Expose´ zur Studienarbeit

Extraktion von räumlichen und zeitlichen Informationen aus Webtexten

Nora Popp

Oktober 2007

Betreuer: Bastian Quilitz, Prof. Ulf Leser, Kathrin Poser

HU Berlin, Institut für Informatik

1 Hintergrund

Seit der PC in fast alle Firmen und auch viele Privathaushalte eingezogen ist, werden immer mehr Informationen, die früher ausschließlich auf Papier festgehalten wurden, digitalisiert. Dabei handelt es sich sowohl um neue Daten, die fast ausschließlich elektronisch verarbeitet und gespeichert werden, als auch um ältere Daten, die nachträglich digitalisiert werden.

Mit der Verbreitung des Internet ist die Möglichkeit entstanden, viele dieser Daten der Öffentlichkeit oder zumindest bestimmten Benutzergruppen zugänglich zu machen. Im Laufe der Zeit ist das World Wide Web zu einer riesigen Plattform des Datenaustauschs und nicht nur zu einer möglichen, sonder zu einer wesentlichen Quelle der Informationsgewinnung geworden.

Um mit der entstandenen Flut an Daten, die über das Internet erhältlich ist, überhaupt umgehen zu können, und die jeweils relevanten Informationen zu finden, müssen die vorhanden Daten gefiltert werden. Ein besonderes Problem stellt dabei die Vielzahl von unstrukturierten Texten dar. Sie bieten zwar mengenmäßig das größte Informationspotential, sind aber viel schwieriger als strukturierte Daten auf ihren Informationsgehalt hin zu prüfen.

2 Problemstellung

In vielen dieser unstrukturierten Texte finden sich Angaben zu Raum und Zeit, Informationen, die in ganz verschiedenen Problemfeldern von großem Interesse und Nutzen sind. So ist es zum Beispiel beim Katastrophenmanagement von großer Wichtigkeit, so genaue Informationen wie möglich über den Ort und die Zeit eines Unglücks oder einer Maßname zu erhalten, um dann schnell und effektiv reagieren zu können.

Im Laufe der vergangenen Jahre hat sich gezeigt, dass, im Falle einer Katastrophe, oft private Internetnutzer sehr schnell Informationen dazu in Form von Texten und/oder Bildern veröffentlichen [1]. Diese Informationen sind meist sehr aktuell, da sie nicht erst über den Umweg einer Nachrichtenzentrale oder andere öffentliche Anlaufstellen verfügbar gemacht werden müssen.

Da die textlichen Informationen aber zumeist in unstrukturierter Form vorliegen, und nicht unbedingt offensichtlich relevante Schlagwörter enthalten, ist es oft schwierig bis unmöglich, sie zu finden und zu nutzen. Gerade Ortsangaben können auf sehr unterschiedliche Art und Weise erfolgen, in Form von Koordinaten, Straßennamen oder Städtenamen, von sehr genau beschrieben, bis grob umrissen. Und auch Zeitangaben lassen sich nicht unbedingt über Schlagworte in einer Suchmaske mit einer Katastrophe in Verbindung bringen.

Aus diesem Grund ist es wichtig, Wege zu finden, Informationen über Ort und Zeit eines Geschehens aus Texten extrahieren zu können. Sind diese Daten einmal ermittelt, ist es möglich, sie einem übergeordneten Ort- und Zeitrahmen zuzuordnen und somit nutzbar zu machen.

3 Zur Studienarbeit

3.1 Theoretische Betrachtungen

Zunächst soll ein Überblick über prinzipielle Möglichkeiten zur Informations-extraktion von Orts- und Zeitangaben aus unstrukturierten Texten gegeben werden. Dabei werden verschiedene statistische und computerlinguistische Herangehensweisen betrachtet. Eine wesentliche Grundlage für die Extraktion von bestimmten Informationen bildet die Aufbereitung der Texte. Dazu gehören zum Beispiel die Tokenisierung, die lexikalische Analyse und die (Eigen-) Namenerkennung.
Des weiteren sollen Beispielprojekte vorgestellt werden, in denen Methoden zur Informationsextraktion bereits erfolgreich angewendet werden, wie z.B. in GeoTracker [2] und SPIRIT [3].

3.2 Praktischer Versuch

Im zweiten Teil der Studienarbeit soll eine Methode zur Extraktion von räumlicher und zeitlicher Information implementiert und evaluiert werden.

3.2.1 Methode

Da es sich bei den Informationsquellen um unstrukturierte Texte handelt, müssen zunächst alle einzelnen Token erkannt werden. Diese werden aufgrund verschiedener Regeln bestimmt. Zum Beispiel kann ein Token aus einer Folge von Buchstaben bestehen, die links und rechts von einem Leerzeichen begrenzt ist.

Für das Finden von räumlichen Informationen soll im Rahmen dieser Studienarbeit ein "einfacher" Stringvergleich implementiert werden. Daher müssen die Texte auch nicht spezieller aufgearbeitet werden. Das Hauptaugenmerk bei der Suche nach räumlichen Informationen wird auf Ortsangaben liegen, es sollen also Städtenamen gefunden werden. Um diese zu erkennen, muss ein entsprechendes Wörterbuch erstellt werden. Weiterhin wird eine Liste mit häufig verwendeten Begriffen zur näheren Ortsbestimmung wie *in der Nähe von, südlich, nördlich, ...* angelegt. Diese Liste wird sowohl deutsche als auch englische Begriffe enthalten.

Auch die Suche nach Zeitangaben wird, soweit möglich, über einen Stringvergleich erfolgen, und auch hierfür wird eine deutsch/englische Wortliste angelegt. Darin finden sich dann Monatsnamen, Wochentage und Begriffe zur Zeitbestimmung wie *heute, morgen, letztes Jahr, ...*
Für das Erkennen von Uhrzeitangaben und Jahreszahlen hingegen müssen Regeln aufgestellt werden, die das Format der zu erkennenden Angaben bestimmen.

3.2.2 Korpus

Damit die entwickelte Methode auf Texten getestet werden kann, muss ein Testkorpus erstellt werden. Dafür sollen Daten von flickr.com herangezogen werden. flickr.com ist hauptsächlich eine Plattform für Fotos, aber viele dieser Fotos dort sind mit Beschreibungen versehen und aus diesen soll ein Korpus zusammengestellt werden. Um die Güte der implementierten Methode prüfen zu können, muss das Testkorpus manuell nach allen räumlichen und zeitlichen Informationen die gefunden werden können durchsucht werden.

Literatur

[1] Fahland, D., Gläßer, T. M., Quilitz, B., Weißleder, S. and Leser, U.:
HUODINI - Flexible Information Integration for Disaster Management
4th International Conference on Information Systems for Crisis Response and Management (ISCRAM), Delft, NL (to appear)
http://www.informatik.huberlin.de/forschung/gebiete/wbi/research/publications/2007/huodini_final.pdf

[2] Yih-Farn Chen, Giuseppe Di Fabbrizio, David Gibbon, Rittwik Jana, Serban Jora:
GeoTracker: Geospatial and Temporal RSS Navigation;
http://www2007.org/papers/paper530.pdf

[3] Paul Clough, Mark Sanderson, Hideo Joho: SPIRIT, Spatially-Aware Information Retrieval on the Internet
http://www.geospirit.net/publications/SPIRIT_WP6_D15_geomarkup_revised_FINAL.pdf