CARL VON OSSIETZKY universität OLDENBURG

Fakultät II – Informatik, Wirtschafts- und Rechtswissenschaften
Department für Informatik

Geographically Focused Web Information Retrieval

Dissertation zur Erlangung des Grades eines
Doktors der Naturwissenschaften

vorgelegt von

Dipl.-Inform. Dirk Ahlers

Gutachter:

Prof. Dr. Susanne Boll, Universität Oldenburg
Prof. Dr. Andreas Henrich, Universität Bamberg

Tag der Disputation: 13. Dezember 2010

Abstract

Search engines are the preferred means to retrieve information from the World Wide Web. Geographic Information Retrieval extends the concept of search to enable specific access to location-based information. While modern search engines already achieve satisfying results for a wide range of queries, the keyword-based approach is not well equipped to handle the more complex search for location-based information. Yet, even within common search engine, up to 20% of queries have a geospatial character. On the other hand, the Web itself already is a large source of rich location-based information. The focused processing of this geospatial Web content constitutes a huge opportunity for search. Hence, the goal of this thesis is the investigation into efficient methods for the creation of a focused geospatial search engine. Addressing the main topics of crawling, parsing, and ranking, it is able to exploit location references within common, unstructured Web content. This work uses methods and approaches from the field of Geographic Information Retrieval, adapts them, and develops them further. Contrary to keyword-based search, the work considers location references in their semantic sense. Location references are extracted from common Web documents. Their identification and subsequent geographical mapping allows the use of geospatial information as an additional dimension of Web search and facilitates a geographical view of the Web. The major contribution of this thesis is a geospatial focus in two aspects. First, efficient strategies are developed to guide a Web crawler specifically towards geospatially relevant Web pages. For this resource discovery process, the approach of focused crawling is adapted to geospatial information. This allows the crawler to restrict itself to pages containing references from a specific geographic region. The necessary crawl strategy is based on the classification and prediction of location-relevant pages on the Web. Second, the thesis develops a geoparser that can extract geographical information at the high granularity of addresses. It allows for a verified extraction of location references at a building-level. By incorporating external domain knowledge, it achieves a high quality or georeferenced results. This work is completed by evaluations of the developed approaches, the proposal of a new ranking approach, and a workable architecture suitable for the development of multiple prototypical applications.

Zusammenfassung

Suchmaschinen sind heute das meistgenutzte Mittel, um Zugang zu Informationen im World Wide Web zu gewinnen. Geografisches Information Retrieval erweitert das Konzept der Suche, um auch den Zugang zu ortsbasierten Informationen zu ermöglichen. Während moderne Suchmaschinen bereits für eine Vielzahl von Anfragen befriedigende Ergebnisse liefern, so gilt dies nur bedingt für ortsbasierte Informationen, da diese sich oftmals nur schwer über eine Stichwortsuche auffinden lassen. Dennoch beinhalten bereits bis zu 20% aller Anfragen an Suchmaschinen einen Ortsbezug. Im Gegenzug beinhaltet das Web ebenfalls eine Vielzahl von hochwertigen Informationen, die ebenfalls einen Ortsbezug und damit einen räumlichen Kontext aufweisen. In der fokussierten Aufbereitung von geographischen Inhalten zur gezielten Beantwortung entsprechender Anfragen liegt daher ein großes Potential. Ziel der Dissertation ist die Untersuchung von effizienten Verfahren zum Aufbau einer geographisch fokussierten Suchmaschine, um den Ortsbezug von frei verfügbaren, allgemeinen Webseiten zu nutzen. Dabei werden die Aspekte Crawling, Parsing und Ranking berücksichtigt. Die Arbeit baut auf Grundlagen des Geographischen Information Retrieval auf, dessen Verfahren genutzt und weiterentwickelt werden. Im Gegensatz zum Information Retrieval, das vornehmlich auf Stichworte abzielt, werden relevante Ortsinformationen dabei explizit in ihrer geographischen Semantik berücksichtigt. Durch die Erschließung von im Volltext von Webseiten enthaltenen Informationen, die Erkennung und Extraktion von Ortsbezügen und deren genaue geographische Positionierung erlaubt die Arbeit die Nutzung der geographischen Dimension als Ordnungskriterium neben dem Volltext der Webseiten und ermöglicht somit eine geographische Sicht auf das WWW. Der zentrale Beitrag der Arbeit ist die Entwicklung einer geographischen Fokussierung in zwei Teilbereichen. Zum einen werden effiziente Strategien entwickelt, um einen Webcrawler fokussiert auf ortsbezogene Seiten zu leiten und damit verstärkt relevante Dokumente aufzufinden. Hierzu wird das so genannte Focused Crawling auf die geographische Suche adaptiert, um die Suchmaschine auf einen Ausschnitt des Web mit den gewünschten Ortsinformationen einschränken zu können. Dies basiert auf der Bewertung und Prognose des Auftretens relevanter Seiten (Crawling). Zum anderen wird ein geographischer Parser für die Extraktion von Ortsinformationen in einer hohen Granularität entwickelt, der das präzise, adressgenaue Identifizieren von Ortsinformationen und die Verortung von Ergebnissen ermöglicht. Es wird dabei ein verifizierender Parser realisiert, der unter anderem durch die Einbindung externen Domänenwissens eine hohe Ergebnisqualität gewährleistet (Parsing). Analyseverfahren zur Nutzung der geographischen Dimension der Daten für die Anfrageverarbeitung (Ranking) sowie die Darstellung einer geeigneten Architektur für die geographische Suchmaschine runden die Arbeit ab.

Contents

Weitere Informationen zur Dissertation finden sich unter

http://dhere.de/diss