und Support, 2004).

4.2.3 Realisierung der Funktion “prefix cut“

Name	Typ	Ort	Verwendung
Search_Letter	String	lokal	Speichert einen Buchstaben des Suchlexems.
Processing_ Lexeme_Length	Integer	lokal	Speichert die Anzahl Buchstaben des Processing_Lexeme
Flag_Prefix_Found	Integer	global	Zeigt an, ob im Suchlexem ein Präfix gefunden wurde. 0: kein Präfix gefunden 1: Präfix gefunden
Searchlexeme	String	global	Speichert das bei Vorhandensein einer Präfixmarkierung von “prefix cut“ verkürzte Suchlexem.
Processing_Lexeme	String	lokal	Speichert das an die Funktion übergebene Suchlexem sowie das durch die Funktion “prefix cut“ reduzierte Suchlexem.

Tabelle 4.03: Datenlexikon der “prefix cut“-Funktion

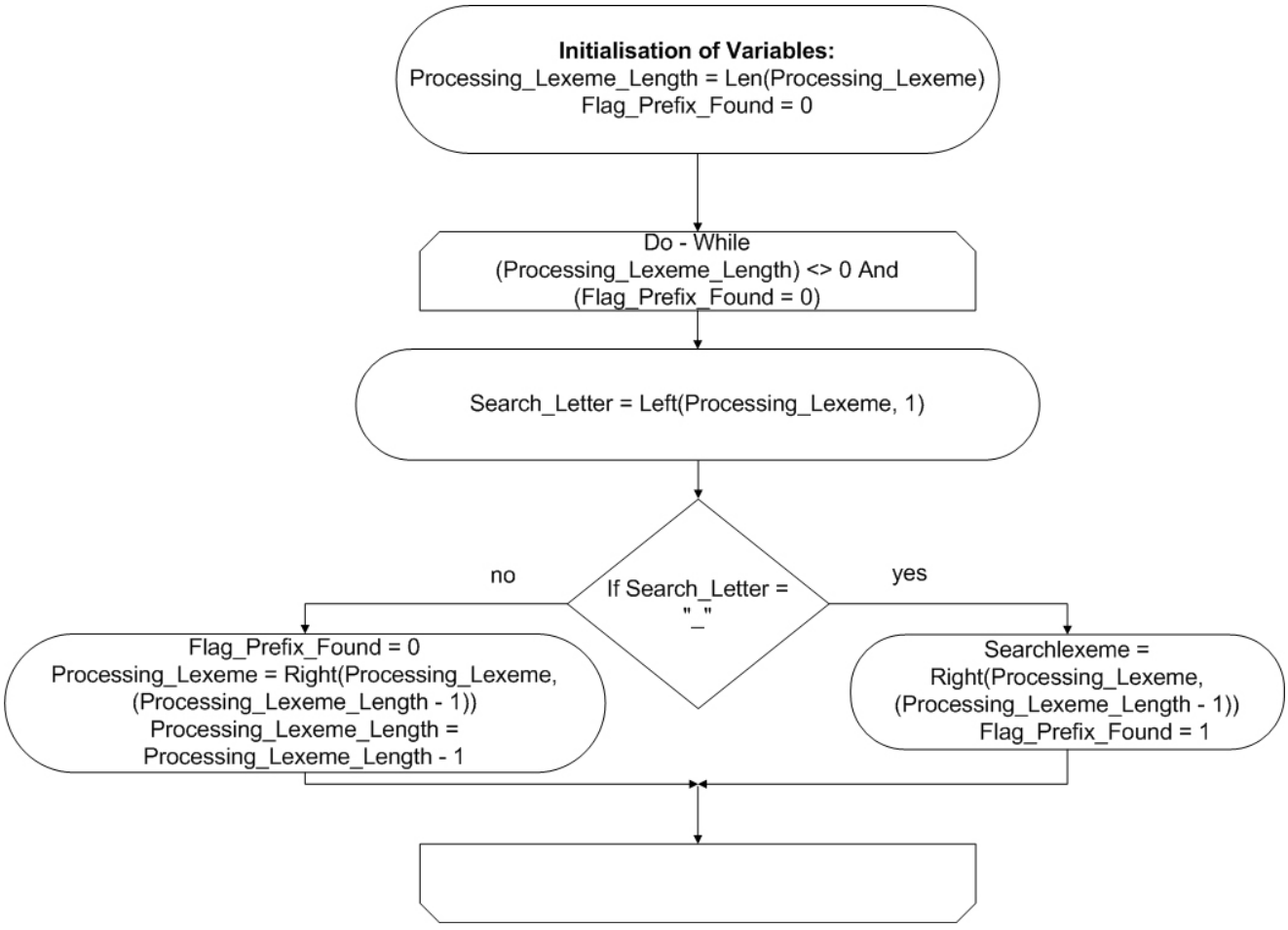


Abbildung 4.04: Ablaufplan der Implementierung der “prefix cut“-Funktion

Beim Aufruf durch die “Main“-Funktion wird der “prefix cut“-Funktion eine Variable vom Typ String übergeben. Diese entspricht dem in 3.3.4 beschriebenen akzeptierten Suchlexem.
Nach der Deklaration gemäß Datenlexikon (Tabelle 4.03) wird die Variable “Processing_Lexeme_Length“ mittels der Funktion “len“[14] mit der Suchlexemlänge initialisiert, und das “Flag_Prefix_Found“ erhält den Wert “0“, d.h. keine Präfixmarkierung.
Nachfolgend wird eine Do-While-Schleife gestartet, in deren Schleifenkopf die Einstiegsbedingung “(Processing_Lexeme_Length) <> 0 And (Flag_Prefix_Found = 0)“ geprüft wird. Diese Schleife wird nun solange wiederholt, bis diese Einstiegsbedingung nicht mehr erfüllt ist. Der Schleifenrumpf beginnt mit einer Initialisierung der Variable “Search_Letter = Left(Processing_Lexeme, 1)“, welche mittels der “left“-Funktion[15] realisiert wird.

Beispiel:
Processing_Lexeme = “en_grave“ (implizit durch den Funktionsaufruf)
Processing_Lexeme_Length = “8“
Flag_Prefix_Found = 0

Erster Schleifendurchlauf
Search_Letter = Left(“en_grave“), 1) →“e“
If-Anweisung→“False“ → Else
Flag_Prefix_Found = 0
Processing_Lexeme = Right(“en_grave“), (8-1)) → “n_grave“
Processing_Lexeme_Length = 8 - 1 →“7“

Zweiter Schleifendurchlauf
Search_Letter = Left(“n_grave“), 1) →“n“
If-Anweisung→“False“ → Else
Flag_Prefix_Found = 0
Processing_Lexeme = Right(“en_grave“), (7-1)) → “_grave“
Processing_Lexeme_Length = 7 - 1 →“6“

Dritter Schleifendurchlauf
Search_Letter = Left(“_grave“), 1) →“_“
If-Anweisung→“True“
Searchlexeme = Right(Processing_Lexeme, (6 - 1)) → “grave“
Flag_Prefix_Found = 1

Durch dieses Verfahren wird das gesamte Suchlexem (Processing_Lexeme) Buchstabe für Buchstabe an die Variable “Search_Letter“ übergeben.
Eine If-Anweisung überprüft, ob sich in dieser Variable ein Unterstrich befindet. Wird dieser identifiziert, wird der globalen “Searchlexeme“-Variable mittels der “right“-Funktion[16] der verbleibende Wert der Variablen “Processing_Lexeme“ (alle Buchstaben rechts vom Unterstrich) zugewiesen und das “Flag_Prefix_Found“ auf den

[14] Die len-Funktion behandelt den ihr übergebenen Inhalt einer Variable wie eine Zeichenfolge. Wenn eine leere Variable übergeben wird, wird der Wert Null zurückgegeben, ansonsten wird die Anzahl der sich in der Variable befindlichen Zeichen (numerische Daten, Zeichenfolgendaten, Datumsangaben) zurückgegeben (vgl. Microsoft Office, 2008).

[15] Gibt einen Wert zurück, der ausgehend von der linken Seite einer Zeichenfolge eine angegebene Anzahl von Zeichen enthält (vgl. Microsoft Office, 2008).

[16] Gibt einen Wert zurück, der ausgehend von der rechten Seite einer Zeichenfolge eine angegebene Anzahl von Zeichen enthält (vgl. Microsoft Office, 2008).

Wert “1“ gesetzt, worauf die Schleife terminiert. Im anderen Fall wird das “Flag_Prefix_Found“ auf “0“ gesetzt, die “Processing_Lexeme_Length“ dekrementiert und die String-Variable “Processing_Lexeme“ (welche das reduzierte Suchlexem enthält) mittels der “right“-Funktion mit dem verbleibenden Wert der Variablen “Processing_Lexeme“ (alle Buchstaben ohne den ersten) belegt. Danach beginnt die Schleife erneut, bis die Eintrittsbedingung im Schleifenkopf nicht mehr erfüllt ist.

```
Function prefix_cut(Processing_Lexeme As String) As String

Dim Search_Letter As String
Dim Processing_Lexeme_Length As Integer

Processing_Lexeme_Length = Len(Processing_Lexeme)
Flag_Prefix_Found = 0

Do While ((Processing_Lexeme_Length) <> 0 And (Flag_Prefix_Found = 0))

   Search_Letter = Left(Processing_Lexeme, 1)

      If Search_Letter = "_" Then

      Searchlexeme = Right(Processing_Lexeme, (Processing_Lexeme_Length - 1))
      Flag_Prefix_Found = 1

      Else

        Flag_Prefix_Found = 0
        Processing_Lexeme = Right(Processing_Lexeme, (Processing_Lexeme_Length - 1))
        Processing_Lexeme_Length = Processing_Lexeme_Length - 1

      End If

Loop

End Function
```

Quellcode 4.03: “prefix cut“-Funktion

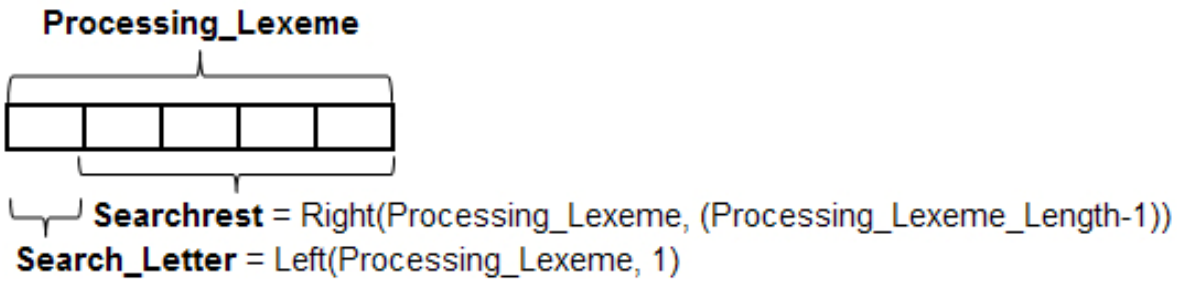


Abbildung 4.04: Verarbeitung der “prefix cut“-Funktion

4.2.4 Realisierung der Funktion “longest match“

Name	Typ	Ort	Verwendung
rs	DAO. recordset	lokal	Speichert das Ergebnis einer SQL-SELECT-Anweisung. Mithilfe der in VBA für DAO.recordset-Variablen bereitgestellten Funktion “Recordcount“ lässt sich einfach feststellen, wie viele Einträge in der Variable gespeichert sind, d.h. ob der an die Variable übergebene SQL-String mögliche Ausgaberesultate erzielte.
Searchalternative _1	String	lokal	
Searchalternative _2	String	lokal	
Searchalternative _3	String	lokal	
Processing _Lexeme_Length	Integer	lokal	Speichert die Anzahl Buchstaben des Processing_Lexeme
Flag_Match_Found	Integer	global	Zeigt an, ob eine Übereinstimmung des (reduzierten) Suchlexems mit einem Regellexem gefunden wurde. 0: keine Übereinstimmung gefunden 1: Übereinstimmung gefunden
Searchlexeme	String	global	Speichert das bei Vorhandensein einer Präfixmarkierung von “prefix cut“ verkürzte Suchlexem.
Match_Lexeme	String	lokal	Speichert das mit einem Registereintrag übereinstimmende veränderte Suchlexem.
Processing _Lexeme	String	lokal	Speichert das an die Funktion übergebene Suchlexem sowie das durch die “longest match“-Funktion reduzierte Suchlexem.
Que_Rules	SQL Abfrage	global	Que_Rules ist eine SQL-Text-Variable in ACCESS. Diese Variable wird im Laufe des Programms so lange verändert, bis sie den Text (SQL-String) der endgültigen, für die Ausgabe des Suchresultats relevanten SQL-Abfrage enthält. Diese SQL-Abfrage wird im Laufe des Programms nur an einer einzigen Stelle ausgeführt, nämlich am Programmende durch die Zuweisung (Me.Rule_Lexemes.RowSource="Que_Rules") an das Listenfeld. Diese Abfrage ist verantwortlich für die Ausgabe der Regellexeme und ist eine Selektion über alle 17 Spalten der Tabelle “Rules“.

Tabelle 4.04: Datenlexikon der “longest match“-Funktion

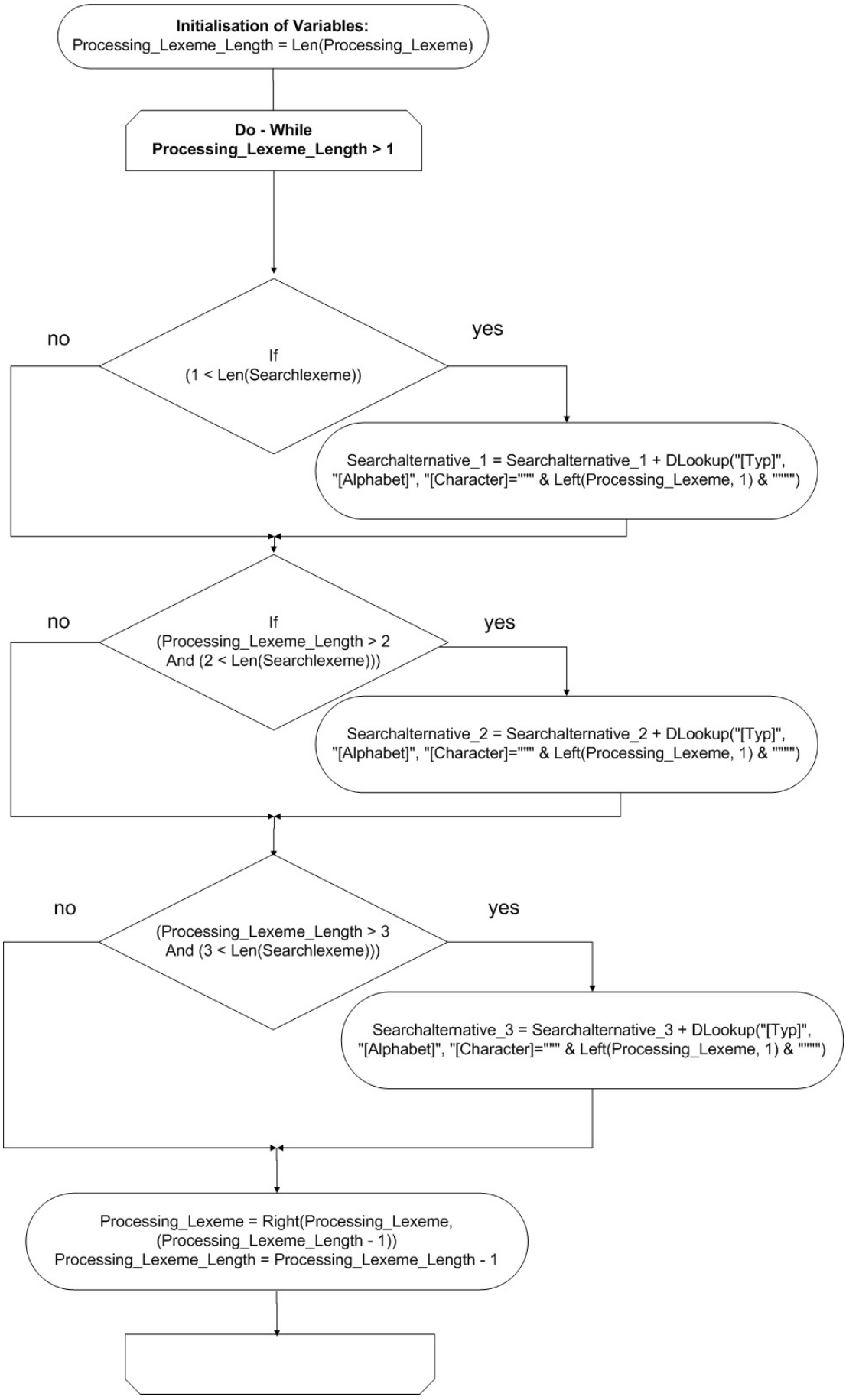


Abbildung 4.05: Ablaufplan der Implementierung der “longest match“-Funktion: Generierung von Jokerzeichen

Abbildung 4.06: Ablaufplan der Implementierung der “longest match“-Funktion: rechtsbündiges Auffüllen mit Buchstaben

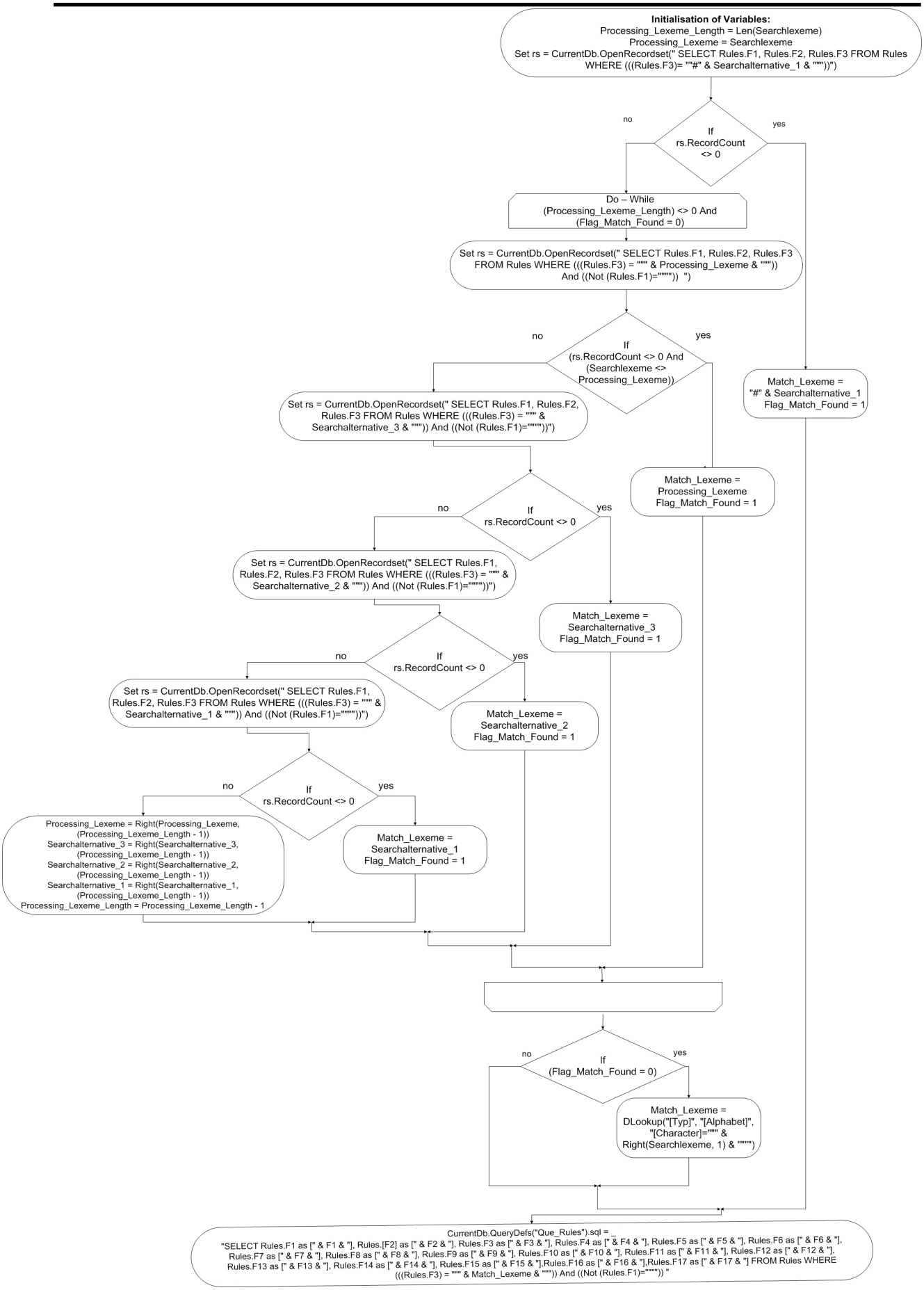


Abbildung 4.07: Ablaufplan der Implementierung der “longest match“-Funktion: Prüfen der Suchalternativen

Beim Aufruf durch die “Main“-Funktion wird der “longest match“-Funktion eine Variable vom Typ String übergeben. Diese entspricht dem in 3.3.4 beschriebenen akzeptierten Suchlexem. Nach der Deklaration gemäß Datenlexikon (Tabelle 4.04) wird die Variable “Processing_Lexeme_Length“ mit der Länge des an die Funktion übergebenen Suchlexems initialisiert.

4.2.4.1 Generierung der Suchalternativen

Nun beginnt das Generieren der Suchalternativen mit Jokerzeichen (Abb. 4.05 und 4.06). Dazu wird eine Do-While-Schleife gestartet, in deren Schleifenkopf die Einstiegsbedingung “Processing_Lexeme_Length > 1“ geprüft wird.
Es folgen drei If-Anweisungen, welche in Abhängigkeit der an die Funktion übergebenen Suchlexemlänge und der Länge des Verarbeitungslexems überprüfen, ob eine entsprechende Suchalternative generiert bzw. wie viele Jokerzeichen angefügt werden. Ist die jeweilige Bedingung erfüllt, wird den Suchalternativen (“Searchalternative_1“, “Searchalternative_2“, “Searchalternative_3“) ihr eigener Wert, d.h. der Wert aus dem vorangegangenen Schleifendurchlauf, konkateniert mit dem Jokerzeichen (C / V), welches mittels der “DLookup“-Funktion“[17] ermittelt wird, übergeben. Die “DLookup“-Funktion“ sucht einen Eintrag in der Tabellenspalte “Typ“ aus der Tabelle “Alphabet“, bei dem der Wert in der Spalte “Character“ gleich dem linken Buchstaben der Variablen “Processing_Lexeme“ ist.

Beispiel für “Searchalternative“ 2:
Processing_Lexeme = “present“
Processing_Lexeme_Length = Len(“present“) →“7“

Erster Schleifendurchlauf
If (7 > 2 And (2 < Len(“present“)))) →“True“

Searchalternative_2 = NULL + DLookup("[Typ]", "[Alphabet]", "[Character]=""" & Left(“present“, 1) & """")
d.h. Searchalternative_2 = NULL + DLookup("[Typ]", "[Alphabet]", "[Character]="p“)
d.h. Searchalternative_2 = NULL + “C“

…
Processing_Lexeme = Right(Processing_Lexeme, (7 - 1)) →“resent“
Processing_Lexeme_Length = 7 - 1 →“6“

Zweiter Schleifendurchlauf
If (6 > 2 And (2 < Len(“Present“)))) →“True“
Searchalternative_2 = “C“ + DLookup("[Typ]", "[Alphabet]", "[Character]=""" & Left(“resent“, 1) & """")
d.h. Searchalternative_2 = “C“ + DLookup("[Typ]", "[Alphabet]", "[Character]="r“)
d.h. Searchalternative_2 = “C“ + “C“
…
Processing_Lexeme = Right(Processing_Lexeme, (6 - 1)) →“esent“
Processing_Lexeme_Length = 6 - 1 →“5“
usw.

Mit diesem Verfahren werden je Schleifendurchlauf die Suchalternativen nach und nach mit Jokerzeichen gefüllt, wobei das Anfügen der Jokerzeichen je nach Suchalternative an unterschiedlichen Positionen beendet

[17] Die DLookup-Funktion wird verwendet, um einen Wert eines bestimmten Feldes aus einer angegebenen Datensatzgruppe, definiert durch eine Tabelle, eine Abfrage oder einen SQL-Ausdruck, zu entnehmen (vgl. Microsoft Office, 2008).

wird (Suchalternative_1: Anzahl Buchstaben des Suchlexems - 1, Suchalternative_2: Anzahl Buchstaben des Suchlexems - 2, Suchalternative_3: Anzahl Buchstaben des Suchlexems - 3).
Nach der Schleife werden in drei If-Anweisungen die Suchalternativen mit der “right“-Funktion um die “fehlenden“ echten Buchstaben des Suchlexems ergänzt (Abb. 4.06), sofern eine Suchalternative erstellt wurde. Ansonsten wird diese mit “NULL“ belegt.

4.2.4.2 Überprüfung der Suchalternativen

Nach der Generierung der Suchalternativen wird der Wert der “Processing_Lexeme_Length“ auf die Länge des “Searchlexeme“ und das “Processing_Lexeme“ auf “Searchlexeme“ zurückgesetzt.
Zunächst wird der Variable “rs“ das Ergebnis einer SQL-SELECT-Anweisung übergeben. Diese SQL-Anweisung selektiert die ersten drei Spalten der Tabelle “Rules“. Es werden nur jene Datensätze in der Variablen “rs“ gespeichert, welche in der Spalte “F3“ (Synthetic lexical base) den Eintrag der Suchalternative_1 mit vorange-stelltem #-Zeichen enthalten.
Eine If-Anweisung überprüft, ob die Anzahl (“RecordCount“) der in der Variablen gespeicherten Einträge ungleich “0“ ist. Wenn diese Bedingung erfüllt ist, wird der Wert der Variable “Match_Lexeme“ auf den Wert der Suchalternative_1 mit vorangestellten #-Zeichen gesetzt, und das “Flag_Match_Found“ erhält den Wert “1“. Die nachfolgende Do-While-Schleife wird in diesem Fall komplett übersprungen.

Ist diese Bedingung nicht erfüllt, d.h. wurden keine Einträge gespeichert, dann wird eine Do-While-Schleife gestartet, in deren Schleifenkopf als Einstiegsbedingung ((Processing_Lexeme_Length) <> 0 And (Flag_Match_Found = 0)) geprüft wird. In dieser Schleife werden gemäß der Reihenfolge der Bedingungshierarchie zuerst Verarbeitungslexem (Processing_Lexeme), dann Suchalternative_3, Suchalternative_2 und Suchalternative_1 abgearbeitet. Die Vorgehensweise ist immer die gleiche. Der Variable “rs“ wird das Ergebnis einer SQL-SELECT-Anweisung übergeben, welche über die ersten drei Spalten der Tabelle “Rules“ selektiert und nur jene Datensätze speichert, die in der Spalte “F3“ (Synthetic lexical base) den Wert des reduzierten Verarbeitungslexems (“Processing_Lexeme“) bzw. der reduzierten Suchalternativen und zusätzlich einen Eintrag in der Spalte “F1“ (Cluster) enthält.
Nachfolgend wird mittels einer If-Anweisung überprüft, ob die Anzahl der gespeicherten Datensätze ungleich “0“ ist.
In diesem Fall, d.h. befinden sich Datensätze in der Variable “rs“, erhält die Variable “Match_Lexeme“ den Wert der jeweiligen reduzierten Suchalternative bzw. des reduzierten Verarbeitungslexems, und die Variable “Flag_Match_Found“ erhält den Wert “1“, welches dazu führt, dass kein weiterer Schleifeneintritt erfolgt. Im anderen Fall passiert nichts, und der Programmcode des Schleifenrumpfs wird fortgesetzt. Wurde keine der If-Bedingungen erfüllt, erfolgen folgende Wertzuweisungen im Else-Zweig der Suchalternative_1, und solange die Schleifeneintrittsbedingung erfüllt ist, beginnt die Do-While-Schleife erneut.

Processing_Lexeme = Right(Processing_Lexeme, (Processing_Lexeme_Length - 1))
Searchalternative_3 = Right(Searchalternative_3, (Processing_Lexeme_Length - 1))
Searchalternative_2 = Right(Searchalternative_2, (Processing_Lexeme_Length - 1))
Searchalternative_1 = Right(Searchalternative_1, (Processing_Lexeme_Length - 1))
Processing_Lexeme_Length = Processing_Lexeme_Length - 1

Beispiel:
Processing_Lexeme_Length = Len(“present“) →“7“
Processing_Lexeme = “present“
Set rs = CurrentDb.OpenRecordset(" SELECT Rules.F1, Rules.F2, Rules.F3 FROM Rules WHERE (((Rules.F3)= ""#" & “CCVCVCt“ & """))")→“0“
If rs.RecordCount <> 0 →“False“ → Else

Erster Schleifendurchlauf
Set rs = CurrentDb.OpenRecordset(" SELECT Rules.F1, Rules.F2, Rules.F3 FROM Rules WHERE (((Rules.F3) = """ & “Present“ & """)) And ((Not (Rules.F1)=""""")) ")→“0“
If (rs.RecordCount <> 0 And (“Present“ <> “Present“)) →“False“ → Else

Set rs = CurrentDb.OpenRecordset(" SELECT Rules.F1, Rules.F2, Rules.F3 FROM Rules WHERE (((Rules.F3) = """ & “CCVCent“ & """)) And ((Not (Rules.F1)="""""))")
If rs.RecordCount <> 0 →“False“ → Else

Set rs = CurrentDb.OpenRecordset(" SELECT Rules.F1, Rules.F2, Rules.F3 FROM Rules WHERE (((Rules.F3) = """ & “CCVCVnt“ & """)) And ((Not (Rules.F1)=""""))")
If rs.RecordCount <> 0 →“False“ → Else

Set rs = CurrentDb.OpenRecordset(" SELECT Rules.F1, Rules.F2, Rules.F3 FROM Rules WHERE (((Rules.F3) = """ & “CCVCVCt“ & """)) And ((Not (Rules.F1)=""""))")
If rs.RecordCount <> 0 →“False“ → Else

Processing_Lexeme = Right(Processing_Lexeme, (7 - 1)) →“resent“
Searchalternative_3 = Right(Searchalternative_3, (7 - 1)) →“CVCent“
Searchalternative_2 = Right(Searchalternative_2, (7 - 1)) →“CVCVnt“
Searchalternative_1 = Right(Searchalternative_1, (7 - 1)) →“CVCVCt“
Processing_Lexeme_Length = 7 - 1

Wenn die Do-While-Schleife durchlaufen ist, ohne vorzeitig abzubrechen, und somit weder das bis auf den letzten Buchstaben reduzierte Verarbeitungslexem noch die reduzierten Suchalternativen eine Übereinstimmung mit den Registereinträgen ergaben (Flag_Match_Found ungleich 1), wird der letzte Buchstabe des Suchlexems mittels der “DLookup“-Funktion in ein Jokerzeichen umgewandelt und der Variablen “Match_Lexeme“ zugewiesen (Suchalternative 4 exit).

Beispiel nach dem letzten erfolglosen Schleifendurchlauf:

If (Flag_Match_Found = 0) →“True“
Match_Lexeme = DLookup("[Typ]", "[Alphabet]", "[Character]=""" & Right(“Present“, 1) & """)→“C“

Im letzten Schritt wird die für die Ausgabe verantwortliche Abfrage “Que_Rules“ mit einer SQL-Anweisung überschrieben. Diese SQL-Anweisung selektiert alle Spalten der Tabelle “Rules“ von “F1“ bis “F17“ und wählt jene Zeilen aus, bei denen der Eintrag in der Spalte “F3“ (Synthetic lexical base) dem Wert der Variablen “Match_Lexeme“ entspricht und ein Eintrag in der Spalte “F1“ (Cluster) existiert. Desweiteren wird jeder selektierten Spalte als Überschrift der Wert aus der gleichnamigen globalen Variable zugewiesen (siehe 4.2).

```
‘ Suchalternativen generieren

Function longest_match(Processing_Lexeme As String) As String

Dim rs As DAO.Recordset
Dim Searchalternative_1 As String
Dim Searchalternative_2 As String
Dim Searchalternative_3 As String
Dim Processing_Lexeme_Length As Integer
Dim Match_Lexeme as String
Processing_Lexeme_Length = Len(Processing_Lexeme)

Do While Processing_Lexeme_Length >  1

If (1 < Len(Searchlexeme)) Then
Searchalternative_1 = Searchalternative_1 + DLookup("[Typ]", "[Alphabet]", "[Character]=""" &
Left(Processing_Lexeme, 1) & """)
Else
End If

If (Processing_Lexeme_Length > 2 And (2 < Len(Searchlexeme))) Then
Searchalternative_2 = Searchalternative_2 + DLookup("[Typ]", "[Alphabet]", "[Character]=""" &
Left(Processing_Lexeme, 1) & """)
Else
End If
```

```
If (Processing_Lexeme_Length > 3 And (3 < Len(Searchlexeme))) Then
Searchalternative_3 = Searchalternative_3 + DLookup("[Typ]", "[Alphabet]", "[Character]=""" &
Left(Processing_Lexeme, 1) & """")
Else
End If
Processing_Lexeme = Right(Processing_Lexeme, (Processing_Lexeme_Length - 1))
Processing_Lexeme_Length = Processing_Lexeme_Length - 1

Loop

'Auffüllen mit Buchstaben

If (1 < Len(Searchlexeme)) Then
Searchalternative_1 = Searchalternative_1 + Right(Searchlexeme, (1))
Else
Searchalternative_1 = ""
End If

If (2 < Len(Searchlexeme)) Then
Searchalternative_2 = Searchalternative_2 + Right(Searchlexeme, (2))
Else
Searchalternative_2 = ""
End If

If (3 < Len(Searchlexeme)) Then
Searchalternative_3 = Searchalternative_3 + Right(Searchlexeme, (3))
Else
Searchalternative_3 = ""
End If

'Prüfen der Suchalternativen

Processing_Lexeme_Length = Len(Searchlexeme)
Processing_Lexeme = Searchlexeme

Set rs = CurrentDb.OpenRecordset(" SELECT Rules.F1, Rules.F2, Rules.F3 FROM Rules
          WHERE (((Rules.F3)= ""#" & Searchalternative_1 & """))")

If rs.RecordCount <> 0 Then
Match_Lexeme = "#" & Searchalternative_1
Flag_Match_Found = 1
Else

     Do While ((Processing_Lexeme_Length) <> 0 And (Flag_Match_Found = 0))

        Set rs = CurrentDb.OpenRecordset(" SELECT Rules.F1, Rules.F2, Rules.F3 FROM Rules
                  WHERE (((Rules.F3) = """ & Processing_Lexeme & """)) And ((Not (Rules.F1)=""""))  ")

        If (rs.RecordCount <> 0 And (Searchlexeme <> Processing_Lexeme)) Then
        Match_Lexeme = Processing_Lexeme
        Flag_Match_Found = 1
        Else

        Set rs = CurrentDb.OpenRecordset(" SELECT Rules.F1, Rules.F2, Rules.F3 FROM Rules
                  WHERE (((Rules.F3) = """ & Searchalternative_3 & """)) And ((Not (Rules.F1)=""""))")

           If rs.RecordCount <> 0 Then
```

```
        Match_Lexeme = Searchalternative_3
        Flag_Match_Found = 1
        Else

        Set rs = CurrentDb.OpenRecordset(" SELECT Rules.F1, Rules.F2, Rules.F3 FROM Rules
                 WHERE (((Rules.F3) = """ & Searchalternative_2 & """)) And ((Not (Rules.F1)=""""))")

           If rs.RecordCount <> 0 Then
           Match_Lexeme = Searchalternative_2
           Flag_Match_Found = 1
           Else

           Set rs = CurrentDb.OpenRecordset(" SELECT Rules.F1, Rules.F2, Rules.F3 FROM Rules
                    WHERE (((Rules.F3) = """ & Searchalternative_1 & """)) And ((Not (Rules.F1)=""""))")

              If rs.RecordCount <> 0 Then
              Match_Lexeme = Searchalternative_1
              Flag_Match_Found = 1
              Else

              Processing_Lexeme = Right(Processing_Lexeme, (Processing_Lexeme_Length - 1))
              Searchalternative_3 = Right(Searchalternative_3, (Processing_Lexeme_Length - 1))
              Searchalternative_2 = Right(Searchalternative_2, (Processing_Lexeme_Length - 1))
              Searchalternative_1 = Right(Searchalternative_1, (Processing_Lexeme_Length - 1))
              Processing_Lexeme_Length = Processing_Lexeme_Length - 1

              End If         '(Searchalternative_1)
           End If          '(Searchalternative_2)
        End If           '(Searchalternative_3)
     End If            '(Processing_Lexeme)

    Loop

     If (Flag_Match_Found = 0) Then
     Match_Lexeme = DLookup("[Typ]", "[Alphabet]", "[Character]=""" & Right(Searchlexeme, 1) & """")
     Else
     End If     '(Searchalternative_4 exit)
 End If  '(#Searchalternative_1)

CurrentDb.QueryDefs("Que_Rules").sql = _
" SELECT Rules.F1 as [" & F1 & "], Rules.[F2] as [" & F2 & "], Rules.F3 as [" & F3 & "], Rules.F4 as [" & F4 & "],
Rules.F5 as [" & F5 & "], Rules.F6 as [" & F6 & "], Rules.F7 as [" & F7 & "], Rules.F8 as [" & F8 & "], Rules.F9 as ["
& F9 & "], Rules.F10 as [" & F10 & "], Rules.F11 as [" & F11 & "], Rules.F12 as [" & F12 & "], Rules.F13 as [" & F13
& "], Rules.F14 as [" & F14 & "], Rules.F15 as [" & F15 & "],Rules.F16 as [" & F16 & "],Rules.F17 as [" & F17 & "]
FROM Rules WHERE (((Rules.F3) = """ & Match_Lexeme & """)) And ((Not (Rules.F1)=""""))  "

End Function
```

Quellcode 4.04: “longest match“-Funktion

4.2.5 Zusatzfunktion für den Datenimport

Nach dem Betätigen des Importbuttons auf der Benutzeroberfläche (siehe 4.1.1, Abb. 4.01) wird mittels der Accessfunktion “DoCmd.Close acForm“ das Startformular geschlossen. Dies ist erforderlich, damit die dahinterliegende SQL-Abfrage und die darunterliegende Tabelle problemlos verändert werden können.
Eine Variable “Path“ vom Typ String wird definiert, in welcher der durch den Benutzer über den Microsoft Programmbrowser ausgewählte Dateipfad gespeichert wird. Mittels der Accessfunktion “DateiOeffnen“ wird der Microsoft Programmbrowser geöffnet, welcher standardmäßig den Pfad zu den “Eigenen Dateien“ aktiviert.

Nun kann der Benutzer einen beliebigen Pfad auswählen, in dem sich entweder seine excelbasierte Alphabettabelle oder sein excelbasiertes Register mit den Regellexemen der jeweiligen Sprache befindet. Diese Entscheidung trifft der Benutzer im Formularfeld “Select“. Nun wird dessen Wert mit dem Befehl “Me.Select“ abgefragt. Eine If-Anweisung überprüft, ob es sich bei dem Wert um “Alphabet“ handelt.

Wenn dies der Fall ist, wird die alte Alphabet-Tabelle entfernt (DoCmd.RunSQL "DROP TABLE Alphabet") und die neue mittels der Funktion “DoCmd.TransferSpreadsheet acImport“ importiert. In den Parametern dieser Funktion steht der neue Name der Tabelle “Alphabet“, der Pfad, wo die zu importierende Datei zu finden ist, und “True“ oder “False“ dafür, ob die Spalten-Namen der Exceldatei als Tabellenspaltennamen übernommen werden sollen.

Andernfalls (nicht Alphabet, sondern Register) wird die alte Tabelle “Head“ (in dieser Tabelle werden die Spaltenbeschriftungen der Lexemregister gespeichert) entfernt. Durch die Funktion “DoCmd.TransferSpreadsheet acImport“ wird nun die neue “Head“-Tabelle nach der oben beschriebenen Struktur importiert. Der zusätzliche Parameter “A1:Q1“ beschränkt den Import auf die erste Zeile der 17 Spalten “A“ bis “Q“.
Desweiteren wird die alte Tabelle “Rules“ aus der Datenbank entfernt und anschließend die neue mit der gleichen Prozedur beginnend ab Zeile zwei importiert.

Nun wird das “Start“-Formular, welches der Benutzeroberfläche entspricht, erneut geöffnet und das “Import“-Formular geschlossen. Nach Abschluss dieser Funktion befinden sich die neuen Datenbestände in der Datenbank.

```
Private Sub Import_Click()
DoCmd.Close acForm, "Start"
Dim Path As String
Path = DateiOeffnen("C:Eigene Dateien", "Datei öffnen")

If Me.Select = "Alphabet" Then

'Alphabet
DoCmd.RunSQL "DROP TABLE Alphabet"
DoCmd.TransferSpreadsheet acImport, acSpreadsheetTypeExcel97, "Alphabet", Path, True

Else

'Head
DoCmd.RunSQL "DROP TABLE Head"
DoCmd.TransferSpreadsheet acImport, acSpreadsheetTypeExcel97, "Head", Path, False, "A1:Q1"

'Rules
DoCmd.RunSQL "DROP TABLE Rules"
DoCmd.TransferSpreadsheet acImport, acSpreadsheetTypeExcel97, "Rules", Path, False, "A2:Q"

End If

DoCmd.OpenForm "Start", acNormal
DoCmd.Close acForm, "Import"
End Sub
```

Quellcode 4.05: Importfunktion

5. Anwendungsmöglichkeiten

Das Ergebnis dieser Forschungsarbeit, d.h. die technische Umsetzung des SMIRT-Algorithmus in einer benutzerfreundlichen Programmversion, kann in folgenden Bereichen hilfreiche Verwendung finden. Dem Sprachwissenschaftler liefert es auf Knopfdruck strukturelle Informationen über das Beugungsverhalten eines bestimmten Verbs in einem untersuchten Verbalsystem. Desweiteren erhält er statistische Informationen über das abweichende Verhalten (es gibt Ausnahmen, aber die Mehrheit sind regelmäßig flektierende Lexeme) von Verben im jeweiligen Verbalsystem. Dieser schnelle Informationszugang kann einem Sprachlernenden ebenfalls das Lernen einer Sprache bzw. das Lernen von Verben einer Sprache erleichtern. Der Unterricht für Sprachlernende könnte hierdurch anschaulicher gemacht werden.
Während herkömmliche Übersetzungsprogramme lediglich ihre Listeneinträge durchgehen und nur, sofern sich das gesuchte Lexem darin wiederfinden lässt, ein Resultat liefern, macht sich dieser Algorithmus, mit Hilfe der durch Data Mining generierten inhomogenen Cluster, die statistische Wahrscheinlichkeit von Klassenzuordnungen zu Nutze.
Eine Sprache wie auch deren Verben unterliegt einem Veränderungsprozess. Daher kommen neue Verben in das Verbalsystem einer Sprache, für die der Algorithmus ebenfalls in der Lage ist, richtige Aussagen zu treffen.

Beispiele (eingedeutschte Verben):
googeln → homC *eln: ~eln, ~elte, ge~elt* (*googeln, googelte, gegoogelt*)
downloaden → basC2 *den: ~den, ~det, ~dete, ge~det* (*downloaden, downloadet, downloadete, gedownloadet*)

Weitere Vorteile ergeben sich aus dem in 4.1.1 beschriebenen Programmkonzept. Da die drei Bereiche (MVC) entkoppelt sind, lassen sich diese problemlos durch andere ersetzen. Beispielsweise könnte eine webbasierte Realisierung wie folgt aussehen:
View: HTML-Seite im Webbrowser
Controller: Umsetzung der Programmlogik in HTML, PHP oder Java
Model: beliebige SQL-Datenbank (mySQL, Oracle, Microsoft SQL Server)

Literaturverzeichnis

Alpar, Paul; Niedereichholz, Joachim: Data Mining im praktischen Einsatz. Braunschweig u.a.: Vieweg u.a. 2000.

Ester, Martin; Sander, Jörg: Knowledge discovery in databases. Techniken und Anwendungen. Berlin: Springer 2000.

Hesse, Wolfgang; Mayr, Heinrich C.: Modellierung in der Softwaretechnik. Informatik Spektrum 31 (2008) 377-393.

Holl, Alfred: Romanische Verbalmorphologie und relationentheoretische mathematische Linguistik. Axiomatisierung und algorithmische Anwendung des klassischen Wort-und-Paradigma-Modells.
Tübingen: Niemeyer 1988.

Holl, Alfred: The inflectional morphology of the Swedish verb with respect to reverse order: analogy, pattern verbs and their key forms.
Arkiv för nordisk filologi 116 (2001) 193-220.

Holl, Alfred: Nutzen und Tücken von Analogieschlüssen in der Verbalmorphologie: Rückläufige Ähnlichkeit als tertium comparationis in ausgewählten romanischen und germanischen Sprachen.
In: Heinemann, S.; Bernhard, G.; Kattenbusch, D. (ed.): Roma et Romania. Festschrift für Gerhard Ernst zum 65. Geburtstag.
Tübingen: Niemeyer 2002, 151-167.

Holl, Alfred: Datenanalyseverfahren der Informatik (Data Mining) als Grundlage einer didaktischen Darstellung der französischen Verbalmorphologie.
In: Bernhard, G.; Kattenbusch, D.; Stein, P. (ed.): Namen und Wörter. Festschrift für Josef Felixberger zum 65. Geburtstag.
Regensburg: Haus des Buches 2003, 107-119.

Holl, Alfred; Behrschmidt, André; Kühn, Alexander: Rückläufige Register zur russischen und deutschen Verbalmorphologie. Aufbereitung mit Datenanalyseverfahren der Informatik (Data Mining).
Regensburg: Roderer 2004
[= Studia et exempla linguistica et philologica, Series V: Lexica, Tom. 4].

Holl, Alfred; Pavlidis, Stilianos; Urban, Reinhard: Rückläufiges Wörterbuch zur alt- und neugriechischen Verbalmorphologie. Aufbereitung mit Datenanalyseverfahren der Informatik (Data Mining).
Regensburg: Roderer 2006
[= Studia et exempla linguistica et philologica, Series V: Lexica, Tom. 5].

Holl, Alfred; Maroldo, Sara; Urban, Reinhard: The inflectional morphologies of the Swedish noun, the Swedish verb and the English verb. Reverse dictionaries based upon data mining methods.
Växjö: Växjö University Press 2007
[= Mathematical modelling in physics, engineering and cognitive sciences, vol. 12].

Holl, Alfred; Suljic, Ivan: Rückläufiges Wörterbuch zur kroatischen Verbalmorphologie. Aufbereitung mit Datenanalyseverfahren der Informatik (Data Mining).
Regensburg: Roderer 2009, Forthcoming
[= Studia et exempla linguistica et philologica, Series V: Lexica, Tom. 6].

Internet-Links

Middendorf, Stefan; Singer, Reiner; Heid, Jörn: Programmierhandbuch und Referenz für die JavaTM-2-Plattform, Standard Edition, 3. Auflage 2002.
http://www.dpunkt.de/java/Programmieren_mit_Java/Oberflaechenprogrammierung/40.html

Zimmer, Wolfgang: Datenbanklösung Access Datenbanksysteme, 2006.
http://www.access-hilfe.de/

Microsoft Hilfe und Support, 2004.
http://support.microsoft.com/kb/875287/de

Microsoft Office, 2008.
http://office.microsoft.com/de-at/access/HA012288741031.aspx