H. Wandke, S. Kain & D. Struve (Hrsg.): Mensch & Computer 2009: Grenzenlos frei!?
München: Oldenbourg Verlag, 2009, S. 415-418

Semantisch unterstütze Informations-extraktion aus Dokumentenmengen

Philipp Heim, Timo Stegemann, Steffen Lohmann, Jürgen Ziegler

Interaktive Systeme, Universität Duisburg-Essen

Zusammenfassung

Die Informationssuche in einer Vielzahl von Dokumentenquellen und die Extraktion relevanter Inhalte ist im digitalen Zeitalter eine weit verbreitete Aufgabe. Dieser Beitrag beschreibt ein System, dass alle Teilaktivitäten der textbasierten Informationsextraktion durchgängig unterstützt. Besondere Kennzeichen des Systems sind die semantische Erweiterung von Suchtermen, die Dokumenten-übergreifende Visualisierung von Fundstellen und die einfache Extraktion und Klassifizierung von Textstellen, die sich anschließend strukturiert weiterverarbeiten lassen.

1 Einleitung

Eine häufige Aufgabe moderner Wissensarbeiter besteht im Zusammentragen von Informationen aus unterschiedlichen Quellen. Oft sind die benötigten Informationen über mehrere Dokumente verteilt und innerhalb der Dokumente in andere Inhalte eingebettet. Dies erschwert die manuelle Extraktion von Informationen, da viel Zeit benötigt wird, um alle Dokumente zu durchsuchen. Eine Alternative bieten hier vollautomatische Extraktionsverfahren, deren Ergebnisse jedoch in vielen Fällen unvollständig oder fehlerhaft sind, da sie die Semantik natürlicher Sprache zumeist nicht in ausreichendem Maße erkennen (Over & Liggett 2002). Ein weiteres Problem ist, dass die automatische Informationsextraktion meist nur wenig transparent gestaltet ist, so dass für den Nutzer unklar bleibt, auf welche Weise die Dokumente durchsucht und ob wirklich alle relevanten Informationen extrahiert wurden.

Vielversprechender sind in diesem Zusammenhang Ansätze, die die Informationssuche durch semi-automatische Verfahren unterstützen, beispielsweise im Bereich der semantischen Suche (Wie et al. 2008). Allgemein lässt sich jedoch ein Mangel an durchgängigen Lösungen feststellen, die den Nutzer von der Formulierung der Suchanfrage über die Visualisierung der Fundstellen bis zur Extraktion relevanter Informationen unterstützen. Das in diesem Beitrag vorgestellte System zielt auf eine solche durchgängige Lösung ab und bietet darüber hinaus semantische Unterstützung bei der Formulierung von Suchanfragen.

2 Systembeschreibung

Das System ist in Adobe AIR[1] implementiert. Zur Erweiterung von Suchanfragen nutzt es Web-Services des Projektes Deutscher Wortschatz (PDW 2007). Die Web Services werden ausschließlich für die semantische Erweiterung der Suchanfragen eingesetzt. Die Aufbereitung der Dokumente und die Suche in diesen finden lokal auf dem Computer statt, auf dem das System installiert ist. Somit wird verhindert, dass beispielsweise unternehmenskritische Textinformationen nach außen gelangen.

Entsprechend den Teilaktivitäten bei der Informationsextraktion gliedert sich die Benutzeroberfläche des Systems in drei Bereiche (s. Abb. 1): Der linke Bereich dient der semantisch unterstützen Erstellung von Suchanfragen (A), im Hauptbereich werden die Suchergebnisse in ihrem Kontext dargestellt (B) und der untere Bereich zeigt die extrahierten Informationen (C). Im Folgenden werden die drei Bereiche ausführlicher beschrieben.

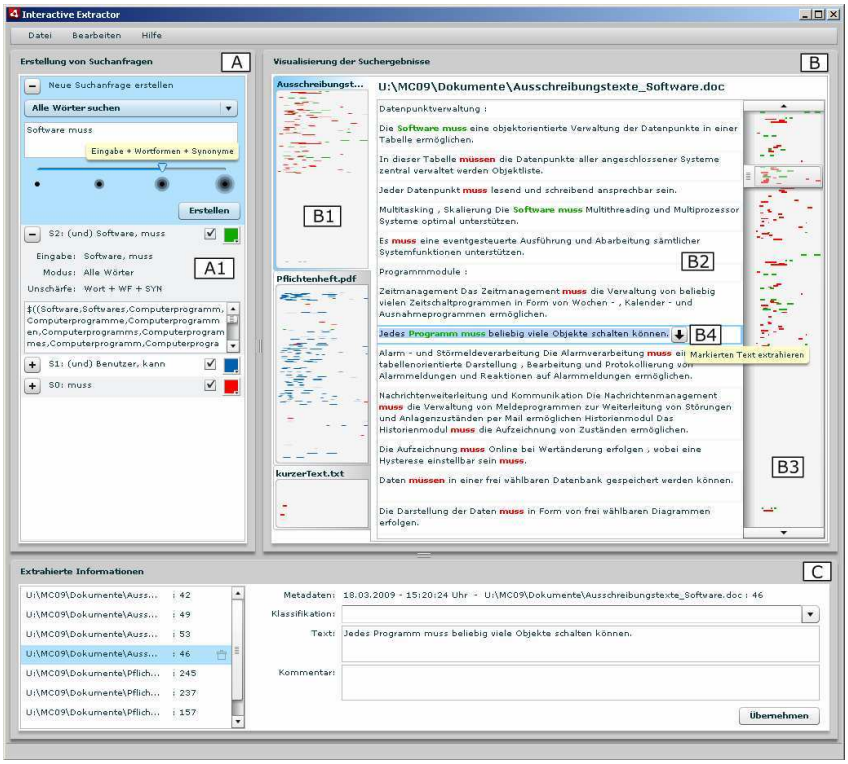

Abbildung 1: Benutzeroberfläche des Systems

[1] Adobe Air: http://www.adobe.com/de/products/air

2.1 Semantisch unterstützte Erstellung von Suchanfragen

Nachdem der Nutzer mehrere Dokumente ausgewählt hat, die er durchsuchen möchte, unterstützt ihn das System bei der Erstellung von Suchanfragen. Eine Suchanfrage wird dabei vom Nutzer zunächst wie üblich in einem oder mehreren Wörtern formuliert. Anschließend kann die Suchanfrage in drei Stufen semantisch erweitert werden.

Die erste Stufe berücksichtigt zusätzlich zu den eingegebenen Suchwörtern alle ihre Wortformen. Hierdurch werden auch Suchergebnisse gefunden, die nicht exakt der eingegebenen Wortform entsprechen. In der zweiten Stufe werden darüber hinaus alle Synonyme der Suchwörter und deren Wortformen bei der Suche berücksichtigt. Die dritte Stufe bezieht neben Synonymen zusätzlich Unterbegriffe aus einem Thesaurus ein, so dass auch spezifischere Varianten der Suchwörter gefunden werden können.

Die gewünschte Stufe der semantischen Erweiterung kann der Nutzer über einen Schieberegler selbst bestimmen (Abb. 1, A). Die Unschärfe der Suche nimmt durch Einbeziehung weiterer Wortformen, Synonyme und Begriffsklassen zu. Für den Nutzer bleibt dennoch jederzeit ersichtlich, welche Wörter und Wortformen in die Suchanfrage eingegangen sind, da die resultierende Liste von Suchtermen nach ihrer Erstellung in einem Textfeld angezeigt wird (A1). Falls der Nutzer möchte, kann er die Liste selbst anpassen, indem er einzelne Wörter entfernt oder weitere ergänzt. Der Benutzer kann zudem die Visualisierung der Fundstellen zu einzelnen Suchanfragen ein- und ausschalten sowie deren farbliche Repräsentation nach Belieben verändern.

2.2 Visualisierung von Suchergebnissen

Die Fundorte der Suchanfragen sind in den Dokumenten in der jeweiligen Farbe gekennzeichnet. Über die vertikal angeordneten Miniaturansichten (B1) erhält der Nutzer, ähnlich wie im System TileBars (Hearst 1995), einen Gesamtüberblick über die Fundstellen in allen Dokumenten. Die Höhe der Miniaturansichten repräsentiert den Umfang der Dokumente. Wählt der Nutzer ein Dokument aus, wird es mit den markierten Fundstellen in der Detailansicht dargestellt (B2).

Wie üblich kann über den Schieberegler des Scrollbalkens der angezeigte Ausschnitt des Dokuments verändert werden. Um die Navigation zu den Fundstellen zu erleichtern, werden diese zusätzlich auch im Scrollbalken angezeigt (B3). Der Bereich des Scrollbalkens, der innerhalb des Schiebereglers liegt, entspricht dabei dem in der Detailansicht dargestellten Ausschnitt des Dokuments.

2.3 Extraktion und Klassifikation von Textstellen

Findet der Nutzer relevante Textstellen, kann er diese auf einfache Weise aus dem jeweiligen Dokument extrahieren und beliebig klassifizieren. Sobald er eine Textstelle in der Detailansicht markiert, wird eine Schaltfläche eingeblendet (B4), mittels der er diese kopieren kann. Da bei der Textaufbereitung eine Satzsegmentierung durchgeführt wird, lassen sich ganze Sätze besonders einfach extrahieren. Die Textinformation wird zusammen mit weiteren Me-

tadaten wie dem Dokumentenpfad, dem Fundort, dem aktuellen Datum und der Uhrzeit in die Sammlung extrahierter Informationen in den unteren Bereich (C) übernommen. Dort kann der Nutzer die extrahierten Informationen klassifizieren und anschließend strukturiert als XML-Datei exportieren.

3 Fazit

Das vorgestellte System bietet eine durchgängige Unterstützung für alle Teilaktivitäten bei der textbasierten Informationsextraktion aus Dokumentenmengen. Insbesondere zeichnet sich das System durch die folgenden Funktionalitäten aus:

- Eine graduelle, semantische Anreicherung von Suchanfragen in drei verschiedenen Intensitätsstufen, die vom Nutzer kontrolliert werden.
- Die gleichzeitige Visualisierung von Fundstellen in mehreren Dokumenten und von verschiedenen Suchanfragen.
- Eine schnelle und einfache Extraktion und Klassifikation von relevanten Textteilen zusammen mit einer Rückverfolgbarkeit zu den jeweiligen Fundorten.

In ersten Anwendungsfällen wurde das System grundsätzlich als hilfreich bewertet. Die allgemeine Form der semantischen Sucherweiterung liefert allerdings bisher noch sehr breite und wenig fachbezogene Unterstützung. Hier ist noch eine Anpassung des verwendeten Wortschatzes auf die jeweilige Domäne bzw. den Unternehmenskontext notwendig.

Literaturverzeichnis

Hearst, M. (1995). *TileBars: Visualization of Term Distribution Information in Full Text Information Access*. In Proceedings of CHI'95, S. 59-66.

Over, P. & Liggett, W. (2002). *Introduction to DUC-2002: an Intrinsic Evaluation of Generic News Text Summarization Systems*. In Proceedings of the Workshop on Automatic Summarization (DUC 2002).

Projekt Deutscher Wortschatz (2007). Universität Leipzig. http://wortschatz.uni-leipzig.de/.

Wei, W., Barnaghi, P.& Bargiela, A. (2008). Search with Meanings: An Overview of Semantic Search Systems. *International Journal of Communications of SIWN, 3*, S. 76-82.

Kontaktinformationen

Philipp Heim, Timo Stegemann, Steffen Lohmann, Jürgen Ziegler

E-Mail: {philipp.heim | timo.stegemann | steffen.lohmann | juergen.ziegler}@uni-due.de