

# InteractiveExtractor: Durchgängige Unterstützung bei der Extraktion von anforderungsrelevanten Informationen

Philipp Heim<sup>1</sup>, Timo Stegemann<sup>1</sup>, Steffen Lohmann<sup>1</sup>, Jürgen Ziegler<sup>1</sup>,  
Haiko Cyriaks<sup>2</sup>, Horst Stolz<sup>2</sup>

<sup>1</sup>Universität Duisburg-Essen, Interaktive Systeme und Interaktionsdesign  
Lotharstraße 65, 47057 Duisburg  
{philipp.heim | timo.stegemann | steffen.lohmann | juergen.ziegler}@uni-due.de

<sup>2</sup>ISA Informationssysteme GmbH  
Azenbergstraße 35, 70174 Stuttgart  
{cyriaks | stolz}@isa.de

**Kurzfassung:** Bei der Erhebung von Anforderungen muss häufig eine Vielzahl bereits vorhandener, in Form und Inhalt verschiedenartiger Dokumente berücksichtigt werden. Dieser Beitrag beschreibt das System *InteractiveExtractor*, das im SoftWiki-Projekt entwickelt wurde und die Extraktion von anforderungsrelevanten Informationen aus bestehenden Dokumenten durchgängig unterstützt. Besondere Kennzeichen des Interactive-Extractor sind die semantische Erweiterung von Sucheingaben, die dokumentenübergreifende Visualisierung von Fundstellen und die einfache Extraktion und Klassifizierung von anforderungsrelevanten Textstellen, die sich anschließend strukturiert weiterverarbeiten lassen.

## 1 Extraktion anforderungsrelevanter Informationen

Das SoftWiki-Projekt verfolgt einen integrierten Requirements-Engineering-Ansatz, der bei der Anforderungserhebung unterschiedlichste Informationsquellen berücksichtigt: Neben der direkten Beteiligung von Stakeholdern in der kollaborativen SoftWiki-Umgebung [\[Querverweis\]](#) sollen anforderungsrelevante Informationen möglichst umfassend auch aus bereits existierenden Dokumentenbeständen, wie beispielsweise Anwendungsfall- und Systembeschreibungen, Gesprächsprotokollen oder Kunden-E-Mails, gewonnen werden (vgl. [\[Querverweis\]](#)). Die manuelle Identifizierung und Extraktion von anforderungsrelevanten Informationen in diesen Dokumentenbeständen ist sehr zeitaufwendig und ohne eine geeignete Systemunterstützung aus ökonomischer Sicht häufig nicht umfassend möglich.

Eine Alternative bieten hier vollautomatische Extraktionsverfahren, deren Ergebnisse jedoch in vielen Fällen unvollständig oder fehlerhaft sind, da sie die Semantik natürlicher Sprache zu meist nicht in ausreichendem Maße erkennen [OL02]. Ein weiteres Problem ist, dass die automatische Informationsextraktion in der Regel nur eine geringe Transparenz aufweist, so dass für den Nutzer oftmals unklar bleibt, auf welche Weise die Dokumente durchsucht und ob wirklich alle relevanten Informationen extrahiert wurden.

Vielversprechender sind in diesem Zusammenhang Ansätze, die die Suche durch semi-automatische Verfahren unterstützen, beispielsweise im Bereich der semantischen Suche [Wei08]. 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 anforderungsrelevanter Informationen unterstützen. Das im SoftWiki-Projekt entwickelte System *InteractiveExtractor* zielt auf eine solche durchgängige Lösung ab und bietet darüber hinaus semantische Unterstützung bei der Formulierung von Suchanfragen.

## 2 Bedienung des InteractiveExtractor

Die Benutzeroberfläche des InteractiveExtractor gliedert sich in drei Bereiche (siehe Abb. 1): Der linke Bereich dient der semantisch unterstützten 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 Funktionen in diesen drei Bereichen ausführlicher beschrieben.

### 2.1 Semantisch unterstützte Erstellung von Suchanfragen

In einem ersten Schritt wählt der Benutzer die Dokumente aus, die er auf anforderungsrelevante Informationen durchsuchen möchte. Dies können beliebige Textdokumente in verschiedenen Dateiformaten sein, die entweder lokal auf dem Rechner oder im Internet bzw. Intranet verfügbar sind (siehe Abschnitt 3.2). Anschließend unterstützt der InteractiveExtractor den Benutzer bei der Erstellung von Suchanfragen. Eine Suchanfrage wird dabei vom Nutzer zunächst wie üblich in einem oder mehreren Wörtern formuliert, die per Konjunktion oder Disjunktion miteinander verknüpft werden können. 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. In dem in Abb. 1 dargestellten Beispiel wird das Eingabewort ‚Software‘ durch die Synonyme ‚Computerprogramm‘ und ‚Programm‘ erweitert. Zusätzlich berücksichtigt die Suchanfrage die entsprechenden Wortformen der Wörter.

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. Er kann zudem die Visualisierung der Fundstellen zu einzelnen Suchanfragen ein- und ausschalten sowie deren farbliche Repräsentation nach Belieben verändern.

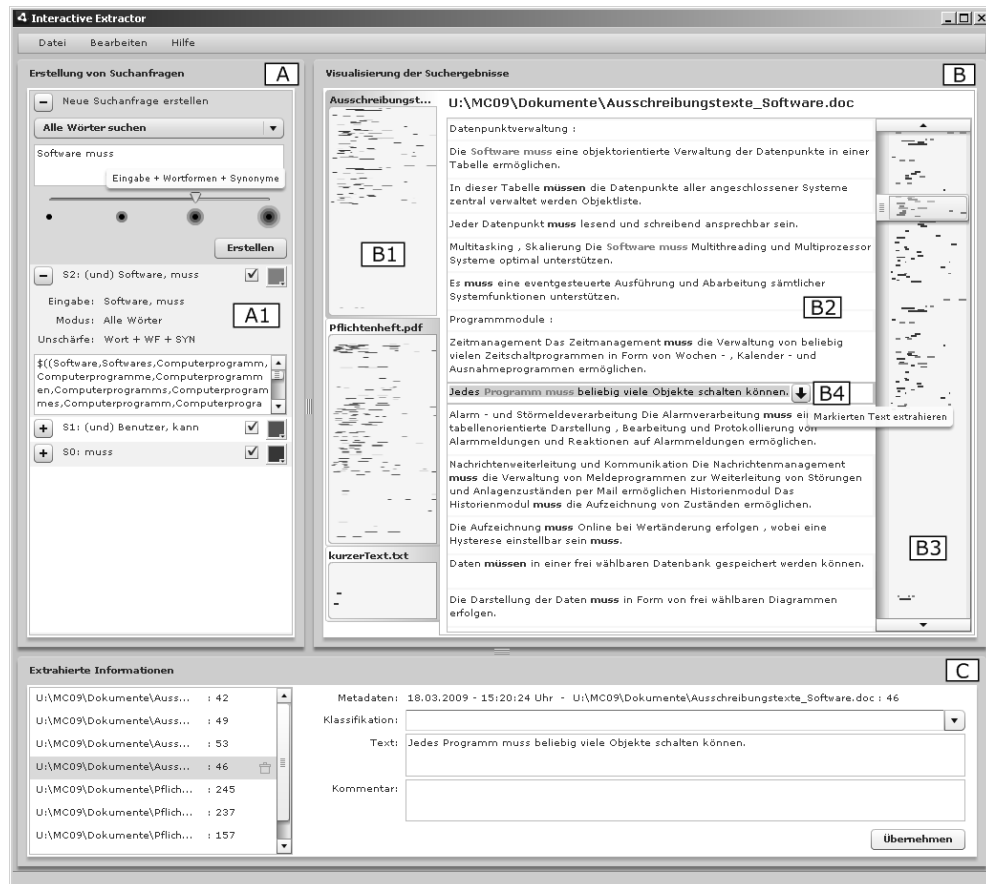


Abbildung 1: Benutzeroberfläche des InteractiveExtractor

## 2.2 Visualisierung von Suchergebnissen

Die Fundorte der Suchanfragen werden in den Dokumenten in der jeweiligen Farbe gekennzeichnet. Über die vertikal angeordneten Miniaturansichten (B1) erhält der Nutzer 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 Detailsicht 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 Detailsicht dargestellten Ausschnitt des Dokuments.

### 2.3 Extraktion und Klassifikation von anforderungsrelevanten Textstellen

Findet der Nutzer anforderungsrelevante Textstellen, kann er diese auf einfache Weise aus dem jeweiligen Dokument extrahieren und klassifizieren. Sobald er eine Textstelle in der Detailansicht markiert, wird eine Schaltfläche eingeblendet (B4), mittels der er diese kopieren kann. Die Textinformation wird zusammen mit weiteren Metadaten 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 zusätzlich mit Schlagwörtern versehen und anschließend als XML-Datei exportieren. Darüber hinaus ist ein Export im RIF-Format<sup>1</sup> (vgl. auch [Querverweis]) möglich, so dass die extrahierten Informationen anschließend über den im SoftWiki-Projekt entwickelten Konverter [Querverweis] in die SoftWiki-Webplattform [Querverweis] überführt und dort für die kollaborative Bearbeitung der Anforderungen herangezogen werden können.

## 3 Architektur des InteractiveExtractor

Der InteractiveExtractor wurde als Client-Server-Architektur realisiert (vgl. Abb. 2). Die geladenen Dokumente werden von einem Textmining-Server aufbereitet und vorgehalten, der als ActiveX<sup>2</sup>-Komponente umgesetzt wurde. Der Client ist in Adobe AIR<sup>3</sup> implementiert und kommuniziert über einen Proxy mit der Server-Komponente. Sowohl die Server- als auch die Client-Komponente laufen lokal auf dem Rechner des Nutzers. Auf das Internet wird nur zugegriffen, um über die Web Services des Projekts Deutscher Wortschatz<sup>4</sup> die vom Nutzer eingegebenen Suchanfragen semantisch zu erweitern. Die Suche in den Dokumenten findet ausschließlich lokal innerhalb der Textmining-Komponente auf dem Rechner des Nutzers statt. Die Dokumente werden somit nicht über ein Netzwerk übertragen, was verhindert, dass unternehmenskritische Informationen nach außen gelangen.

### 3.1 Nutzung der Web Services

Zur semantischen Erweiterungen der Suchanfragen werden drei Web Services des Projekts Deutscher Wortschatz genutzt, die Synonyme, Wortformen und Thesaurus-Informationen zu angefragten Wörtern zurückgeben (vgl. Abb. 2). Die Kommunikation mit den Web Services erfolgt per SOAP<sup>5</sup>.

Die einzelnen Stufen der semantischen Erweiterung bauen aufeinander auf, so dass für eine Erweiterung der Suchanfrage auf einer höheren Stufe auch die Ergebnisse der darunterliegen-

---

<sup>1</sup> Requirements Interchange Format (RIF): <http://www.automotive-his.de/rif>

<sup>2</sup> ActiveX: <http://www.activex.com/>

<sup>3</sup> Adobe Air: <http://www.adobe.com/de/products/air>

<sup>4</sup> Projekt Deutscher Wortschatz: <http://wortschatz.uni-leipzig.de>

<sup>5</sup> SOAP: <http://www.w3.org/TR/soap12>

den Stufen benötigt werden. Die Web-Service-Anfragen zur höchsten Stufe der semantischen Sucherweiterung verlaufen demnach wie folgt:

1. Abfrage der Wortformen des Eingabewortes
2. Abfrage der Synonyme des Eingabewortes
  - 2.1. Abfrage der Wortformen für jedes zuvor erhaltene Synonym
3. Abfrage der Begriffe aus dem Thesaurus für das Eingabewort
  - 3.1. Abfrage der Wortformen für jeden zuvor erhaltenen Begriff

Wurden mehrere Begriffe eingegeben, wird der Vorgang für jeden dieser Begriffe separat ausgeführt. Das Eingabewort und alle Ergebnisse der Web Services zu diesem Wort werden disjunkt verknüpft und bilden die Suchbegriffsmenge. Bei mehreren Eingabewörtern können die entstandenen Suchbegriffsmengen entweder per Konjunktion oder Disjunktion verknüpft werden, woraus sich die Suchanfrage ergibt (vgl. Abschnitt 2.1). Anschließend wird die Suchanfrage als Zeichenkette formatiert über den lokalen Proxy an den Textmining-Server übergeben, der dann die Suche in den aufbereiteten Dokumenten durchführt.

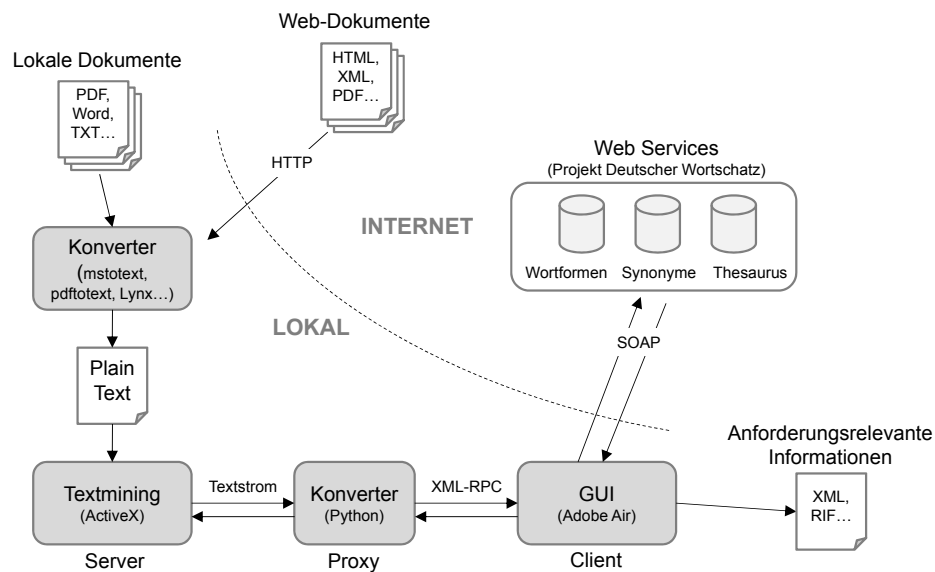


Abbildung 2: Architektur des InteractiveExtractor

### 3.2 Aufbereitung der Dokumente

Der modulare Aufbau der Textmining-Komponente ermöglicht die einfache Anbindung externer Konverter für unterschiedliche Dokumententypen. In der aktuellen Implementierung unterstützt der InteractiveExtractor neben einfachen (ASCII-)Textdokumenten unter anderem auch

den Import von Word-, PDF- oder HTML-Dokumenten (unter Verwendung der Werkzeuge *mstotext*, *pdfiotext* und *Lynx*, vgl. Abb. 2). Webseiten können direkt über die Eingabe ihrer URL importiert werden.

Die Textmining-Komponente arbeitet nach dem Prinzip eines modularen, konfigurierbaren Tokenizers. Der Zeichen-Eingabestrom der zu verarbeitenden Dokumente wird dabei schrittweise in einzelne Einheiten (Token) zerlegt. Der Aufbau der Token bzw. das Zerlegungsmuster kann frei konfiguriert werden. Standardmäßig werden die Dokumente in Sätze und Wörter unterteilt. Denkbar wäre auch eine Aufteilung in Absätze oder andere, logisch sinnvolle Einheiten.

Über die entstandenen Token wird dann ein Index erzeugt. Außerdem wird die Häufigkeit des Auftretens der Wörter ermittelt. Durch Vergleich mit einem optional installierbaren Referenzwortschatz (der fachspezifisch oder allgemeinsprachlich sein kann) lässt sich so signifikant häufiges Auftreten von Wörtern zur Terminologieextraktion (Verschlagwortung) eines Dokumentes nutzen. Die Textmining-Komponente ist auch in der Lage, Kookkurrenzen in Form von einem überdurchschnittlich häufigen, gemeinsamen Auftreten von Wörtern in den Texten zu ermitteln. Hierbei können sowohl Nachbarschafts- als auch Satzkoookkurrenzen berechnet werden (vgl. [\[Querverweis\]](#) und [\[HQW06\]](#)). Diese beiden Funktionen werden vom Client der aktuellen Implementierung des InteractiveExtractors jedoch noch nicht genutzt.

## 4 Fazit

Der InteractiveExtractor bietet eine durchgängige Unterstützung bei der Extraktion anforderungsrelevanter Informationen aus Dokumenten. Insbesondere zeichnet er sich durch die folgenden Funktionalitäten aus:

- Paralleler Zugriff auf mehrere Dokumente unterschiedlicher Dateiformate
- Graduelle, semantische Anreicherung von Suchanfragen in drei verschiedenen Intensitätsstufen, die vom Nutzer kontrolliert werden
- Visualisierung von Fundstellen unterschiedlicher Suchanfragen in mehreren Dokumenten gleichzeitig
- Unterstützung bei der Extraktion atomarer Informationseinheiten durch Segmentierung und Indexierung der Dokumenteninhalte
- Klassifikation der extrahierten Inhalte und automatische Ergänzung von Metadaten, die eine Rückverfolgbarkeit bis zu den Fundorten ermöglichen
- Export der extrahierten Informationen im XML- oder RIF-Format zur Weiterverarbeitung in der SoftWiki-Webplattform (vgl. [\[Querverweis\]](#)) oder anderen Requirements-Engineering-Tools

In ersten Anwendungsfällen aus dem SoftWiki-Projekt wurde der InteractiveExtractor grundsätzlich als hilfreich bewertet. Die allgemeine Form der semantischen Sucherweiterung liefert allerdings bisher sehr breite und wenig fachbezogene Unterstützung. Hier ist noch eine Anpas-

sung des verwendeten Wortschatzes auf die jeweilige Domäne bzw. den Unternehmenskontext notwendig. Eine sinnvolle Quelle für einen angepassten Wortschatz könnten außerdem die Anforderungstexte und Tags aus der kollaborativen SoftWiki-Umgebung bilden [\[Querverweis\]](#).

## **Literatur**

- [Hqw06] Heyer, G., Quasthoff, U. & Wittig, T.: *Text Mining: Wissensrohstoff Text – Konzepte, Algorithmen, Ergebnisse*. W3L, Herdecke/Bochum, 2006.
- [OL02] Over, P. & Liggett, W. Introduction to DUC-2002: an Intrinsic Evaluation of Generic News Text Summarization Systems. In: *Proceedings of the Workshop on Automatic Summarization (DUC 2002)*.
- [Wei08] Wei, W., Barnaghi, P. & Bargiela, A. Search with Meanings: An Overview of Semantic Search Systems. *International Journal of Communications of SIWN*, 3, 2008; S. 76-82.