Jens Gräf, Andreas Henrich, Volker Lüdecke, Christoph Schlieder

Geografisches Information Retrieval

Im klassischen Information Retrieval (IR) werden Suchanfragen gestellt, die sich auf den Inhalt von Dokumenten beziehen. Das geografische IR (GIR) bezieht daneben den räumlichen Aspekt ein, mit dem nur solche Webseiten gefunden werden sollen, die einen bestimmten geografischen Kontext besitzen. Dazu müssen bestehende IR-Systeme um entsprechende Komponenten erweitert werden, die den geografischen Kontext einer Webseite erkennen, geografische Anfragen verarbeiten und bei der Ergebnispräsentation die räumlichen Zusammenhänge darstellen können. In diesem Artikel soll ein Überblick über die Problemstellungen in diesem Gebiet gegeben und die grundsätzlichen Komponenten eines geografischen Information-Retrieval-Systems vorgestellt werden.

1 Einführung

Nach einer Studie von [Sanderson & Kohler 2004] weisen etwa 20% aller Suchanfragen im Web einen geografischen Kontext auf. Dieser kann entweder ausdrücklich formuliert sein und in Form geografischer Terme in der Anfrage vorliegen oder implizit im Informationswunsch des Anfragenden enthalten sein. Herkömmliche Suchdienste können dies nur anhand der formulierten Ausdrücke berücksichtigen, die über boolesche Operatoren mit anderen Anfragetermen verknüpft werden (z.B. »Biergarten« UND »Bamberg«). Geografische Suchdienste sollen darüber hinaus geografische Aspekte einer Suchanfrage einbeziehen und entsprechende Mechanismen zur Verfügung stellen. Die Bedeutung geografischer Suchdienste für das WWW wird durch die Bestrebungen der großen Anbieter unterstrichen, lokale Suchfunktionen anzubieten, wie beispielsweise Google local[1] oder die lokale Suche von Yahoo[2].

Viele Webseiten verfügen über einen geografischen Kontext, bei dem zwischen *Location* und *Locality* unterschieden werden kann. Die *Location* einer Webseite umfasst den geografischen Ort oder auch eine Menge von Orten, auf die in der Webseite Bezug genommen wird. Oft wird dies auch als der geografische Fokus bezeichnet. Die *Locality* dient einer Beschreibung der geografischen Bedeutung oder auch der Reichweite des geografischen Bezugs [Gravano et al. 2003]. Dieses Konzept wird für eine Differenzierung der Webseiten im Hinblick darauf benutzt, ob deren Inhalt hauptsächlich für eine lokal oder regional eingegrenzte Nutzergruppe relevant ist oder ob dieser als ortsunabhängig von Interesse und von globaler Bedeutung angesehen werden kann. Darüber hinaus kann eine *Provider Location* identifiziert werden, die den Ort des Anbieters einer Webseite bezeichnet [Wang et al. 2005]. In der Literatur wird bisweilen auch zwischen *Content Location* (Ortsbezug des Seiteninhaltes) und *Serving Location* (Ortsbezug des anvisierten Nutzerkreises) unterschieden. Es findet sich insgesamt keine eindeutige Begriffsverwendung. Wir werden im Folgenden *Location* und *Locality* in dem oben skizzierten Sinne verwenden.

Möchte ein Informationssuchender eine Webseite von lokaler Bedeutung finden, wird er bei klassischen Suchmaschinen üblicherweise zusätzliche Terme in die Anfrage aufnehmen, die relevante Orte benennen. Dabei werden nur diejenigen Dokumente im Ergebnis enthalten sein, die diese Ortsbezeichnungen auch als eigenständiges Wort beinhalten. So werden bei einer Anfrage »Wandern UND Franken« aufgrund der mangelnden Umsetzung geografischer Konzepte in klassischen Suchmaschinen keine Webseiten in der Ergebnismenge enthalten sein, die sich mit Wandern in der Rhön, im Spessart o.Ä. beschäftigen. Zudem werden für die Relevanzbeurteilung in den heute gebräuchlichen Suchmaschinen im Internet häufig Algorithmen eingesetzt, die auf Zitationsverfahren beruhen. Bei diesen Verfahren wird die Relevanz von Webseiten unter anderem danach beurteilt, von wie vielen anderen Webseiten diese jeweils mit Hyperlinks referenziert werden [Brin & Page 1998]. »Nur« lokal relevante Webseiten, die für einen Nutzer bei der Informationssuche besonders interessant wären, werden durch diese gängigen Algorithmen jedoch aufgrund der wenigen auf sie verweisenden Links benachteiligt.

Spezialisierte geografische Suchmaschinen sollten dem Benutzer erlauben, den geografischen Aspekt seines Informationswunsches separat zu formulieren. Dabei können geografische Bezüge auch in der Ergebnisdarstellung visualisiert werden und ein geografisches Browsing ermöglichen.

Im Folgenden werden grundlegende Ansätze, Techniken und Probleme einer geografisch fokussierten Suche im Internet vorgestellt.

2 Bestimmung des geografischen Kontextes

Für die Umsetzung einer geografisch fokussierten Suche ist es notwendig, den geografischen Kontext der Webseiten im Internet zu ermitteln. Zur Erkennung der *Location* und *Locality* von Ressourcen im Web existieren unterschiedliche Ansätze und Techniken, die in unterschiedlichem Maße für diese Aufgabenstellung geeignet und mit jeweils individuellen Problemen behaftet sind. Dabei können vier Grundrichtungen der Ansätze unterschieden werden, die auf unterschiedlichen Informationsquellen beruhen: der Netzwerktechnologie, dem Inhalt, der Linkstruktur und den Metadaten.

Netzwerktechnologie-basierte Ansätze versuchen dazu, die Wegewahl der IP-Vermittlungsschicht, das Domain Name System (DNS) oder Datenbanken mit Informationen über Domain-Registrierungen auszunutzen, um eine Webseite geografisch einzuordnen. Diese Ansätze wurden in ersten einfachen Systemen verwendet und weisen zahlreiche Schwächen auf. Insbesondere wird häufig der technische Standort eines Servers wenig über den dort gespeicherten Inhalt aussagen. Daher sei an dieser Stelle nur auf die Arbeiten von [Buyukkokten et al. 1999], [Jones et al. 2002], [McCurley 2001] und [Markowetz et al. 2004] hingewiesen. Wir werden uns somit im Folgenden auf den Inhalt, die Linkstruktur und Metadaten konzentrieren.

1. *http://local.google.com/*
2. *http://de.search.yahoo.com/*

2.1 Inhaltsanalyse und Data Mining

2.1.1 Problemstellung und terminologische Grundlagen

Ansätze der Inhaltsanalyse und des Data Mining analysieren den Inhalt einer Webseite, um ihren geografischen Kontext zu erfassen. Als Quellen dienen dafür zum einen die in numerischer Form vorliegenden Adressangaben (z.B. Postleitzahlen) und Telefonnummern und zum anderen natürlichsprachlich angegebene geografische Ortsbezeichnungen wie beispielsweise Städtenamen, Regionen, Berge und Länder.

Die terminologische Bezeichnung der Phasen des Inhaltsanalyseprozesses zur Ermittlung des geografischen Kontextes ist in der Literatur nicht einheitlich. Allerdings lassen sich übereinstimmend die Teilschritte der Extraktion der Indikatoren für den geografischen Kontext, deren Weiterverarbeitung und geografische Zuordnung und schließlich eine sich daran anschließende Nutzbarmachung für darauf aufsetzende Anwendungen unterscheiden.

Die erste Phase beinhaltet eine Auswertung des Inhalts der Webseiten und die Extraktion von Indikatoren für den geografischen Kontext wie Adressangaben und Ortsbezeichnungen. In Übereinstimmung mit der von [McCurley 2001] und [Pouliquen et al. 2004] verwendeten Terminologie wird diese Phase im Folgenden als *Geoparsing* bezeichnet.

Im Anschluss daran erfolgt die Überprüfung und Weiterverarbeitung der extrahierten Indikatoren, insbesondere deren Einordnung in das geografische Koordinatensystem. Zusammen mit der nachfolgenden rechnerinternen Repräsentation des geografischen Kontextes und dessen Nutzbarmachung für Anwendungen (wie beispielsweise Visualisierung, Browsing und auch geografisch fokussierte Suche) stellen diese Teilschritte eine Form des *Geocoding* dar, die häufig als *Grounding* bezeichnet wird.

2.1.2 Erkennung von Adressangaben und Telefonnummern

Adressangaben in Form von Postleitzahlen und Straßennamen im Inhalt von Webseiten bilden einen wichtigen Indikator für deren geografischen Kontext. Problematisch gestaltet sich die Tatsache, dass die Formate für Adressen weltweit stark variieren. So unterscheidet sich beispielsweise das System der US Zip-Codes deutlich von dem der deutschen Postleitzahlen. Erschwerend kommt hinzu, dass Adressangaben auf Webseiten oft nur relativ angegeben sind, d. h. ohne Angabe des zugehörigen Landes.

Um dieses Problem und vor allem auch die Gefahr einer Verwechslung mit anderen numerischen Bezeichnungen wie beispielsweise Artikelnummern zu vermindern, schlagen [Markowetz et al. 2005] die zusätzliche Bestätigung von extrahierten Adressangaben mit validierenden Termen (*validator terms*) vor. Diese dienen dazu, die korrekte Erkennung eines geografischen Indikators zu bestätigen. Dies bedeutet, dass eine im Rahmen des Geoparsing erkannte Postleitzahl nur im Falle des Auftretens einer zugehörigen Stadt- oder Dorfbezeichnung bei Einhaltung eines gewissen Maximalabstandes im Text geokodiert wird. Dies dient vor allem der Vermeidung von *false-positives* und somit auch der Erhöhung der *Precision*[1] bei einer darauf aufbauenden geografisch fokussierten Suche.

Eine extrahierte Adressangabe muss daraufhin in ein geografisches Koordinatensystem eingeordnet werden. Für diesen Zweck werden spezielle Datenbanken eingesetzt. In Abhängigkeit von der gewünschten systeminternen Repräsentation des geografischen Kontextes können dabei nicht ausschließlich geografische Punktkoordinaten verwendet werden, zumal Postleitzahlen und Zip-Codes in der Realität unregelmäßig eingegrenzte Flächen von durchaus vielen Quadratkilometern umfassen können. Mit Hilfe einer vom *US Postal Service* entwickelten Datenbank, in der der jeweilige Geltungsbereich von Zip-Codes in Form von Polygonen gespeichert ist, wurde zumindest für die USA mittlerweile eine verbesserte Repräsentation von Regionen im Rahmen des Geocoding ermöglicht [McCurley 2001].

Weitere numerische Indikatoren für den geografischen Kontext von Webseiten sind dort enthaltene Telefonnummern. Diese sind hierarchisch strukturiert: Gemäß internationalen Standards sollten Telefonnummern mit dem länderspezifischen Code beginnen, gefolgt von einem ortsbezogenen Code und schließlich der individuellen Rufnummer. In der Realität werden allerdings durchaus viele unterschiedliche Formate zur Darstellung verwendet, weshalb für eine erfolgreiche Extraktion von Telefonnummern das Einpflegen von entsprechenden Regeln in den verwendeten Geoparser unerlässlich ist [McCurley 2001]. Die Gefahr einer Verwechslung von Telefonnummern mit anderen numerischen Zeichenfolgen ist aufgrund der Tatsache, dass auf diese oft keine weiteren geografischen Indikatoren wie Städtenamen im Text als validierende Terme folgen, wesentlich größer als bei Postleitzahlen.

Mit Hilfe des Geoparsing und Geocoding von Adressen und Telefonnummern kann nach [McCurley 2001] etwa 10% aller Webseiten ein geografischer Kontext zugeordnet werden.

2.1.3 Erkennung von geografischen Ortsbezeichnungen

Eine sehr wichtige Quelle für die geografische Einordnung von Ressourcen im Web stellt die Erkennung von geografischen Konzepten innerhalb ihres Inhalts dar. Natürlichsprachliche Texte auf Webseiten enthalten vielfach räumliche Referenzen – beispielsweise auf Städte, Berge, Seen, Regionen und Länder. Durch die Extraktion und Erkennung dieser geografischen Eigennamen (Toponyme) kann den jeweiligen Webseiten im Rahmen des Geocoding ein Kontext zugeordnet werden.

In der Phase des Geoparsing werden aus dem zu analysierenden Text einer Webseite diejenigen Terme extrahiert, die potenziell geografische Konzepte repräsentieren. Für diesen Zweck können einfache geografische Wörterbücher herangezogen werden, in denen die Bezeichnungen für geografische Orte hinterlegt sind.

Als Ergebnis dieses Teilschrittes wird der natürlichsprachliche Langtext auf eine Folge von Termen reduziert, die als Indikatoren für den geografischen Kontext dieser Webseite angesehen werden können. Dabei können jedoch zahlreiche Doppeldeutigkeiten auftreten.

Geografische vs. nicht geografische Bedeutung

Eine *Geo-nongeo-Ambiguität* bedeutet, dass die Zeichenfolge eines geografischen Ortes auch ein anderes, nicht geografisches Konzept repräsentieren kann.

1. Die Precision als Gütemaß bei Suchmaschinen gibt Auskunft über den Anteil relevanter Dokumente innerhalb der Suchresultate.

So kann die Zeichenfolge »Oder« in einem deutschen Text sowohl den Fluss als auch die Konjunktion bezeichnen. Häufig müssen auch Eigennamen von Toponymen unterschieden werden. Des Weiteren können Begriffe in unterschiedlichen Sprachen unterschiedliche Bedeutungen tragen. Für die Handhabung dieses Problems werden in der Literatur unterschiedliche Vorgehensweisen vorgeschlagen.

In dem von [Pouliquen et al. 2004] verwendeten Ansatz bildet das Erkennen der jeweils verwendeten Sprache den ersten Schritt zur Lösung dieses Problems. Um Mehrdeutigkeiten zwischen Termen verschiedener Sprachen zu verringern, werden beim Geoparsing nur diejenigen Begriffe extrahiert, die in der jeweiligen Textsprache oder in der lokal verwendeten Sprache ein geografisches Konzept repräsentieren. So würden in einem deutschen Text die Terme »Brüssel« und »Bruxelles« extrahiert werden, nicht jedoch die englische Bezeichnung »Brussels«.

Daneben versuchen andere Ansätze aus den Wörtern der Umgebung eines Begriffes Rückschlüsse darüber zu ziehen, ob dieser Begriff ein geografisches Konzept repräsentiert. [Densham & Reid 2003] verwenden etwa 300 verschiedene reguläre Ausdrücke, die aus Trainingsdaten gewonnen wurden, um potenzielle Toponyme zu identifizieren. In einem zweiten Durchgang wird versucht, *false-positives* zu eliminieren, indem die Wahrscheinlichkeit des gleichzeitigen Auftretens unterschiedlicher Ortsnamen betrachtet wird und außerdem Patterns angewendet werden, um Toponyme von Eigennamen zu unterscheiden. Zusätzlich werden Begriffe in der Umgebung eines Toponym-Kandidaten als Entscheidungshilfe verwendet (wie z.B. *river* oder *shire*), wobei ein *part-of-speech-tagger* zum Einsatz kommt (siehe [Brill 1994]).

Während beim Ansatz von [Pouliquen et al. 2004] die Auflösung dieser Mehrdeutigkeiten erst im Rahmen des Geocoding stattfindet, vertreten [Markowetz et al. 2005] die Auffassung, dass das Problem der *Geo-nongeo-Ambiguität* bereits vollständig bei der Extraktion potenzieller geografischer Eigennamen zu lösen ist. In ihrem Ansatz wird ähnlich wie in [Densham & Reid 2003] zu diesem Zweck bei der Erkennung von Städtebezeichnungen zwischen starken und schwachen Termen (*strong terms* und *weak terms*) unterschieden. Als stark werden alle Terme bezeichnet, die nahezu ausschließlich Städte repräsentieren. Schwache Terme repräsentieren nur in Verbindung mit starken ein geografisches Konzept. So bilden der schwache Term »Bad« und der starke Term »Kissingen« im Deutschen eine Mehrwortgruppe, die gemeinsam ein geografisches Konzept darstellt. »Kissingen« bezeichnet dabei ausschließlich die unterfränkische Stadt, wogegen »Bad« auch eine andere Semantik tragen kann. Schwache Terme werden zur Verringerung der *Geo-nongeo-Ambiguität* ausschließlich dann extrahiert, wenn sie in einem gewissen individuell festzulegenden Maximalabstand von einem mit ihnen assoziierten starken Term im Text vorkommen.

Zur Unterscheidung zwischen Eigennamen und geografischen Konzepten werden hierbei zusätzlich zu dem bereits eingeführten Begriff der validierenden Terme auch so genannte ausschließende Terme (*killer terms*) in die Entscheidung mit einbezogen, ob ein Indikator extrahiert werden soll oder nicht. Starke Terme, die mit einem validierenden Term verknüpft sind, werden nur dann extrahiert, wenn beide Terme innerhalb eines gewissen Maximalabstands auftreten. Umgekehrt wird ein starker Term ignoriert, wenn der zugehörige ausschließende Term in dessen unmittelbarer Nähe im Text auftritt. Kommen so genannte *general killers* wie beispielsweise »Herr«, »Frau« oder »Professor« im Text vor, so werden alle potenziellen geografischen Indikatoren in deren unmittelbarer Nähe ignoriert.

Bei allen Ansätzen bilden die extrahierten geografischen Indikatoren den Input für das anschließende Grounding. In diesem Teilschritt werden die übergebenen Terme auf physische Orte abgebildet und z.B. mit Hilfe von Längen- und Breitengraden in das räumliche Koordinatensystem eingeordnet.

Dafür werden oft spezielle Datenbanken, so genannte *Gazetteers*, eingesetzt (vgl. z.B. Alexandria Digital Library Gazetteer [Hill 2000]). Diese Gazetteers entsprechen geografischen Wörterbüchern, in denen Orte und deren physische Koordinaten auf der Erdoberfläche verzeichnet sind. Zusätzlich sollten möglichst einige Schreibweisenvarianten der Ortsbezeichnungen sowie Beziehungen zu anderen Konzepten, wie beispielsweise dem Land, in dem der Ort liegt, in den Datenbanken enthalten sein [Pouliquen et al. 2004].

Der Hauptbestandteil des Matching-Prozesses besteht darin, die extrahierten Toponyme in den Gazetteers nachzuschlagen und den darin enthaltenen physischen Positionen mit einer individuell festzulegenden Zutreffenswahrscheinlichkeit zuzuordnen. Diese Wahrscheinlichkeit gibt an, wie gut ein Indikator mit einem in der Datenbank enthaltenen Ortsbegriff übereinstimmt. Das Hauptproblem, das dazu führt, dass Bezeichnungen nicht immer eindeutig geografischen Orten zugeordnet werden können, ist das Auftreten einer so genannten *Geo-geo-Ambiguität* [Amitay et al. 2004].

Geografische Doppeldeutigkeit

Dieses Phänomen ist auf Homonyme zurückzuführen. »London« bezeichnet etwa die Hauptstadt Großbritanniens und auch eine Reihe anderer Kleinstädte auf dem Globus. Der Term »Victoria« vereinigt in sich sowohl das Problem der *Geo-nongeo-* als auch der *Geo-geo-Ambiguität*. So kann mit diesem Begriff ein Personenname, ein Stadtteil von Hongkong oder auch ein großer See in Afrika gemeint sein.

Für eine Auflösung der *Geo-geo-*Mehrdeutigkeit werden verschiedene Ansätze verfolgt. Eine grundlegende Annahme ist jedoch allen Ansätzen gemeinsam, nämlich die Hypothese des *single sense of discourse* (vgl. [Gale et al. 1992]). Demnach wird eine Bezeichnung in einem zusammenhängenden Text niemals für verschiedene Konzepte verwendet. Träfe diese Hypothese zu, bedeutete dies im Kontext des geografischen Information Retrieval, dass bei einer mehrmaligen Verwendung derselben geografischen Bezeichnung im Inhalt einer Webseite mit hoher Wahrscheinlichkeit immer derselbe physische Ort gemeint wäre.

Als Hilfsmittel zur Auflösung der *Geo-geo-Ambiguität* eines Toponyms können die anderen im Text der Webseite enthaltenen Indikatoren dienen. Können diese eindeutig einem physischen Ort zugeordnet werden, so ergeben sich daraus unter Umständen Anhaltspunkte, die für die Auflösung des mehrdeutigen Toponyms genutzt werden können. So ist es beispielsweise bei der Identifikation

der eindeutigen räumlichen Indikatoren »China« und »Hongkong« sehr wahrscheinlich, dass mit der Bezeichnung »Victoria« der zugehörige Stadtteil gemeint ist. *False-positives* sind hierbei selbstverständlich keineswegs auszuschließen. Diese Vorgehensweise ist natürlich von der Existenz zusätzlicher, identifizierbarer Toponyme im Text abhängig.

In [Leidner et al. 2003] werden dazu alle Toponyme in einem Gazetteer nachgeschlagen, der alle möglichen physischen Orte samt deren Koordinaten liefert. Gleichlautende Orte werden dabei in einer so genannten Konfusionsmenge zusammengefasst. In einem nächsten Schritt wird das kartesische Produkt aller Konfusionsmengen gebildet, das alle möglichen Kombinationen von eindeutigen Orten enthält. Nun wird anhand der Koordinaten für jedes Element ein Polygon berechnet und angenommen, dass jeweils die Orte in einem Dokument gemeint sind, die das kleinste Polygon bilden.

Andere Ansätze wie [Markowetz et al. 2005] oder [Rauch et al. 2003] verwenden zur Auflösung von Mehrdeutigkeiten statistische Methoden. Im Falle einer mehrdeutigen Zuordnung eines Terms zu verschiedenen geografischen Orten wird diese Bezeichnung beim Geocoding auf den größten oder wichtigsten der möglichen Orte gematcht. Zur Bestimmung der Bedeutsamkeit eines Ortes werden dabei beispielsweise die Einwohnerzahlen von Ortschaften herangezogen, die ebenfalls in Gazetteers enthalten sein können. Bei Vorkommen der Bezeichnung »London« wird diese beispielsweise der Hauptstadt von Großbritannien zugeordnet, sofern zusätzliche Indikatoren keinen Anlass für eine andere Einschätzung geben.

2.1.4 Repräsentation des geografischen Kontextes

Nach der geografischen Zuordnung der einzelnen im Inhalt einer Webseite enthaltenen Indikatoren ist eine geeignete Form der Repräsentation des gesamten geografischen Kontextes dieser Seite notwendig.

Da Webseiten Bezug auf mehrere unterschiedliche Locations nehmen können, wird anstelle einer Anzahl von Punktkoordinaten oft eine integrierte Form der Repräsentation dieser verschiedenen Indikatoren verwendet. In [Brinkhoff et al. 1993] werden einige Approximationen miteinander verglichen. Am weitesten verbreitet sind die so genannten MBRs (Minimum Bounding Rectangles, auch: Minimum Bounding Boxes), da sie aufgrund ihrer simplen Struktur besonders einfach und effizient zu handhaben sind. Daneben können auch polygonale und andere Repräsentationsformen (konvexe Hüllen, Ellipsen) verwendet werden, die jeweils spezifische Vor- und Nachteile für das jeweilige Anwendungsszenario aufweisen.

In diesem Zusammenhang wird für Ressourcen im Web in [Markowetz et al. 2005] auch von einem *geografischen Fußabdruck* einer Webseite gesprochen. In diesem Fußabdruck sind alle Locations enthalten, für die eine Webseite als relevant erachtet wurde. Die Relevanz wird dabei nicht binär, sondern in Form einer relativen Relevanzwahrscheinlichkeit für jede Location angegeben. Für die Repräsentation des geografischen Fußabdrucks wurde in diesem Ansatz ein geometrisches Rasterdatenmodell gewählt. Dabei wird die Erdoberfläche mit Hilfe eines Gitters abgebildet und für die zu indexierenden Webseiten die geografische Relevanzwahrscheinlichkeit für jedes einzelne Feld im Gitter angegeben. Diese Daten werden anschließend im Index abgelegt und können für eine Anfragebearbeitung genutzt werden.

2.1.5 Bestimmung des geografischen Fokus

Ein wesentlicher Aspekt ist neben der Location auch der geografische Fokus, d.h. der Gültigkeitsbereich einer Webseite. Dieser Gültigkeitsbereich (*geographical scope*) ist der räumlich begrenzte Umkreis, für den die Informationen einer Webseite relevant sind.

Für die Berechnung der Locality von Ressourcen im Web wurde ein auf die USA begrenzter Prototyp entwickelt, der ausgehend von den extrahierten geografischen Indikatoren den *scope* der Webseiten in die in Abbildung 1 dargestellte geografische Hierarchie einordnet [Ding et al. 2000]. Je umfassender die geografische Bedeutung einer Webseite ist, desto höher wird sie innerhalb der Hierarchie eingeordnet. Hat eine Webseite beispielsweise den Gültigkeitsbereich Kalifornien, so wird sie im zugehörigen Knoten CA der Hierarchie gespeichert. Dabei werden automatisch alle dazugehörigen Blattknoten, wie hier beispielsweise San Francisco und Los Angeles, mit erfasst.

Einen Algorithmus, der die räumliche Verteilung der in einer Webseite enthaltenen geografischen Referenzen nutzt, schlagen [Ding et al. 2000] für die Berechnung der Locality vor. Dabei ist für die Zuordnung des geografischen Gültigkeitsbereichs einer Webseite zu einem Knoten in der Hierarchie vor allem maßgeblich, dass ein sehr großer Anteil der im Inhalt der Webseite enthaltenen Indikatoren entweder auf den im Knoten repräsentierten geografischen Ort selbst oder auf dessen Söhne in der Baumhierarchie verweist.

Da die dargestellte Hierarchie auf Städte und Bundesstaaten innerhalb der USA begrenzt ist, wäre für eine präzisere Erfassung der Locality von Seiten im globalen Internet eine Ausweitung über die USA hinaus und ein Einbezug zusätzlicher geografischer Konzepte wie beispielsweise Seen, Berge, Regionen oder Ähnlichem notwendig.

2.2 Analyse der Linkstruktur

Mittels Verfahren der Inhaltsanalyse und des Data Mining kann durch die Extraktion numerischer und natürlichsprachlicher geografischer Indikatoren zahlreichen Seiten im Web ein geografischer Kontext zugewiesen werden. Für eine geografische Einordnung der von der Inhaltsanalyse noch nicht erfassten Webseiten und eine präzisere Berechnung der

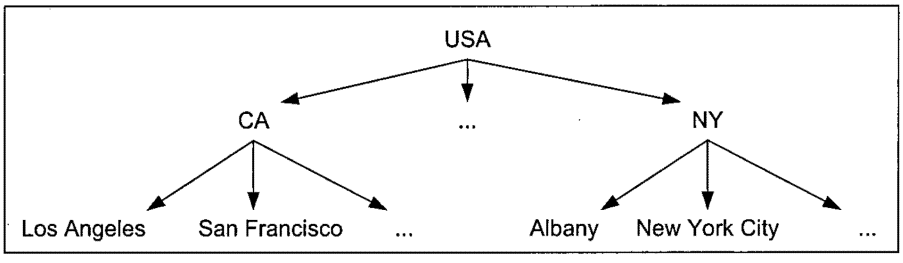

Abb. 1: Hierarchische Organisation von Locations [Ding et al. 2000]

Locality von Ressourcen im Web stellen Verfahren der Linkanalyse ein zusätzliches Instrument dar.

2.2.1 Ermittlung des impliziten geografischen Kontextes

Im Rahmen der Inhaltsanalyse kann Webseiten durch geografische Indikatoren ein expliziter räumlicher Kontext zugeordnet werden. Nicht alle Webseiten enthalten jedoch solche Indikatoren. Diesen Webseiten kann ein impliziter geografischer Kontext zugewiesen werden, wenn sie über Hyperlinks mit Webseiten mit explizitem geografischem Kontext verknüpft sind [McCurley 2001].

Diese Zuordnung basiert auf zwei Annahmen. Die erste Annahme lautet, dass sich zwei Webseiten, die miteinander verlinkt sind, ähnlicher sind als zwei zufällig ausgewählte Seiten im Web [Davison 2000]. Dies bedeutet, dass bei der Verlinkung von zwei Seiten auch eine hohe Wahrscheinlichkeit für einen ähnlichen geografischen Kontext besteht. Die zweite Hypothese besagt, dass zwei Webseiten innerhalb derselben *Website* eine höhere Ähnlichkeit zueinander aufweisen als zu anderen, zufällig ausgewählten Seiten des Internets. [Markowetz et al. 2005] fügen dem Geocoding deshalb eine weitere Phase hinzu, die als *geo propagation* bezeichnet wird.

Im Rahmen dieses Teilschritts wird Seiten ein impliziter geografischer Kontext von anderen Webseiten vererbt, wenn sie mit diesen über Links verbunden sind. Die Vererbung kann dabei über mehrere Links hinweg erfolgen, wobei die Wahrscheinlichkeit einer richtigen Zuordnung mit jedem weiteren Link abnimmt. Dies soll durch die Begrenzung der Anzahl der Vererbungsstufen und die Einführung eines Abschwächungsfaktors berücksichtigt werden, mit dem die Relevanzwahrscheinlichkeit für den impliziten Kontext im geografischen Fußabdruck der betreffenden Seiten angepasst wird.

Auf diese Weise kann nahezu allen Ressourcen im Web ein geografischer Kontext zugeordnet werden. Besonders hilfreich gestaltet sich dabei die Tatsache, dass in vielen Ländern für Websites eine spezielle Seite mit Kontaktinformationen der verantwortlichen Personen vorgeschrieben ist, das Impressum. Dieses kann im Rahmen der genannten Intrasite-Hypothese für die Zuordnung eines impliziten räumlichen Kontextes zu den anderen Seiten der Website genutzt werden.

2.2.2 Ermittlung der Locality von Webseiten

Eine Analyse der Linkstruktur bietet zudem weitere wertvolle Hinweise auf die Locality beziehungsweise den *geographical scope* von Webseiten. Zur Unterscheidung, ob eine Webseite hauptsächlich lokal, regional, national oder auch global relevant ist, bietet deren Linkstruktur ein geeignet erscheinendes Kriterium.

Die Grundannahme dieses Ansatzes ist, dass eine lokal relevante Webseite viele andere Webseiten in ihrer Umgebung referenziert und umgekehrt auch von vielen anderen räumlich nahen Webseiten referenziert wird [Markowetz et al. 2004]. Bei Webseiten von globalem Interesse wird sich dies in ähnlicher Form in einer Menge von weltweit gestreuten Verweisen auf diese Seite niederschlagen.

[Ding et al. 2000] machen ebenfalls von Verfahren der Linkanalyse Gebrauch, um den *geographical scope* einer Webseite in ihre in Abbildung 1 dargestellte Hierarchie einzuordnen. Eine Webseite w ist dabei für einen bestimmten räumlichen Bereich b relevant, wenn ein hoher Anteil der Webseiten, die auf w verweisen, ebenfalls dem Bereich b oder einem seiner Subbereiche zugeordnet ist. Dies wird auch als die Stärke des relativen räumlichen Interesses an der Webseite bezeichnet.

2.3 Metadaten und Ontologien

2.3.1 Metadaten

Im Rahmen des geografischen Information Retrieval wurde vorgeschlagen, dass Autoren im Code ihrer Webseiten zusätzliche Angaben über den geografischen Kontext dieser Seiten hinterlegen.

Der *Dublin Core Metadata Standard* sieht die Angabe von räumlichen Daten zur Beschreibung des geografischen Kontextes im *coverage tag* vor [Markowetz et al. 2004]. In diesem *Tag* kann die Location der Webseite mit einer geografischen Ortsbezeichnung, mit Punktkoordinaten oder auch in Form von polygonal begrenzten Regionen angegeben werden.

Ähnliches verfolgt auch der Ansatz der *geotags*. Mit Hilfe dieses speziell für die Belange des geografischen Information Retrieval entwickelten Ansatzes kann die Position der Locations einer Webseite noch genauer als im Dublin Core Standard spezifiziert werden.

Einen Schritt weiter geht der Standard einer *Geographical Markup Language* [McCurley 2001]. Mit dieser Unterspache von XML können nicht nur die räumlichen Daten der vorher genannten Standards angegeben werden, sondern es wird auch eine Modellierung der Struktur der geografischen Daten ermöglicht. Um bei einer Verarbeitung von Metadaten Schwierigkeiten zu vermeiden, die durch Ambiguitäten entstehen können, ist für den Autor bei der Angabe von Ortsbezeichnungen die Kenntnis der für einen Ort festgelegten Vorzugsbenennung nötig. Für diesen Zweck sind spezielle geografische Thesauri verfügbar, die eine terminologische Kontrolle der verwendeten Begriffe ermöglichen sollen.

Mit den Technologien des *Semantic Web*, deren Entwicklung und Standardisierung vom *World Wide Web Consortium* (W3C) vorangetrieben wird, entstehen gänzlich neue Möglichkeiten, Metadaten zu beschreiben und zu verwerten, die gerade für raumbezogene Metadaten von Bedeutung sind [Berners-Lee et al. 2001]. Für die semantische Suche verliert die – auf klassische Suchmaschinen zutreffende – Argumentation, Metadaten seien wegen zu hoher Kosten und zu leichter Manipulierbarkeit wenig interessant, ihre Berechtigung. Zwei Gründe sind hierfür ausschlaggebend, die sich schon auf der Ebene des *Resource Description Framework* (RDF), d.h. der sprachlichen Beschreibungsebene für Metadaten im Semantic Web, nachvollziehen lassen:

1. RDF-Metadaten sind maschinenlesbar und damit, zumindest prinzipiell, auch maschinell generierbar. Gerade der Raumbezug kann oft automatisch erfasst und zu niedrigen Kosten durch Metadaten beschrieben werden. Heute schon verbinden beispielsweise Softwarelösungen verschiedener Anbieter Digitalkameras mit GPS-Geräten und erzeugen so zu jedem Foto einen EXIF-Metadatensatz (*Exchangeable Image File Format*), der die geografischen Koordinaten des Aufnahmepunktes enthält und leicht in RDF konvertiert werden kann.

2. Mit RDF ist eine verteilte Metadatenhaltung möglich und intendiert. Damit kann im Prinzip jeder Metadaten zu einem Web-Dokument erstellen, nicht nur der Autor des Dokuments: »*Anyone should be able to freely add information about an existing resource, using any vocabulary they please*« [Manola et al. 2004]. Der Raumbezug einer Webseite *ex:webpage* lässt sich beispielsweise beschreiben durch das *coverage tag* von Dublin Core und eine Ontologie räumlicher Regionen *ont:*, die Regionen wie *ont:bavaria* oder *ont:franconia* definiert. Wenn der Autor der Webseite den Raumbezug selbst großzügig mit *ex:webpage dc:coverage ont:bavaria* angibt, kann dies an anderer Stelle mit *ex:webpage dc:coverage ont:franconia* korrigiert werden. Eine Manipulation durch den Autor wird so erheblich erschwert.

Die semantische Herausforderung bei der Zusammenführung verteilter Metadaten oder bei der Bearbeitung von Suchanfragen über verteilten Datenbeständen besteht darin, dass diese in verschiedenen Ontologien beschrieben sein können. Man kann z.B. kaum davon ausgehen, dass eine geografische Bezeichnung wie »Süddeutschland« im Web einheitlich konzeptualisiert ist. Für Übersetzungen zwischen Ontologien verwendet man Inferenzdienste, die die auf RDF folgende sprachliche Beschreibungsebene des Semantic Web bereitstellt, nämlich die Modellierungsprache *Ontology Web Language* (OWL) [McGuinness & van Harmelen 2004] zusammen mit einem geeigneten Theorembeweiser.

Raumbezogene Metadaten erfordern spezielle Inferenzdienste, weil geometrische Beziehungen Teil ihrer Semantik sind. Egenhofer spricht in Bezug auf diese spezielle semantische Funktionalität vom *Geospatial Semantic Web* [Egenhofer 2002]. Eine wichtige Aufgabe besteht darin, die Funktionalität der Ontologien mit der Funktionalität von GIS zu verbinden, um die ontologische Modellierung von numerischen Berechnungen zu entlasten. Einen Überblick über aktuelle Forschungsarbeiten auf dem Gebiet vermittelt [Rodriguez et al. 2005]. Noch haben sich die Lösungsansätze nicht in Standards verfestigt – derzeit wird aber unter dem Arbeitstitel »OWL-SPACE« an einer Ontologie für räumliche Konzepte gearbeitet.

2.3.2 Ontologische Modellierung räumlicher Konzepte

Das Ziel weiterführender Ansätze ist es somit, die räumliche Position der Locations auf Webseiten nicht nur angeben zu können, sondern durch die Modellierung der Struktur des geografischen Kontextes diesen auf einer höheren semantischen Ebene für Maschinen »verständlich« zu machen. Erreicht werden soll dieses Ziel durch die Verwendung von geografischen Ontologien.

[Jones et al. 2001] stellen in ihrem Ansatz ein Konzept einer geografischen Ontologie vor, mit deren Hilfe es ermöglicht werden soll, bei der Anfragebearbeitung nicht nur die Anfrageterme selbst, sondern auch das dahinter liegende Informationsbedürfnis des Nutzers zu berücksichtigen. Diese Ontologie (siehe Abb. 2) modelliert die Struktur der geografischen Konzepte, sowohl in Form von quantitativen als auch qualitativen Daten. Auf ähnliche Weise sind grundlegende geografische und geoinformatische Konzepte in verschiedenen Top-Level-Ontologien formalisiert. Am umfangreichsten sind die Interoperabilitätsstandards des Open Geospatial Consortium (OGC), insbesondere die Simple Feature Specification [Ryden 2005], die geometrische Grundkonzepte definiert.

Die terminologischen Bezeichnungen von geografischen Orten werden in einem kontrollierten Vokabular abgelegt und Vorzugsbenennungen für diese definiert. Die Position von räumlichen Konzepten wird mit geografischen Koordinaten festgelegt und die Beziehungen zu anderen räumlichen Konzepten qualitativ formuliert. Dabei wird zwischen der Taxonomie (Subklassenbeziehung) und der Partonomie (Ganzes-Teile-Beziehung) unterschieden.

In einem Prototyp, der diese Konzepte umsetzt – dem *Ontologically-Augmented Spatial Information System* (OASIS) –, werden auch die Historie von geografischen Orten und der räumliche Kontext von bedeutenden kulturellen oder geschichtlichen Artefakten modelliert. Es werden Dokumente präsentiert, die thematisch zur Anfrage relevant sind und deren räumlicher geografischer Kontext sich in geringem euklidschem Abstand zum Ortsbezug der Anfrage befindet.

Eine ähnliche Vorgehensweise wird auch im *SPIRIT-Projekt* eingeschlagen [Jones et al. 2002]. Ziel dabei ist die Erweiterung herkömmlicher geografischer Gazetteers und Thesauri um weitere Eigenschaften wie räumliche Beziehungen, um die darin enthaltenen Informationen über das geografische Vokabular und die Struktur geografischer Konzepte in einer Ontologie abzulegen. Auch hier sollen bei der Bearbeitung einer Anfrage Dokumente sowohl mit synonym bezeichneten als auch räumlich nahen geografischen Orten gefunden werden.

Zu diesem Zweck werden bekannte Methoden der Anfrageerweiterung herangezogen (wie sie auch auf Basis von Thesauri bekannt sind), um die in der Anfrage des Nutzers spezifizierten Orte um Varianten, Synonyme oder Ähnliches zu ergänzen. Um den hierfür benötigten geografischen Kontext der Webseiten für die Suche zugreifbar zu machen, wird dieser mit Hilfe von Ansätzen des maschinellen Lernens automatisiert aus den Seiten extrahiert und in Form von Metadaten hinzugefügt.

3 Anfragebearbeitung bei einer geografisch fokussierten Suche

Nach den oben dargestellten Techniken, um den geografischen Kontext der Webseiten im Internet zu ermitteln und für die geografisch fokussierte Suche nutzbar zu machen, soll im Folgenden ein kurzer Überblick über spezielle Probleme und Lösungsansätze der Anfragebearbeitung von geofokussierten Suchmaschinen gegeben werden. Dabei werden vor allem Fragestellungen der Anfragerepräsentation, des Rankings und der Präsentation des Anfrageergebnisses beleuchtet.

3.1 Problemstellungen des Rankings

Bei einer geografisch fokussierten Suche stellt sich für das Ranking im Vergleich zu klassischen Suchmaschinen ein zusätzliches Problem. Denn neben einer thematischen Komponente der Anfrage ist auch der geografische Aspekt im Ranking zu berücksichtigen.

Für ein Ranking der Ergebnisse muss die Suchmaschine nun sowohl die Relevanz der Dokumente zum gesuchten Thema als auch zum gewünschten geografi-

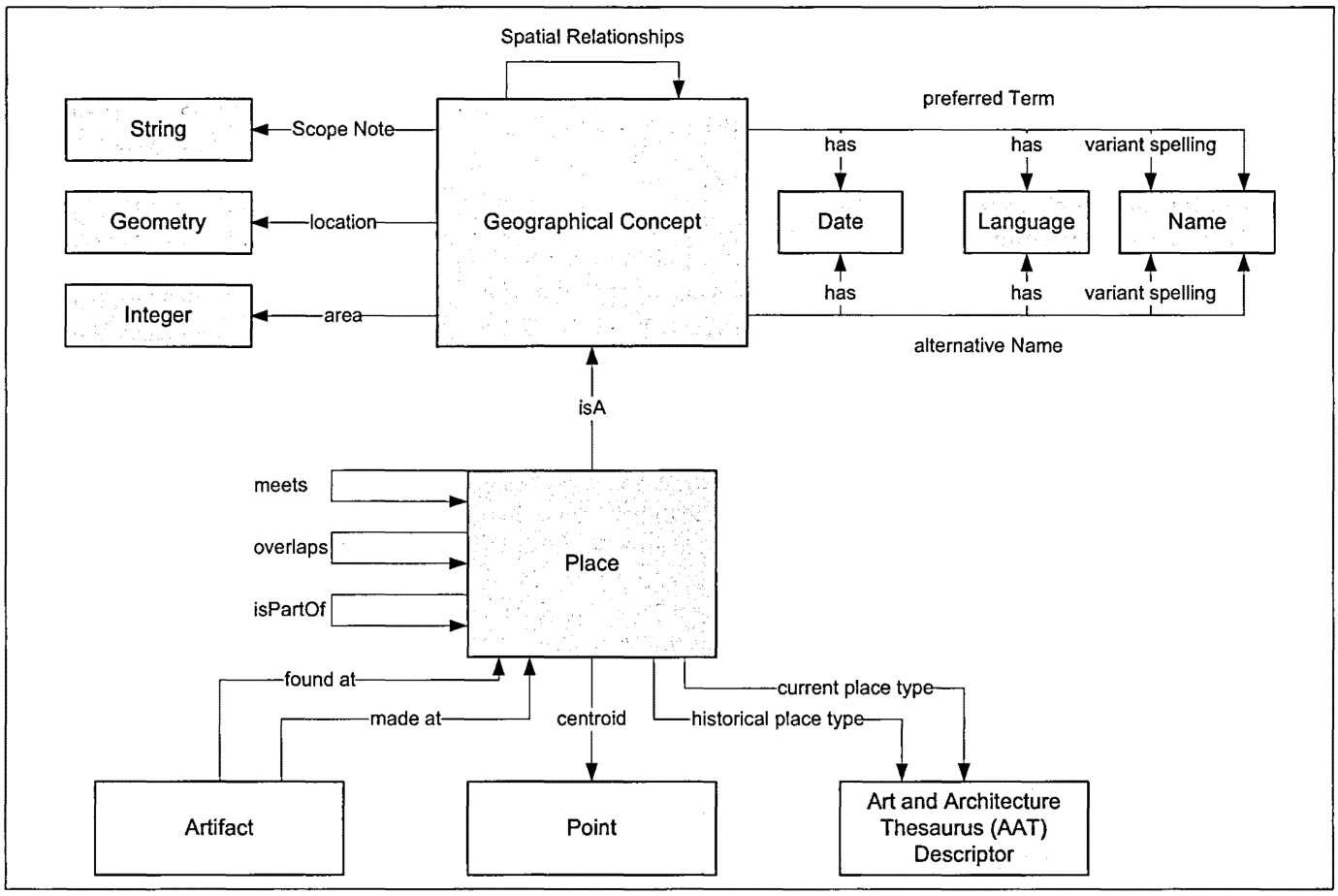

Abb. 2: Modellierung von geografischen Konzepten (nach [Jones et al. 2001])

schen Kontext berücksichtigen. In der Literatur werden dazu verschiedene, zueinander ähnliche Vorgehensweisen vorgeschlagen.

Dabei können grundsätzlich drei verschiedene Rankingmodelle zum Einsatz kommen. Zunächst kann der geografische Aspekt der Anfrage als Filterkriterium dienen. Dabei werden alle Dokumente hinsichtlich ihrer inhaltlichen Übereinstimmung zur Anfrage bewertet, deren geografischer Fokus eine gewisse Mindestübereinstimmung mit der geografischen Komponente der Anfrage aufweist. Umgekehrt kann auch die inhaltliche Komponente als Filter dienen, so dass nur Dokumente nach geografischer Ähnlichkeit gerankt werden, die ein Mindestmaß an inhaltlicher Übereinstimmung zur Anfrage aufweisen. Als drittes generisches Modell kann ein Mischranking zum Einsatz kommen, bei dem beide Komponenten gewichtet in das Ergebnisranking einfließen.

In [Markowetz et al. 2004] wird ein als *dynamisches Balancieren* bezeichneter Ansatz vorgestellt, der ein filterbasiertes Ranking verwendet: Bei der Berechnung des Anfrageergebnisses findet der Anfrageprozessor im Index r Dokumente, die sowohl thematisch als auch räumlich für die Anfrage relevant wären. Von diesen r Dokumenten werden die n zum gesuchten Thema relevantesten Webseiten dem Nutzer präsentiert. Das Ranking dieser n Webseiten erfolgt gemäß der räumlichen Nähe zum gewünschten geografischen Kontext. Das Balancieren erfolgt dabei durch Festsetzung der beiden Parameter, was dem Nutzer als Instrument dienen kann, seine individuelle Präferenzen auszudrücken.

Die Rankingverfahren für räumliche Nähe hängen dabei stark von der verwendeten Repräsentation des geografischen Kontextes ab. Bei Punktkoordinaten kann im einfachsten Fall z.B. die euklidische Distanz als lineares Distanzmaß zum Einsatz kommen. Bei anderen räumlichen Repräsentationen kann die Überlappung berücksichtigt werden. Binäre Betrachtungen, ob eine Überlappung vorliegt, ob die Anfragerepräsentation vollständig innerhalb der Dokumentrepräsentation liegt oder anders herum, liefern dabei für ein sortierendes Ranking allerdings ungenügende Informationen. Dazu können Maße verwendet werden, die den Grad der Überlappung ermitteln (vgl. z.B. [Beard & Sharma 1997], [Walker et al. 1992]).

In [Larson & Frontiera 2004] wird eine Übersicht über GeoIR-Rankingverfahren gegeben und eine Evaluierung vorgenommen. Dort wird auch ein weiterer Ansatz vorgestellt, der die Idee des probabilistischen Rankings verfolgt.

Der erwähnte Ansatz des dynamischen Balancierens wird in [Markowetz et al. 2005] weiterentwickelt. Die *thematische* Relevanz eines Dokumentes zur Nutzeranfrage wird mit einem so genannten Dokumenten-Fußabdruck angegeben. Der gewünschte geografische Kontext wird in Form eines *räumlichen* Fußabdrucks formuliert. Dabei besteht die Möglichkeit, entweder einen scharf abgegrenzten oder auch einen fließend übergehenden räumlichen Bereich für den zu suchenden geografischen Fußabdruck zu wählen. Die Suchmaschine liefert Webseiten aus der Schnittmenge des geografischen und des thematischen Fußabdrucks zurück, wobei das Ranking des Anfrage-

ergebnisses gemäß einer Punktzahl erfolgt, die mit Hilfe der beiden gewichteten »Fußabdrücke« berechnet wird. Der Nutzer hat dabei die Wahl, entweder den geografischen oder den thematischen Fußabdruck bei der Berechnung stärker zu gewichten und so eine für ihn nützliche Balance der beiden Suchkriterien zu erzielen.

Sowohl für die Formulierung des Informationswunsches als auch für die Präsentation des Anfrageergebnisses bestehen bei der geografisch fokussierten Suche verschiedene Darstellungs- und Visualisierungsmöglichkeiten, über die im folgenden Abschnitt ein kurzer Überblick gegeben werden soll.

3.2 Navigation und Visualisierung

Während bei klassischen Suchmaschinen die Anfrage vom Nutzer in ein Textfeld eingegeben und das Ergebnis in Form einer eindimensionalen Liste der gefundenen Webseiten präsentiert wird, bieten sich für die geografisch fokussierte Suche andere Möglichkeiten zur Gestaltung der Nutzeroberfläche an.

Für den Zugriff auf Informationen mit geografischem Charakter haben sich schon in der Vergangenheit *multi-modale Nutzeroberflächen* bewährt [Jones et al. 2002]. Dabei besteht für den Nutzer die Möglichkeit, entweder textuell oder grafisch navigierend mit der Suchmaschine zu kommunizieren. Der gewünschte räumliche Kontext kann textuell mit Hilfe geografischer Ortsbezeichnungen oder durch Operationen auf interaktiven Karten an die Suchmaschine übergeben werden. Der Anwender sollte räumliche Schwerpunkte der Suche durch das Markieren von Orten auf der Karte setzen, Areas-of-Interest definieren und Zoom-Operationen durchführen können, um den Detaillierungsgrad der dargestellten Informationen zu beeinflussen.

Auch die Ergebnisse einer Suche können auf einer Karte dargestellt werden, wobei sowohl die räumliche Abdeckung als auch der inhaltliche Relevanzgrad visualisiert werden können. Wenn Anfrage und Ergebnispräsentation dieselbe Kartenfunktion nutzen, kann dem Benutzer eine Browsingmöglichkeit geboten werden.

Für die Umsetzung wird in [McCurley 2001] und [Pouliquen et al. 2004] die besondere Eignung des SVG-Formats für die Darstellung interaktiver Karten im Rahmen der geofokussierten Suche herausgestellt. So könnten auf SVG-Grafiken Zoom-Operationen des Nutzers auf der Clientseite ohne Neuberechnung der Karte auf dem Server umgesetzt werden. [McCurley01] schlägt vor, die Vorteile des Hypertextes für einen räumlich navigierenden Zugriff auf Webseiten zu nutzen. So bestünde für den Nutzer die Möglichkeit, sich zu einem Suchergebnis weitere Webseiten anzeigen zu lassen, die einen ähnlichen geografischen Kontext besitzen. In dem von ihnen verwendeten Browser wird durch das Betätigen eines »What's nearby?«-Buttons neben dem traditionellen Browserfenster eine interaktive Karte geöffnet, auf welcher der Nutzer zu räumlich ähnlichen Webseiten navigieren kann.

4 Zusammenfassung und Ausblick

In diesem Artikel wurden die grundsätzlichen Aspekte einer geografischen Suche vorgestellt. Geografisches Information Retrieval ist ein verhältnismäßig junges Forschungsfeld, dem jedoch zunehmend Interesse gewidmet wird. Wesentliche Schwerpunkte der Forschung sind dabei die Erfassung eines geografischen Kontextes von Dokumenten, das Erstellen und Nutzen geografischer Ontologien, die Konzeption einer Nutzerschnittstelle sowie das Ranking von Ergebnisdokumenten zu Anfragen, die sowohl geografische als auch thematische Anforderungen enthalten.

Der *Workshop on Geographic Information Retrieval* [GIR 2005], der 2005 zum zweiten Mal stattfand, ist eine der Plattformen zum Informationsaustausch zwischen Forschenden und Anwendern in diesem Bereich. Aktuell werden dort die Themenbereiche Visualisierung in GIR, Komponenten von GIR-Systemen, Anwendungen von GIR und neue Ansätze im geografischen Information Retrieval behandelt.

Literatur

[Amitay et al. 2004] *Amitay, Einat; Nadav, Har'El; Sivan, Ron; Soffer, Aya:* Web-a-where: geotagging web content. In: SIGIR '04: Proceedings of the 27th annual international conference on Research and development in information retrieval, New York, NY, USA, ACM Press, 2004, S. 273-280.

[Beard & Sharma 1997] *Beard, Kate; Sharma, Vyjayanti:* Multidimensional ranking for data in digital spatial libraries. Int. J. on Digital Libraries, 1(2):153-160, 1997.

[Berners-Lee et al. 2001] *Berners-Lee, Tim; Hendler, James; Lassila, Ora:* The semantic web. Scientific American, May 2001, S. 35-43.

[Brill 1994] *Brill, Eric:* Some advances in transformation-based part of speech tagging. In: AAAI '94: Proceedings of the 12th national conference on Artificial intelligence (vol. 1), Menlo Park, CA, USA, American Association for Artificial Intelligence, 1994, S. 722-727.

[Brin & Page 1998] *Brin, Sergey; Page, Lawrence:* The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1-7):107-117, 1998.

[Brinkhoff et al. 1993] *Brinkhoff, Thomas; Kriegel, Hans-Peter; Schneider, Ralf:* Comparison of approximations of complex objects used for approximation-based query processing in spatial database systems. In: ICDE, 1993, S. 40-49.

[Buyukkokten et al. 1999] *Buyukkokten, Orkut; Cho, Junghoo; Garcia-Molina, Hector; Gravano, Luis; Shivakumar, Narayanan:* Exploiting geographical location information of web pages. In: WebDB (Informal Proceedings), 1999, S. 91-96.

[Davison 2000] *Davison, Brian D.:* Topical locality in the web. In: SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM Press, 2000, S. 272-279.

[Densham & Reid 2003] *Densham, Jan; Reid, James:* A geo-coding service encompassing a geo-parsing tool and integrated digital gazetteer service. In: Workshop on the Analysis of Geographic References at the NAACL-HLT 2003 conference, Edmonton, Canada, May 2003.

[Ding et al. 2000] *Ding, Junyan; Gravano, Luis; Shivakumar, Narayanan:* Computing geographical scopes of web resources. In: 26th International Conference on Very Large Databases, VLDB 2000, Cairo, Egypt, September 10-14, 2000.

[Egenhofer 2002] *Egenhofer, Max J.:* Toward the semantic geospatial web. In: GIS '02: Proceedings of the 10th ACM international symposium on Advances in geographic information systems, New York, NY, USA, ACM Press, 2002, S. 1-4.

[Gale et al. 1992] *Gale, William A.; Church, Kenneth W.; Yarowsky, David:* One sense per discourse. In: HLT '91: Proceedings of the workshop on Speech and Natural Language, Morristown, NJ, USA, Association for Computational Linguistics, 1992, S. 233-237.

[GIR 2005] GIR '05: Proceedings of the 2005 workshop on geographic information retrieval, 2005. General Chair-Chris Jones and General Chair-Ross Purves.

[Gravano et al. 2003] *Gravano, Luis; Hatzivassiloglou, Vasileios; Lichtenstein, Richard:* Categorizing web queries according to geographical locality. In: CIKM '03: Proceedings of

the 12th international conference on Information and knowledge management, New York, NY, USA, ACM Press, 2003, S. 325-333.

[Hill 2000] *Hill, Linda L.:* Core elements of digital gazetteers: Placenames, categories, and footprints. In: ECDL '00: Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries, London, UK, Springer-Verlag, 2000, S. 280-290.

[Jones et al. 2001] *Jones, Christopher B.; Alani, Harith; Tudhop, Douglas:* Geographical information retrieval with ontologies of place. Lecture Notes in Computer Science, 2205:322-335, 2001.

[Jones et al. 2002] *Jones, C.; Purves, R.; Ruas, A.; Sanderson, M.; Sester, M.; van Kreveld, M.; Weibel, R.:* Spatial information retrieval and geographical ontologies – an overview of the spirit project, 2002.

[Larson & Frontiera 2004] *Larson, Ray R.; Frontiera, Patricia:* Spatial ranking methods for geographic information retrieval (gir) in digital libraries. In: Rachel Heery; Liz Lyon (eds.), Research and Advanced Technology for Digital Libraries: 8th European Conference, ECDL 2004, Lecture Notes in Computer Science Series, LNCS 3232, Springer-Verlag, 2004, S. 45-57.

[Leidner et al. 2003] *Leidner, Jochen L.; Sinclair, Gail; Webber, Bonnie:* Counding spatial named entities for information extraction and question answering. In: Proceedings of the Workshop on the Analysis of Geographic References held at the Joint Conference for Human Language Technology and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics 2003 (HLT/NAACL'03), Edmonton, Alberta, Canada, May 2003, S. 31-38.

[Manola et al. 2004] *Manola, Frank; Miller, Eric; McBride, Brian:* Rdf primer. W3c recommendation, World Wide Web Consortium, February 2004; *www.w3.org/TR/rdf-primer/.*

[Markowetz et al. 2004] *Markowetz, Alexander; Brinkhoff, Thomas; Seeger, Bernhard:* Geographic information retrieval. 3rd International Workshop on Web Dynamics, 2004.

[Markowetz et al. 2005] *Markowetz, Alexander; Chen, Yen-Yu; Suel, Torsten; Long, Xiaohui; Seeger, Bernhard:* Design and implementation of a geographic search engine. In: WebDB, 2005, S. 19-24.

[McCurley 2001] *McCurley, Kevin S.:* Geospatial mapping and navigation of the web. In: WWW '01: Proceedings of the 10th international conference on World Wide Web, New York, NY, USA, ACM Press, 2001, S. 221-229.

[McGuinness & van Harmelen 2004] *McGuinness, Deborah L.; van Harmelen, Frank:* Owl web ontology language overview. W3c recommendation, World Wide Web Consortium, February 2004; *www.w3.org/TR/owl-features/.*

[Pouliquen et al. 2004] *Pouliquen, Bruno; Steinberger, Ralf; Ignat, Camelia; De Groeve, Tom:* Geographical information recognition and visualization in texts written in various languages. In: SAC '04: Proceedings of the 2004 ACM symposium on Applied computing, New York, NY, USA, ACM Press, 2004, S. 1051-1058.

[Rauch et al. 2003] *Rauch, Erik; Bukatin, Michael; Baker, Kenneth:* A confidence-based framework for disambiguating geographic terms. In: HLT-NAACL 2003 Workshop on Analysis of Geographic References, 2003, S. 50-54.

[Rodriguez et al. 2005] *Rodríguez, M. Andrea; Cruz, Isabel F.; Egenhofer, Max J.; Levashkin, Sergei (eds.):* GeoSpatial Semantics, First International Conference, GeoS 2005, Mexico City, Mexico, November 29-30, 2005, Proceedings, volume 3799 of Lecture Notes in Computer Science, Springer-Verlag, 2005.

[Ryden 2005] *Ryden, Keith:* Opengis implementation specification for geographic information – simple feature access – part 1: Common architecture (= iso 19125), version 1.1.0. Specification of the open geospatial consortium, World Wide Web Consortium, February 2005; *www.opengeospatial.org.*

[Sanderson & Kohler 2004] *Sanderson, Mark; Kohler, Janet:* Analyzing geographic queries. In: Proceedings of the ACM SIGIR Workshop on Geographic Information Retrieval, Sheffield, UK, 2004.

[Walker et al. 1992] *Walker, D.; Newman, I.; Medyckyj-Scott, D.; Ruggles, C.:* A system for identifying datasets for gis users. In: International Journal of Geographical Information Systems, volume 6, 992, S. 511-527.

[Wang et al. 2005] *Wang, Chuang; Xie, Xing; Wang, Lee; Lu, Yansheng; Ma, Wei-Ying:* Detecting geographic locations from web resources. In: GIR '05: Proceedings of the 2005 workshop on Geographic information retrieval, New York, NY, USA, ACM Press, 2005, S. 17-24.

Jens Gräf studiert Wirtschaftsinformatik an der Otto-Friedrich-Universität Bamberg. Er schreibt seine Diplomarbeit im Bereich des geografischen Information Retrieval.

Andreas Henrich ist Inhaber des Lehrstuhls Medieninformatik an der Otto-Friedrich-Universität Bamberg. Seine Forschungsschwerpunkte liegen im Bereich der Entwicklung multimedialer Anwendungen sowie im Information Retrieval (IR). Im IR richten sich die Arbeiten insbeondere auf die Berücksichtigung von Kontextwissen, die effiziente Anfragebearbeitung sowie die Suche in P2P-Systemen.

Volker Lüdecke studierte Wirtschaftsinformatik an der Otto-Friedrich-Universität Bamberg. Seit seinem Abschluss 2003 ist er als wissenschaftlicher Mitarbeiter bei Prof. Henrich tätig. Er forscht im Bereich des geografischen Information Retrieval.

Christoph Schlieder ist Inhaber des Lehrstuhls für Angewandte Informatik in den Kultur-, Geschichts- und Geowissenschaften an der Otto-Friedrich-Universität Bamberg. Seine Forschungsarbeiten befassen sich mit der Entwicklung semantischer Informationstechnologien (z.B. Geospatial Semantic Web) für kulturwissenschaftliche Anwendungen.

Jens Gräf
Prof. Dr. Andreas Henrich
Dipl.-Wirtsch.-Inf. Volker Lüdecke
Otto-Friedrich-Universität Bamberg
Fakultät für Wirtschaftsinformatik und Angewandte Informatik
Lehrstuhl für Medieninformatik
Feldkirchenstr. 21
96045 Bamberg
jens.graef@stud.uni-bamberg.de
{andreas.henrich, volker.luedecke}@wiai.uni-bamberg.de
www.uni-bamberg.de/wiai/minf

Prof. Dr. Christoph Schlieder
Otto-Friedrich-Universität Bamberg
Fakultät für Wirtschaftsinformatik und Angewandte Informatik
Lehrstuhl für Angewandte Informatik in den Kultur-, Geschichts- und Geowissenschaften
Feldkirchenstr. 21
96045 Bamberg
christoph.schlieder@wiai.uni-bamberg.de
www.kinf.wiai.uni-bamberg.de/