Erschienen in: Proceedings Münsteraner GI-Tage 2003, S. 181-194.

Geodienste im Rahmen der Architektur einer Geo-Suchmaschine

Thomas Brinkhoff
Institut für Angewandte Photogrammetrie und Geoinformatik (IAPG)
Fachhochschule Oldenburg/Ostfriesland/Wilhelmshaven
Kompetenzzentrum für Geoinformatik in Niedersachsen (GiN)
Ofener Str. 16/19, D-26121 Oldenburg
E-Mail: Thomas.Brinkhoff@fh-oldenburg.de
http://www.fh-oow.de/institute/iapg/personen/brinkhoff/

ZUSAMMENFASSUNG

In diesem Beitrag wird ein Geodienst namens „Page2Loc“ vorgestellt, der im Rahmen der Architektur einer Geo-Suchmaschine die Aufgabe hat, für Webseiten die geographischen Geltungsbereiche zu bestimmen. Dazu werden eine Reihe aus der Literatur bekannter Verfahren betrachtet und bewertet. Vier dieser Ansätze werden auf Basis dieser Bewertung abgestuft in den Dienst „Page2Loc“ aufgenommen. Für eine Implementierung dieses Dienstes benötigt man wiederum andere Geodienste, um Ortsbezeichnungen, Postleitzahl- und Vorwahlangaben zu vereinheitlichen und insbesondere auf eine geometrische Beschreibung der Lage und Ausdehnung abzubilden. Zusätzlich können die Ergebnisse einer Link-Analyse berücksichtigt werden, um die so bestimmten Geltungsbereiche zu gewichten. Die Link-Analyse nutzt ihrerseits den Dienst „Page2Loc“.

EINLEITUNG

Während das Internet in seiner Anfangszeit überwiegend globale oder zumindest national relevante Informationen und Dienste bereitgestellt hat, wandelte sich das Informations- und Dienstangebot in den letzten Jahren: Inzwischen dominieren Angebote mit regionaler oder nur lokaler Relevanz. Das typische Beispiel ist der Pizzaservice, der die Bestellung und Auslieferung von Pizzas über das Internet erlaubt. Diese Entwicklung hat Konsequenzen für die Anfragen, die von einer Internet-Suchmaschine zu bearbeiten sind. Im folgenden Beispiel (Markowetz et al., 2003) soll für ein Mountainbike der Marke „Cannondale“ ein Betrieb gesucht werden, bei dem man das Rad zur Inspektion geben kann. Eine Suche bei Google nach „Cannondale“ liefert weltweit über 200.000 Ergebnisse. Schränkt man die Suche auf deutsche Web Sites ein, verbleiben über 6.000 Antworten – offenkundig immer noch zu viele Ergebnisse. Die Suche nach „Cannondale AND Marburg“ beschränkt auf deutsche Angebote liefert 19 Ergebnisse. Dies scheint eine handhabbare Antwortgröße zu sein. Leider sind aber alle 19 Antworten verkehrt, da es in Marburg keinen entsprechenden Fachbetrieb gibt; die Begriffe Cannondale und Marburg befinden sich mehr oder weniger nur zufällig auf den gefundenen Seiten. Die nächstgelegene Firma in Allendorf wird hingegen durch eine solche Suche nicht entdeckt.

Dieses Beispiel macht deutlich, dass Webangebote mit regionaler oder lokaler Relevanz Suchmaschinen erfordern, die in der Lage sind, räumliche Informationen – sowohl auf der Anfrage- als auch auf der Angebotsseite – zu berücksichtigen. Auf der Anfrageseite kann eine solche *Geo-Suchmaschine* entsprechende Eingabefelder vorsehen bzw. im Rahmen von Location-Based Services die aktuelle Position des Anfragenden auswerten. Auf der Angebotsseite sieht eine Lösung auf den ersten Blick recht einfach aus: Die naheliegendste Lösung ist die Ergänzung der Webseiten durch entsprechende Meta-Informationen: Dazu ergänzt der Anbieter seine Webseiten durch *Geo-Tags*, die den *geographischen Geltungsbereich* einer Webseite oder einer Web Site beschreiben. Die bisherigen Erfahrungen zeigen allerdings, dass sich standardisierte Meta-Informationen nur äußerst schwierig und langsam auf breiter Front durchsetzen lassen. Auch neigen Anbieter dazu, solche Informationen zu verfälschen, um sich (vermeintliche) Wettbewerbsvorteile zu verschaffen. Daher sind weitergehende Ansätze notwendig, um den geographischen Geltungsbereich von Webseiten zu bestimmen. Eine Übersicht über solche Verfahren wird etwas später im Abschnitt „Techniken für die Bestimmung des geographischen Geltungsbereichs von Webseiten“ gegeben.

Nach einer kurzen Skizze der „Architektur einer Geo-Suchmaschine“ wird in dem vorliegenden Beitrag der Geodienst „Page2Loc“ näher betrachtet. Für diesen Dienst sind die vorgestellten Verfahren für die Bestimmung des geographischen Geltungsbereichs von Webseiten bewertet und vier Verfahren ausgewählt worden. Diese werden abgestuft – wie in dem Unterabschnitt „Verarbeitung“ dargestellt – in den Dienst „Page2Loc“ integriert.

Eine wichtige Aufgabe im Rahmen von „Page2Loc“ ist die Vereinheitlichung der Resultate der verschiedenen Verfahren und damit von geographischen Geltungsbereichen. Dazu sind zweckmäßigerweise andere Geodienste erforderlich, die Ortsbezeichnungen, Postleitzahl- und Vorwahlangaben vereinheitlichen und auf geometrische Beschreibungen der Lage und Ausdehnung abbilden. Daneben wird eine „Link-Analyse“ in den vorgestellten Geodienst integriert, um Geltungsbereiche gewichten zu können.

Der Beitrag endet mit einer Zusammenfassung und einem Ausblick auf zukünftige Arbeiten.

TECHNIKEN FÜR DIE BESTIMMUNG DES GEOGRAPHISCHEN GELTUNGSBEREICHS VON WEBSEITEN

Die Einleitung hat deutlich gemacht, dass Webangebote mit regionaler oder lokaler Relevanz den Einsatz von Geo-Suchmaschinen erfordern, die in der Lage sind, geographische Informationen – sowohl auf der Anfrage- als auch auf der Angebotsseite – zu berücksichtigen. Für die angebotsseitige Bestimmung des geographischen Geltungsbereichs wurden in der Literatur eine Reihe verschiedener Ansätze vorgeschlagen, die an dieser Stelle in einem Überblick vorgestellt werden sollen.

Geo-Tags

Eine naheliegende Lösung ist die Ergänzung der Webseiten durch entsprechende Meta-Informationen: Dazu ergänzt der Anbieter seine Webseiten durch *Geo-Tags*, die den *geographischen Geltungsbereich* einer Webseite oder einer Web Site beschreiben. So

hat die „Dublin Core Metadata Initiative“ (DCMI) ein sogenanntes *Coverage Element* für diesen Zweck vorgeschlagen (DCMI, 2000). Für dieses Element wurden unterschiedliche Kodierungsverfahren definiert: Punkt- und Rechteckskoordinaten können über die Elemente *DCMI Point* (Cox, 2000a) and *DCMI Box* (Cox, 2000b) beschrieben werden. Als Namen für geographischer Objekte können (für Staaten) *ISO 3166 Codes* oder Bezeichnungen verwendet werden, die über eine Datenbank wie dem *Getty Thesaurus of Geographic Names* festgelegt sind.

Es existieren weitere Vorschläge für die Definition von Geo-Tags. Beispiele hierfür sind der Internet Draft von A. Daviel (2001) und die geographischen Metadaten, die von dem US Federal Geographic Data Committee entwickelt worden sind.

Wie in der Einleitung erwähnt, ist es schwierig, standardisierte Meta-Informationen auf breiter Front durchsetzen und einen Missbrauch durch Anbieter zu verhindern, die sich Wettbewerbsvorteile durch irreführende Angaben verschaffen wollen. Daher müssen für eine Geo-Suchmaschine weitere Ansätze zur Bestimmung der geographischen Geltungsbereiche Berücksichtigung finden.

Top Level Domain .geo

Der Vorschlag von SRI International für die Top Level Domain (TLD) *.geo* kann auch als eine Art Geo-Tag aufgefasst werden. Ein solcher Domain-Name würde den geographischen Geltungsbereich über die Angabe einer Zelle beinhalten. Im Vorschlag wurden Zellen der Größen 10x10 Grad, 1x1 Grad und 1x1 Minute unterschieden. Für die dritte Variante könnte der Name einer Web Site *www.11e21n.3e7n.30e10n.geo* lauten. The TLD *.geo* wurde von ICANN im November 2000 nicht zugelassen.

Die resultierenden Domain-Namen der TLD .geo sind kryptisch und stellen damit die Idee für den Menschen leicht merkbarer Namen auf den Kopf. Auch ist die räumliche Auflösung selbst bei der kleinsten Zellengröße noch sehr grob.

Manuelle Registrierung

Viele Web-Portale erlauben die manuelle Registrierung von Web Sites nach der geographischen Lage des Angebots. Bekannte Beispiele sind private Homepages bei *www.web.de* bzw. Gelbe Seiten im Internet wie z.B. *www.gelbe-seiten.de*. Der Hauptnachteil dieser und ähnlicher Ansätze ist die Unvollständigkeit der Verzeichnisse; ein Benutzer kann sich nicht darauf verlassen, dass (mit hoher Wahrscheinlichkeit) alle relevanten Webseiten gelistet sind bzw. wirklich das nächstgelegene Angebot bestimmt worden ist. Auch können finanzielle Zuwendungen dazu führen, dass ein Angebot andere eventuell geeignetere Webseiten im Ranking überflügelt.

Neben den bekannten Web-Portalen existieren Prototypen für Geo-Suchmaschinen, die nicht das Web eigenständig durchsuchen, sondern auf einer Registrierung und Geokodierung durch die Anbieter beruhen. Beispiele hierfür sind *www.geourl.org* und *geotags.com*. Diese Suchmaschinen decken aber nur einen sehr kleinen Teil des Webs ab, so dass ein Benutzer nur äußerst rudimentäre Suchergebnisse erhält.

Auswertung IP-Adresse

Es gibt verschiedene Möglichkeiten, IP-Adressen auf geographische Positionen abzubilden. Eine verbreitete Anwendung ist die kartenbasierte Visualisierung von Trace

Routes mit Hilfe von Werkzeugen wie *GTrace* (Periakaruppan & Nemeth, 1999). Solche Werkzeuge zeigen den Netzwerkpfad von einem Client-Rechner zu einem spezifizierten Server-Rechner. Die Lagebestimmung der dabei bestimmten IP-Adressen erfolgt u.a. über WhoIs-Einträge und Namen der Gateways (Lakhina et al., 2002).

Für die Bestimmung des geographischen Geltungsbereichs von Webseiten ist es gefährlich, IP-Adressen als Basis zu verwenden, da die meisten Webseiten von Internet-Providern gehostet werden, deren Server sich an einem zentralen Ort befindet (bei der 1&1 Puretec GmbH zum Beispiel in Karlsruhe). Die geographische Lage des Servers hat damit im Regelfall nichts mit der Lage des Web-Anbieters oder dem geographischen Geltungsbereich der Web Site zu tun.

Auswertung DNS

Über den WhoIs-Service können für einen Domain-Namen eine Reihe von Informationen abgerufen und für die Bestimmung des geographischen Geltungsbereichs verwendet werden. Dabei wird zwischen dem Domain-Inhaber (registrant), dem administrativen Ansprechpartner (admin-c), dem technischen Ansprechpartner (tech-t) und dem Zonenverwalter (zone-c) unterschieden. Nach Markowetz et al. (2003) hat sich insbesondere die Auswertung von den Einträgen für den administrativen Ansprechpartner bewährt; für diesen liegt eine Anschrift mit Postleitzahl, Orts- und Straßenangabe vor.

Semantische Analyse der Seiteninhalte und der URLs

Ein weiterer Ansatz ist, den Inhalt von Webseiten syntaktisch und semantisch zu analysieren. Bei der semantischen Analyse wird der Inhalt nach geographischen Bezeichnungen wie Städtenamen oder nach Postleitzahlen oder Vorwahlnummern untersucht (McCurley, 2001) (Daniel, 2002).

Allerdings ist eine solche Analyse recht aufwändig. Auch ist die Sicherheit des Analyseresultats beschränkt. "Oldenburg" kann eine Stadt in Niedersachsen oder Schleswig-Holstein sein oder in Illinois, Indiana, Mississippi oder Texas liegen oder ein Nachname oder Firmenname sein. Nationale Unterschiede erschweren die Analyse von Postleitzahlen und Vorwahlnummern. Auch gibt es Vorwahlnummern ohne geographische Relevanz.

Ein ähnlicher Ansatz ist, die URL einer Web Site zu analysieren. Dieser Ansatz führt oft zu guten Ergebnissen. So funktioniert dieser Ansatz für die Universität Oldenburg (*www.uni-oldenburg.de*). Für die FH Oldenburg/Ostfriesland/Wilhelmshaven (*www.fh-oow.de*) gibt es auch eine Beziehung zwischen der URL und der Lage; diese Beziehung kann aber nicht automatisch extrahiert werden. Auch gibt es irreführende Fälle: So lebt die Schriftstellerin Birgit Oldenburg (*www.birgit-oldenburg.de*) in Minden und nicht in Oldenburg.

ARCHITEKTUR EINER GEO-SUCHMASCHINE

Da keines der genannten Verfahren allein die Bestimmung des geographischen Geltungsbereichs einer Webseite mit hinreichender Sicherheit erlaubt bzw. für alle Seiten ein Ergebnis liefern kann, müssen mehrere Verfahren in geeigneter Weise miteinander kombiniert werden. Dies muss in der *Architektur* einer effektiven Geo-Suchmaschine

berücksichtigt werden. Daneben sind allerdings weitere Bausteine in die Architektur einer Geo-Suchmaschine aufzunehmen:

- Neben dem geographischen Geltungsbereich spielt zur Beurteilung einer Webseite weiterhin die thematische Treffergenauigkeit eine hervorstechende Rolle. Daher müssen diese beiden Aspekte von der Geo-Suchmaschine miteinander verknüpft werden. Dazu sind geeignete Benutzerschnittstellen und Suchalgorithmen einzusetzen, die einerseits eine flexible, im Verlauf der Ergebnisverarbeitung variierbare Gewichtung beider Aspekte erlauben und anderseits keinen großen algorithmischen Aufwand bei einer Veränderung der Gewichtung erfordern (Markowetz et al., 2003).
- Bei der Gewichtung des geographischen Geltungsbereichs muss die lokale Bedeutung einer Webseite von ihrer globalen Bedeutung abgegrenzt werden. Für die Beispielsanfrage zu Beginn war ausschließlich die lokale Bedeutung der gesuchten Webseite(n) von Relevanz.

DER PAGE2LOC-DIENST

In diesem Beitrag soll (nur) auf die Kombination der verschiedenen Verfahren zur Bestimmung von geographischen Geltungsbereichen als Teil einer Geo-Suchmaschine näher eingegangen werden. Dabei wird angestrebt, diese Verfahren in einem sogenannten *„Page2Loc"-Dienst* zu bündeln. Dieser Geodienst soll automatisiert zu einer gegebenen Webseite (einschließlich ihrer URL) einen oder ggf. mehrere geographische Geltungsbereiche bestimmen und – wenn möglich – mit einer Geltungswahrscheinlichkeit versehen.

Auswahl der Verfahren

Zunächst sollen die Verfahren identifiziert werden, die für einen „Page2Loc"-Dienst in Frage kommen. Tabelle 1 stellt die präsentierten Verfahren zur Bestimmung des geographischen Geltungsbereichs von Webseiten gegenüber.

Tab. 1: Vergleich der Verfahren zur Bestimmung des geographischen Geltungsbereichs von Webseiten.

	Verfügbarkeit	Einfachheit	Sicherheit	Genauigkeit
Geo-Tags	-	+	±	+
TLD .geo	--	+	+	-
Manuelle Registrierung	-	+	± / +	+
Auswertung IP-Adresse	+	±	--	+
Auswertung DNS	+	±	+	+
Semantische Analyse	±	-	±	+

Unter „Verfügbarkeit" ist bewertet, inwieweit ein Verfahren verfügbar bzw. zu welchem Grad für gängige Webseiten angewendet werden kann. Die fehlende Verfügbarkeit der Top Level Domain .geo disqualifiziert diese Vorgehensweise. „Einfachheit" bewertet den Aufwand, der notwendig ist, um ein Verfahren automatisiert in einem

Geodienst anwenden zu können. Hier ist die semantische Analyse von URLs und von Seiteninhalten sicherlich das aufwändigste Verfahren. Die manuelle Registrierung wurde in dieser Rubrik zwar mit einem Plus bewertet; allerdings ist dabei zu berücksichtigen, dass damit üblicherweise die Suche von Webseiten zu einem örtlichen Kriterium unterstützt wird. Für den „Page2Loc“-Dienst wird aber genau die umgekehrte Auswertungsrichtung benötigt, die sich allerdings relativ einfach implementieren ließe. Die Kategorie „Sicherheit“ gibt an, inwieweit ein Ergebnis tatsächlich Rückschlüsse auf den geographischen Geltungsbereich einer Webseite erlaubt. Die Auswertung der IP-Adresse wird aufgrund dieses Kriteriums disqualifiziert. Bei der manuellen Registrierung stehen zwei Bewertungen, die davon abhängen, ob die Registrierungsangaben von dritter Seite überprüft werden. Unter dem in der letzten Spalte aufgeführten Aspekt „Genauigkeit“ ist die erzielbare Genauigkeit für geographische Geltungsbereiche gemeint. Hier sind (mit Ausnahme von der TLD .geo) keine signifikanten Unterschiede zu erwarten.

Damit verbleiben vier Verfahren: Geo-Tags, die manuelle Registrierung von Webseiten, die Bestimmung der Informationen über den administrativen Ansprechpartner bezüglich eines Domain-Namens und die semantische Analyse von URLs und von Seiteninhalten in Hinsicht auf ortsbezogene Informationen.

Verarbeitung

Ein zunächst naheliegender Ansatz ist es, die vier genannten Verfahren parallel anzuwenden und dann die Resultate zu kombinieren. Auf einen solche Vorgehensweise sollte aber in Hinsicht auf die Effizienz möglichst verzichtet werden. Stattdessen ist ein Kompromiss zwischen Verarbeitungsgeschwindigkeit und Genauigkeit bzw. Wahrscheinlichkeit des Resultats gesucht.

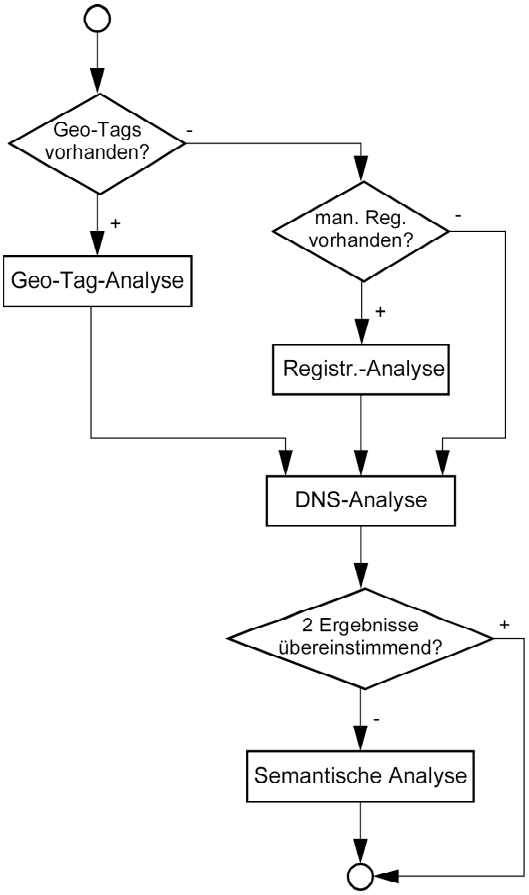

Abb. 1: Ablaufdiagramm zur Kombination von Verfahren für die Bestimmung von geographischen Geltungsbereichen.

Die beiden einfachsten der vier Verfahren (vgl. Tabelle 1) sind die Geo-Tags und die manuelle Registrierung. Sind Geo-Tags vorhanden oder eine manuelle Registrierung verfügbar, sollten sie daher ausgewertet werden. Allerdings beruhen beide Verfahren auf einem ähnlichen Ansatz – nämlich der expliziten Referenzierung der Lage durch den Web-Anbieter. Daher ist die Bestätigung von Geo-Tags durch die Angaben der manuellen Registrierung nicht angebracht. Stattdessen sollte möglichst eine Bestätigung über die Auswertung des administrativen Ansprechpartners erfolgen. Stimmen die Angaben überein, kann der Prozess enden. Wird hier keine Übereinstimmung erzielt oder fehlen sowohl Geo-Tags als auch manuelle Registrierungseinträge, kann eine semantische Analyse in einem letzten Schritt eingesetzt werden; aufgrund des hohen Aufwands sollte dies aber nicht vorher erfolgen. Das Flussdiagramm in Abbildung 1 illustriert den geschilderten Ablauf.

Vereinheitlichung der geographischen Geltungsbereiche

Das Ergebnis des „Page2Loc“-Dienstes sind ein oder mehrere geographische Geltungsbereiche. Um die Resultate dieses Geodienstes in einer Geo-Suchmaschine oder von anderen Anwendungen weiterverarbeiten zu können, ist eine möglichst exakt definierte Spezifikation der Beschreibung der Geltungsbereiche erforderlich. Auch innerhalb des „Page2Loc“-Dienstes tritt diese Anforderung auf: Die Verwendung mehrerer Verfahren zur Bestimmung von geographischen Geltungsbereichen führt dazu, dass die so be-

stimmten Geltungsbereiche unterschiedlich beschrieben sind: Der im Abschnitt „Geo-Tags“ vorgestellte Vorschlag der Dublin Core Metadata Initiative erlaubt bereits unterschiedliche Beschreibungsformen, der Ortbezug des administrativen Ansprechpartners eines Domain-Namens erfolgt über Postleitzahl, Orts- und Straßenangabe und die Ergebnisse der semantischen Analyse decken ebenfalls einen sehr weiten Bereich von Beschreibungen ab.

Um diese Vielfalt einzuschränken, ist die Abbildung von geographischen Angaben in ein einheitliches Modell zur Beschreibung der geographischen Geltungsbereiche erforderlich. Dazu bietet sich die Verwendung von Punkt- oder Flächengeometrien an. Dienstintern kann dies über Geometrien erfolgen, die über das Simple-Feature-Modell des OpenGIS-Konsortiums repräsentiert werden (Open GIS Consortium, 1999). Für die Weitergabe an einen anderen Dienst oder eine Applikation ist insbesondere die Geography Markup Language (GML) (Open GIS Consortium, 2003) eine geeignete Basis zur Beschreibung von geographischen Geltungsbereichen.

Um die verschiedenen Ortsbeschreibungen in Punkt- oder Flächengeometrien abzubilden, ist der Einsatz unterschiedlicher Geodienste angezeigt. Bei der Auswahl und dem Zusammenspiel sind u.a. die folgenden Gesichtspunkte zu berücksichtigen:

- die Mehrdeutigkeit von geographischen Bezeichnungen,
- die Existenz von unterschiedlichen Bezeichnungen und/oder Abkürzungen für geographische Merkmale,
- die Abbildung von geographischen Bezeichnungen, Postleitzahlen, Vorwahlnummern usw. in eventuell nicht übereinstimmende geometrische Gebiete,
- die Bestimmung von Entfernungen zwischen geographischen Geltungsbereichen.

Damit werden folgende Geodienste benötigt:

- *Name2Name:* Dieser Dienst dient dazu, um unterschiedliche Bezeichnungen und/oder Abkürzungen für geographische Merkmale auf ein einheitliches Bezeichnungssystem abzubilden. Aufgrund der Mehrdeutigkeit von geographischen Bezeichnungen kann auch eine Menge von Bezeichnungen das Ergebnis dieses Dienstes sein.
- *Name2Loc:* Die Abbildung von (standardisierten) Ortsbezeichnungen in eine geometrische Beschreibung der Lage und Ausdehnung. Aufgrund der Mehrdeutigkeit von geographischen Bezeichnungen kann auch eine Menge von Geometrien das Ergebnis dieses Dienstes sein. Ein Beispiel für einen solche Dienst ist der ArcWeb Place Finder Web Service von ESRI.
- *PostCode2Loc:* Im Rahmen der Auswertung des administrativen Ansprechpartners für einen Domain-Namen und der semantischen Inhaltsanalyse von Webseiten treten Postleitzahlen auf, die in eine geometrische Beschreibung der Lage und Ausdehnung überführt werden müssen. Als Beispiel für einen entsprechenden Web Service kann der Zip2Loc Web Service von CDYNE Systems genannt werden.
- *DialingCode2Loc:* Analog zum „PostCode2Loc“-Dienst ist auch eine Umsetzung von Vorwahlnummern in eine geometrische Beschreibung erforderlich. Der Web Service DOTS GeoPhone bestimmt für nordamerikanische Telefonnummern den

Zip Code und die Adresse, auf deren Basis wiederum Koordinaten abgeleitet werden können.

Bei der DNS-Analyse und der semantischen Inhaltsanalyse können sehr genaue Lageinformationen (insbesondere Straßen- und Hausnummerangaben) auftreten, die auch in Rahmen der Analyse von geographischen Geltungsbereichen betrachtet werden können. Ein solcher Schritt soll aber an dieser Stelle nicht weiter erörtert werden, da für die meisten Anwendungsfälle eine Auflösung auf Orts-, Postleitzahl- oder Vorwahlbereichsebene als hinreichend erscheint.

Abbildung 2 skizziert die Nutzung der genannten Geodienste innerhalb des „Page2Loc"-Dienstes.

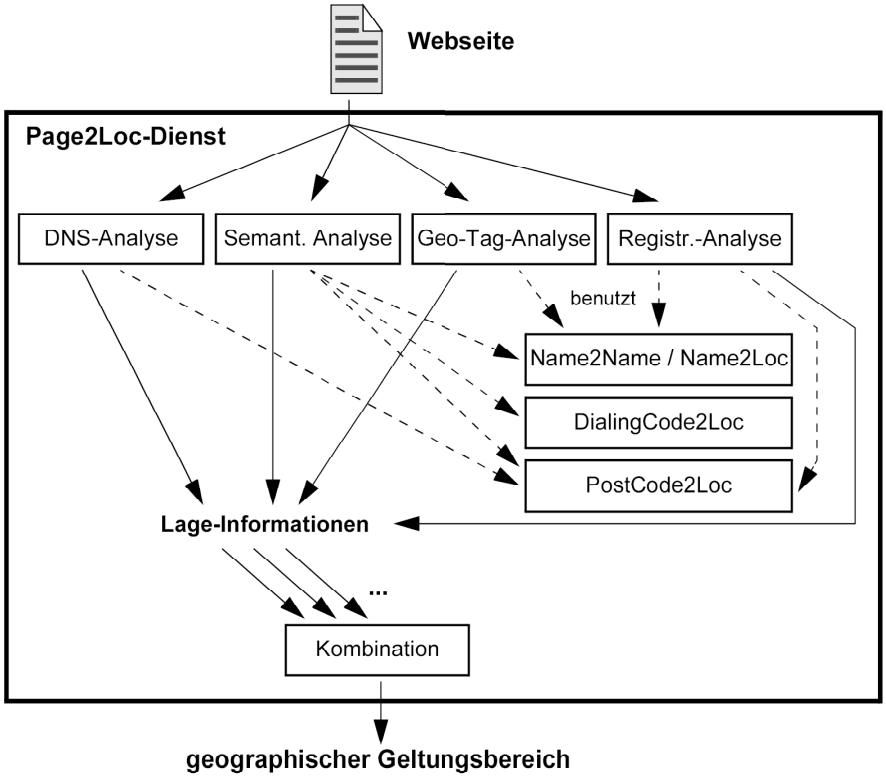

Abb. 2: Nutzung von Geodiensten zur Bestimmung von geographischen Geltungsbereichen.

Link-Analyse

Ein bislang nicht erwähnter Ansatz, um den geographischen Geltungsbereich einer Webseite zu bestimmen bzw. um diesen zu gewichten, ist die Betrachtung der (bekannten oder noch zu bestimmenden) geographischen Geltungsbereiche von Webseiten, die auf die aktuell betrachtete Seite über einen Hyperlink verweisen (Ding et al., 2000). Um auf dieser Basis ein verlässliches Ergebnis zu erhalten, sind zwei Bedingungen zu erfüllen:

- eine gewisse Mindestanzahl von Links aus dem (vermuteten) Geltungsbereich der Webseite sollten auf diese Seite verweisen und

- diese Links sollten über den (vermuteten) Geltungsbereich möglichst gleichverteilt sein.

Um den Ansatz der *Link-Analyse* in den „Page2Loc“-Dienst aufnehmen zu können, muss zusätzlich eine Datenbank für die bislang bekannten Geltungsbereiche eingeführt werden. Die Link-Analyse ruft ihrerseits (u.a.) den „Page2Loc“-Dienst auf. Abbildung 3 skizziert diese Erweiterung.

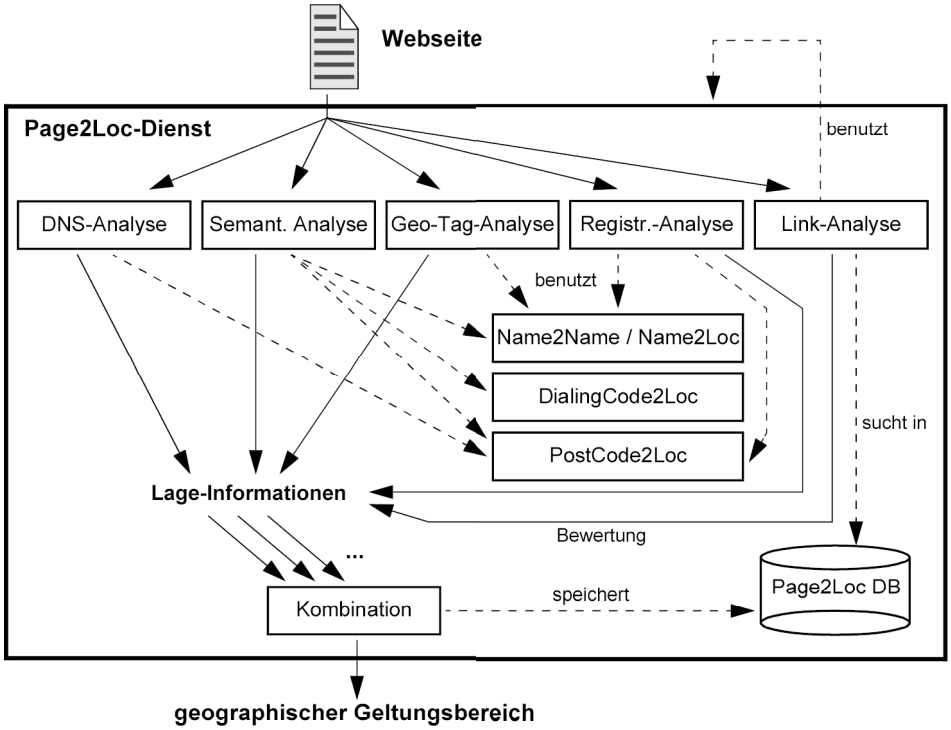

Abb. 3: Integration der Link-Analyse in den „Page2Loc“-Dienst.

SCHLUSSBEMERKUNGEN

In diesem Beitrag wurde der „Page2Loc“-Dienst vorgestellt, der in Rahmen der Architektur einer Geo-Suchmaschine die Aufgabe hat, für eine Webseite den geographischen Geltungsbereich zu bestimmen. In diesen Geodienst werden vier Verfahren integriert:

- die Analyse von Geo-Tags,
- die Auswertung von Angaben einer manuellen Registrierung der Webseite,
- die Auswertung der Angaben von Informationen über den administrativen Ansprechpartner für einen Domain-Namen und
- die semantische Analyse von URLs und von Seiteninhalten in Hinsicht auf ortsrelevante Informationen.

Für die Implementierung des „Page2Loc“-Dienstes werden wiederum andere Geodienste benötigt, um Ortsbezeichnungen, Postleitzahl- und Vorwahlangaben zu vereinheitlichen und insbesondere auf eine geometrische Beschreibung der Lage und Ausdehnung

abzubilden. Zusätzlich wurde eine Link-Analyse in den vorgestellten Dienst integriert, um Geltungsbereiche gewichten zu können. Die Link-Analyse nutzt ihrerseits den Geodienst „Page2Loc“.

Bislang sind einzelne Bausteine, die im Rahmen der Architektur einer Geo-Suchmaschine benötigt werden, separat entworfen und untersucht worden. Diese Bausteine müssen weiter in der vorgestellten Zielrichtung erweitert und zusammen mit noch fehlenden Bausteinen in die Gesamtarchitektur integriert werden. Dabei ist allerdings auch die Verfügbarkeit, Qualität und Effizienz der genutzten Web- und Geodienste von entscheidender Bedeutung, um eine effektive und effiziente Geo-Suchmaschine entwickeln zu können.

LITERATUR

Buyukkokten, O., Cho, J., Garcia-Molina, H., Gravano, L., Shivakumar, N (1999): *Exploiting Geographical Location Information of Web Pages*. Proceedings WebDB 1999.

Cox S. (2000a): *DCMI Point Encoding Scheme.* Recommendation of the Dublin Core Metadata Initiative, July 2000, http://dublincore.org/ documents/dcmi-point/

Cox S. (2000b): *DCMI Box Encoding Scheme*. Recommendation of the Dublin Core Metadata Initiative, July 2000, http://dublincore.org/ documents/dcmi-box/

Daniel E. (2002): *Geographic Search*. Winner of the Google Programming Contest, http://www.google.com/programming-contest/winner.html.

Daviel A.: *Geographic registration of HTML documents*. Internet Draft, April 2001, http://geotags.com/geo/draft-daviel-html-geo-tag-05.html

Ding, J., Gravano L., Shivakumar, N. (2000): *Computing Geographical Scopes of Web Resources*. Proceedings 26th International Conference on Very Large Databases, 445-456.

Dublin Core Metadata Initiative (2000): *Dublin Core Qualifiers, Recommendation*. http://dublincore.org/documents/dcmes-qualifiers/

International Organization for Standardization (ISO) und Deutsches Institut für Normung (DIN): *ISO 3166-1: The Code List*. http://www.din.de/ gremien/nas/nabd/iso3166ma/codlstp1/index.html

J. Paul Getty Trust: *Getty Thesaurus of Geographic Names On Line*. http://www.getty.edu/research/tools/vocabulary/tgn/index.html

Lakhina A., Byers J., Crovella M., Matta I. (2002): *On the Geographic Location of Internet Resources*. Technical Report, Boston University, May 2002.

Markowetz, A., Brinkhoff, Th., Seeger B. (2003): *Spatial-Aware Browsing the Internet*. Eingereicht zur Veröffentlichung.

McCurley K.S. (2001): *Geospatial Mapping and Navigation of the Web*. Proceedings10th International World Wide Web Conference.

Open GIS Consortium Inc. (1999): *OpenGIS Simple Feature Specification for SQL*, Revision 1.1, May 1999, http://www.opengis.org/techno/ implementation.htm

Open GIS Consortium Inc. (2003): *OpenGIS Geography Markup Language (GML) Implementation Specification*, Version 3.0, January 2003, http://www.opengis.org/techno/implementation.htm

Periakaruppan R., Nemeth E (1999): *GTrace - A Graphical Traceroute Tool*. 13th Systems Administration Conference - LISA '99.

SRI International: *The Proposed .geo Top-Level Domain Name*, http://www.dotgeo.net/dotgeo/