Standortdarstellung eines Webseitenbetreibers auf Basis von Open-Source-Komponenten & Google Maps

Frank Rump, Thomas Grensemann

Fachbereich Technik
FH Oldenburg/Ostfriesland/Wilhelmshaven
Constantiaplatz 4
26723 Emden
rump@informatik-emden.de
thomas_grensemann@web.de

Abstract: In diesem Artikel wird die Entwicklung eines Informationssystems auf Basis von Open-Source-Komponenten vorgestellt, das die Darstellung des exakten Standortes eines Webseitenbetreibers einer aufgerufenen Webseite automatisiert. Aktuelle, im Web gängige Systeme gestatten die genaue Darstellung eines Ortes erst nach Eingabe der zugehörigen Adresse oder verwenden zur Bestimmung von Standortdaten GeoIP-Länderdatenbanken und ermöglichen so höchstens eine Genauigkeit in einem Radius von 20 km bis 80 km. Der in dieser Arbeit verwendete Ansatz implementiert eine ortsbasierte Suchmaschine, die Webseiten nach Ortsangaben durchsucht, diese den Webseiten in einer Datenbank zuordnet und zur Darstellung aufbereitet. Die Darstellung selbst erfolgt mithilfe eines Firefox-Add-on und einer zugehörigen Webanwendung.

1 Ausgangssituation

Google Maps, ein von der Firma Google im Februar 2005 gestarteter Internetdienst ermöglicht die weltweite Suche von Standorten oder anderen Objekten (wie Hotels etc.) mittels einer zugehörigen Adresse und erlaubt die Darstellung wahlweise als reine Kartendarstellung, als Satellitenbild oder als Mischform. Eine Zoomfunktion und Navigationselemente, die eine Detailansicht von Objekten sowie die Fortbewegung auf Karten- oder Bildausschnitten erlauben, runden den Dienst ab [STE08]. Mit der bereitgestellten Programmierschnittstelle (Google Maps API) ist es möglich Applikationen zu implementieren, die auf die Karten zugreifen. Diese Applikationen lassen sich direkt in Webseiten integrieren und per Browser ausführen [PLA06].Eine beispielhafte Anwendung, die auf die Technologie von Google Maps zurückgreift, ist z. B. die Firefox-Erweiterung Shazou der amerikanischen Firma Seisan. Die Erweiterung wertet die in der Adressleiste des Browsers stehende URL aus und zeigt mithilfe von Google Maps den Standort des Servers an, auf dem die dargestellte Internetseite gehostet wird. Weil der angezeigte Server-Standort aber in den seltensten Fällen dem Standort des Internetseitenbetreibers entspricht, geben die von Shazou ermittelten Daten keine Auskunft über den wahren Standort des Betreibers wieder. Hier bietet sich in Verbindung mit der 1997

eingeführten Impressumpflicht für deutsche Internetseiten die interessante Möglichkeit, eine Erweiterung für den Firefox-Browser nach dem Vorbild von Shazou zu implementieren, die den genauen Standort des Betreibers einer angezeigte Internetseite wiedergibt. Zur Umsetzung dieser Funktionalität müssen im Vorfeld die Adressinformationen aus dem Impressum oder den Kontaktseiten der Internetseiten ausgewertet sowie die zur Darstellung des Standortes benötigten Geodaten von Google geladen werden. Um die verhältnismäßig zeitaufwendigen Vorgänge nicht erst beim Aufruf des Firefox-Add-ons auszuführen, je nach Gegebenheit ergeben sich bei der Analyse eines Internetauftritts (Suche und Auswertung der Adressdaten) sowie dem Laden der Geodaten Laufzeiten von 10 sec bis 40 sec, werden die Adressinformationen und die Geodaten mithilfe einer Webanwendung aus einer relationalen Datenbank anhand der URL der angezeigten Internetseite bezogen. Zum vorherigen Sammeln und Laden dieser Informationen bietet sich die Open-Source-Suchmaschine Nutch an. Die Suchmaschine erlaubt es mit ihrem Plug-in-System eine Erweiterung zu implementieren, die den erforderlichen Funktionalitäten entspricht (ortsbasierte Suchmaschine). Ferner ermöglicht die Suchmaschine die Installation auf einem Rechner-Cluster, so dass sich das Sammeln der benötigten Daten zusätzlich beschleunigen lässt.

Das Einsatzgebiet des entwickelten Systems ist vor allem im Bereich der automatisierten Routenplanung zu sehen. Denkbar wäre es auch, mithilfe der ortsbasierten Suchmaschine Unternehmensverzeichnisse zu generieren oder Navigationssysteme mit Inhalten zu füllen.

2 Nutch

Die zentrale Komponente des Systems bildet die Open-Source-Suchmaschine Nutch. Nutch ist eine auf Java basierende Anwendung und verwendet zur Indizierung und zum Durchsuchen der Internetseiten die Java-Bibliothek Lucence. Nach eigenen Angaben kann Nutch mehrere Millionen Internetseiten pro Monat sammeln und indizieren und gestattet das Durchsuchen der indizierten Webseiten bis zu 1000-mal pro Sekunde [NUT07]. Zum Betreiben der Suchmaschine werden lediglich eine aktuelle Java-Version und ein Tomcat-Server benötigt. Der Server ist ausschließlich zur Bereitstellung der Weboberfläche erforderlich. Die enorme Geschwindigkeit der Suchmaschine wird durch Prozessaufteilung erreicht. Die Suchmaschine separiert zum einen die komplette Weboberfläche vom Sammeln (Crawlen) und Indizieren der Webseiten, zum anderen werden auch die Prozesse zum Crawlen und Indizieren getrennt voneinander ausgeführt. Letzteres hat unter anderem den Vorteil, dass das Crawlen nicht durch das Indizieren komplexer Internetseiten ausgebremst wird.

Die Suchmaschine verfügt über ein Plug-in-System, mit dessen Hilfe es möglich ist die Software an verschiedenen Stellen um Funktionalitäten zu erweitern. Beispielsweise werden die Verarbeitung diverser Dokumententypen sowie die Filterung der Internet-Adressen und Inhalte durch Plug-ins umgesetzt. Die Plug-ins werden als kleine, abgeschlossene Module programmiert und mithilfe einer XML-Datei installiert ohne am Kernsystem selbst etwas zu ändern [NEB09].

Basis der ortsbasierten Suchmaschine bildet ein Plug-in, das die Domain einer URL vor dem Download durch den Nutch-Crawler nach Ortsangaben aus einem bestimmten, durch die Konfiguration des Plug-in vorgegebenen Postleitzahlbereich durchsucht. Hierzu wird die Domain geladen und im Anschluss nach einem Impressum oder einer Kontaktseite durchsucht. Enthält die Domain eine entsprechende Seite, wird diese nach einer Postleitzahl aus dem vorgegebenen Postleitzahlbereich durchsucht. In Fall einer Übereinstimmung werden dann die Adressdaten ausgelesen, die zur Darstellung des Standortes benötigten Geodaten von Google Maps geladen und die Standortinformationen in einer Datenbank gespeichert.

Die Einschränkung auf einen vorgegebenen Postleitzahlbereich erfolgte während der Entwicklungsphase zu Testzwecken, lässt sich aber bei entsprechender Konfiguration des Plug-ins auf ganz Deutschland ausdehnen.

Abb.1 illustriert die Arbeitsweise der ortsbasierten Suchmaschine.

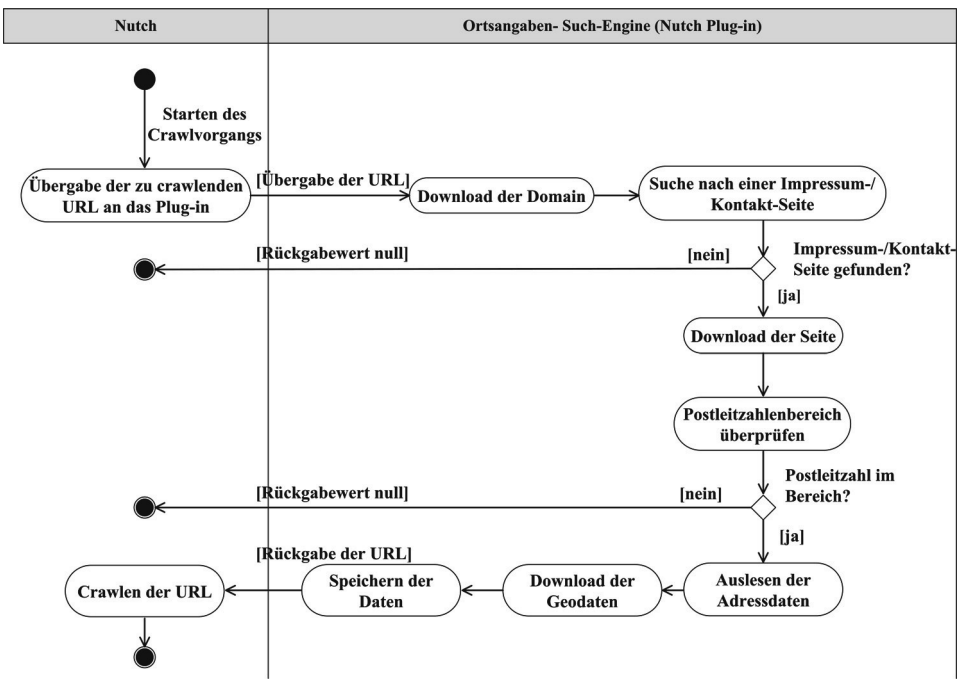

Abbildung 1: Arbeitsweise der Suchmaschine

3 System-Architektur

Die Hauptkomponente des Systems bildet ein Rechner-Cluster auf dem die ortsbasierte Suchmaschine installiert wird. Zusätzlich beinhaltet das System ein Firefox-Plug-in sowie eine zugehörige Webanwendung. Die Datenhaltung wird mithilfe des Datenbankmanagementsystems MySQL realisiert.

Abb. 2 veranschaulicht die Komponenten und ihre Einordnung in das Gesamtsystem.

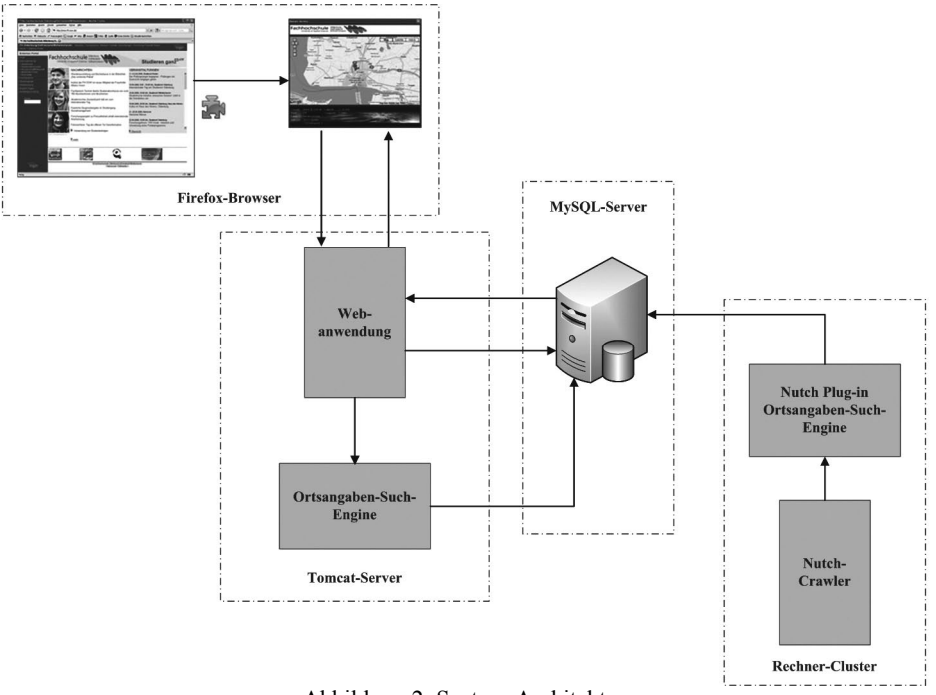

Abbildung 2: System-Architektur

3.1 Firefox-Plug-in

Das Plug-in ermöglicht die Darstellung des Betreibers einer aufgerufenen Internetseite. Hierzu wird die in der Adressleiste des Browser stehende URL ausgelesen und an die Webanwendung übergeben. Im Anschluss wird die durch die Webanwendung generierte Karte in einem gesonderten Browserfenster dargestellt.

3.2 Webanwendung

Die Webanwendung wurde unter Verwendung des Open-Source-Framework Struts implementiert [BMWH05]. Aufgabe der Anwendung ist es, die durch das Firefox-Plug-in gesendete URL entgegenzunehmen und mithilfe der Domain die zur Kartendarstellung erforderlichen Daten aus der Datenbank zu ermitteln. . Im Anschluss wird eine JSP-Seite generiert und an das Plug-In zurückgesendet. Die Seite enthält u.a. eine mit der Google-Maps-Tag-Library generierte Karte. Die Bibliothek ist ein Open-Source-Projekt der US amerikanischen Firma Lamatek und gestattet die Verwendung der Google-Maps-API-Funktionalitäten in JSP-Seiten ohne Verwendung des API-JavaScript-Codes. Das Einbinden der Bibliothek in eine JSP-Seite sowie das Arbeiten mit der Bibliothek erfolgt ohne nennenswerte Probleme äquivalent zu anderen bekannten Bibliotheken.

Für den Fall, dass es trotz der Sammlung von Adress- und Geodaten in der Datenbank dazu kommen sollte, dass für eine angefragte Domain keine Daten vorhanden sind, wird die Ortsangaben-Such-Engine verwendet, um für die betreffende URL die benötigten Daten zu ermitteln.

3.3 Datenbank

Die Speicherung der gesammelten Daten erfolgt mithilfe des Open-Source-Datenbankmanagementsystems MySQL. Gespeichert werden zum einen die Standortinformationen zu den jeweiligen Domains, zum anderen werden die Daten der Domains gespeichert für die keine Ortsangaben gefunden werden. Die Speicherung der Domains ohne Ortsangaben erfolgt aus Gründen der Performanz.

3.4 Rechner-Cluster

Wie kommerzielle Suchmaschinen ermöglicht auch Nutch die Installation auf einem Rechner-Cluster. Die Umsetzung erfolgt mithilfe des Frameworks Hadoop. Hadoop ist wie Nutch ein Open-Source-Projekt der Apache Software Foundation und setzt das von Google entwickelte Framework MapReduce in Java um [DG04]. Zusätzlich implementiert es das Dateisystem „Hadoop Distributed File System“. MapReduce gestattet die Gliederung der Daten in kleine Arbeitsblöcke (Segmente), die dann mithilfe des Dateisystems auf die einzelnen Cluster-Rechner zur Verarbeitung verteilt werden. Anschließend lassen sich die Daten dann wieder mit Hilfe von MapReduce zusammenführen. Hierfür implementiert MapReduce eine Map-Funktion, die die Daten einer Schlüssel-Wert-Liste zuordnet und in Segmente aufteilt, und eine Reduce-Funktion, die die Ergebnisse der einzelnen Cluster-Rechner für den gleichen Schlüssel wieder zusammenführt [HAD08].

Die für das Clustering notwendigen Hadoop-Dateien sind in Nutch integriert, so dass für eine Cluster-Installation Hadoop nicht zusätzlich installiert werden muss.

4 Zusammenfassung

Im Rahmen dieser Arbeit wurde ein Informationssystem entworfen, auf dessen Basis der Standort eines Webseitenbetreibers einer aufgerufenen Webseite automatisch dargestellt werden kann. Grundlage des Systems bildet eine ortsbasierte Suchmaschine, die mithilfe der Open-Source-Suchmaschine Nutch implementiert wurde, sowie ein Firefox-Add-on, das die Darstellung des Standortes realisiert.

Die an das System gestellten Anforderungen konnten im Wesentlichen unter Einsatz der Open-Source-Komponenten erfüllt werden. Problematisch gestaltete sich vor allem die Auswertung der Ortsangaben aufgrund der verschiedenen Darstellungsvarianten in den Internetpräsenzen. Optimierungsmöglichkeiten ergeben sich deshalb im Wesentlichen bei der Adressenauswertung sowie der Performanz der Ortsangaben-Such-Engine.

5 Verwendete Komponenten

Die nachfolgende Tabelle gibt einen Überblick über die verwendeten Open-Source-Komponenten:

System	Beschreibung	URL
Nutch	Suchmaschine	http://lucene.apache.org/nutch/
Apache Tomcat	Servlet-Container	http://tomcat.apache.org/
Apache	Http-Server	http://httpd.apache.org/
MySQL	MySQL-Datenbank	http://www.mysql.com/
phpMyAdmin	Applikation zur Verwaltung von MySQL Datenbanken	http://www.phpmyadmin.net/
Eclipse-WTP	Eclipse für JSP/Web/XML	http://www.eclipse.org/webtool/
XULBooster	Eclipse Plug-In für XUL	http://cms.xulbooster.org/
Firefox	Web-Browser	http://www.mozilla.com/
Dom-Inspector	Firefox Erweiterung	
Struts	Framework zur Entwicklung von Java Web-Applikationen	http://struts.apache.org/
Ant	Build-Werkzeug	http://ant.apache.org/
Google Maps JSP Tag Library	Java Frontend für die Google Maps API	http://www.lamatek.com/GoogleMaps/
dom4j	API für die Verarbeitung von XML-Dokumenten	http://www.dom4j.org/
jaxen	Java XPath Engine	http://jaxen.codehaus.org/

Tabelle 1: Verwendete Open-Source-Komponenten

Literaturverzeichnis

[BMWH05] Adam Bien, Marcel May, Bernhard Wöhrlin und Sven Haiges: Struts. Java Framework für Webanwendungen. Software & Support Verlag, 2005.

[DG04] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. In: OSDI'04: Sixth Symposium on Operating System Dsign and Implementation, San Francisco, CA, December, 2004.

[HAD08] Hadoop, http://hadoop.apache.org/core/, März 2008.

[NEB09] Michael Nebel: Freie Suchsoftware: Aspseek und Nutch. http://www.heise.de/ix/artikel/2005/09/100/, September 2005.

[NUT07] Nutch, http://lucene.apache.org/nutch/docs/de/, April 2007.

[PLA06] Alberto Planas: Aus großer Distanz, http://www.linux-magazin.de/, Januar 2006.

[STE08] Ralf Steyer: Google Web API. M.P. Media-Print Informationstechnologie GmbH, Paborn, 2007.