Semantische Aufbereitung von Dokumentenbeständen zur Gewinnung anforderungsrelevanter Informationen

Haiko Cyriaks[1], Steffen Lohmann[2], Horst Stolz[1], Veli Velioglu[1], Jürgen Ziegler[2]

[1]ISA Informationssysteme GmbH
Azenbergstraße 35, 70174 Stuttgart
{cyriaks, stolz, velioglu}@isa.de

[2]Universität Duisburg-Essen
Abt. Informatik und Angew. Kognitionswissenschaft
Lotharstraße 65, 47057 Duisburg
{lohmann, ziegler}@interactivesystems.info

Abstract: Bei der umfassenden Erhebung von Anforderungen muss häufig eine Vielzahl bereits vorhandener Dokumente berücksichtigt werden. Die Aufbereitung und Auswertung dieser Dokumente kann mit einem hohen Aufwand verbunden sein. Um diese Aktivitäten zu erleichtern, werden im SoftWiki-Ansatz Text Mining-Verfahren eingesetzt, die mittels statistischer und korpuslinguistischer Analysen Dokumentenbestände automatisiert vorverarbeiten. Es werden Worthäufigkeiten berechnet und statistisch signifikante Nachbarschafts- und Satz-Kookkurrenzen identifiziert. Das Ergebnis wird als RDF-Graph ausgegeben und in Form eines semantischen Netzes visualisiert. Hierdurch werden ein thematischer Überblick über den Dokumentenbestand und ein leichterer Zugriff auf Teile davon ermöglicht. Die Visualisierung und aktive Filtermöglichkeiten unterstützen die Identifizierung von anforderungsrelevanten Informationen.

1 Einleitung

Im SoftWiki-Projekt[1] verfolgen wir einen integrierten Requirements Engineering-Ansatz, der bei der Anforderungserhebung unterschiedlichste Informationsquellen berücksichtigt: Neben der direkten Beteiligung von Stakeholdern in der kollaborativen SoftWiki-Umgebung sollen anforderungsrelevante Informationen unter anderem möglichst umfassend auch aus bereits existierenden Dokumentenbeständen wie z.B. Anwendungsfall- und Systembeschreibungen oder Kunden-E-Mails gewonnen werden (vgl. auch [Ha07]). Diese Dokumentenbestände weisen in ihrer Gesamtheit jedoch meist nur einen geringen Strukturierungsgrad und eine große thematische Vielfalt auf. Eine Vorverarbeitung dieser Informationen ist deshalb unabdingbare Voraussetzung für ihre effiziente Integration in die kollaborative SoftWiki-Umgebung. Um den Aufwand der Vorverarbeitung zu reduzieren, sollen die Dokumentenbestände automatisch analysiert und semantisch aufbereitet werden.

[1] http://softwiki.de

Vor diesem Hintergrund werden im SoftWiki-Kontext Text Mining-Verfahren [FS06, HQW06] eingesetzt und weiterentwickelt, die aus Dokumentenbeständen zentrale Begriffe und semantische Beziehungen ermitteln und mit Annotationen versehen. Die im SoftWiki Requirements Engineering Prozess angewandten Verarbeitungsschritte Kollokationsanalyse, RDF-Transformation, Visualisierung und manuelle Bearbeitung sowie die weitere Verwendung der semantischen Struktur werden im Folgenden näher erläutert.

2 Kollokationsanalyse und RDF-Transformation

Eine Kernaktivität im Text Mining ist die Identifizierung von Kollokationen im untersuchten Dokumentenbestand. Der Begriff Kollokation bezeichnet das überdurchschnittlich häufige, gemeinsame Auftreten der gleichen Wörter in einem begrenzten Kontext (Kookurrenz) und ist Indikator für eine semantische Beziehung zwischen diesen Wörtern. Der betrachtete Kontext kann dabei variieren. Im SoftWiki-Ansatz werden alle Kookkurrenzen innerhalb von Sätzen sowie zusätzlich die direkten linken und rechten Nachbarn eines Wortes für die Analyse herangezogen.

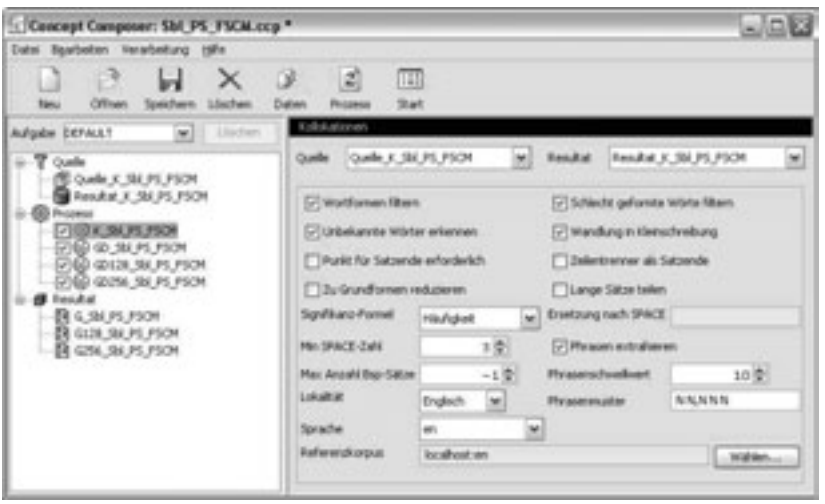

Abbildung 1: ConceptComposer – Parameter für die Kollokationsanalyse

Die Analyse wird mit dem *ConceptComposer* durchgeführt, der zunächst alle Wortformen isoliert, auf ihre Grundformen reduziert und die Häufigkeit ihres Auftretens berechnet. Anschließend werden die Nachbarschafts- und Satzkookkurrenzen ermittelt und mit einem Referenzkorpus abgeglichen. Dieser stellt übliche Häufigkeits- und Kookkurrenzwerte bereit, die durch eine Analyse von etwa zehn Millionen Sätzen ermittelt wurden [HQW02]. Durch den Abgleich mit dem Referenzkorpus lassen sich signifikante Kookkurrenzen identifizieren. Das Ergebnis der Analyse wird in einer relationalen Datenbank abgelegt und bildet die Grundlage für weitere Bearbeitungsschritte.

Über die Benutzeroberfläche des ConceptComposer lässt sich die Kollokationsanalyse anhand verschiedener Parameter konfigurieren (siehe Abbildung 1). Beispielsweise kann definiert werden, wie viele Wörter (bzw. Leerzeichen) ein Satz mindestens enthalten muss, damit er in die Analyse einfließt. Zusätzlich lassen sich Phrasen festlegen, die bei der Analyse berücksichtigt werden sollen. Phrasen bestehen aus mehreren Wörtern und können als Muster (in Form einer Liste von aufeinander folgenden Wortarten) angegeben werden. Treten die gleichen Wörter hinreichend oft entsprechend dem definierten Muster auf, werden sie als Phrase erkannt und extrahiert.

Anschließend werden die ermittelten Begriffe und semantischen Relationen über den Graph Distiller in RDF [Mc04, LS99] transformiert. Abbildung 2 zeigt die Benutzeroberfläche, die diesen Transformationsprozess steuert. Hier lässt sich beispielsweise festlegen, welchen minimalen Signifikanzwert Kollokationen besitzen müssen, damit sie in die RDF-Struktur übernommen werden. Außerdem kann ein Wert für die Anzahl an semantischen Relationen angegeben werden, die eine Wortform in der RDF-Struktur maximal besitzen darf.

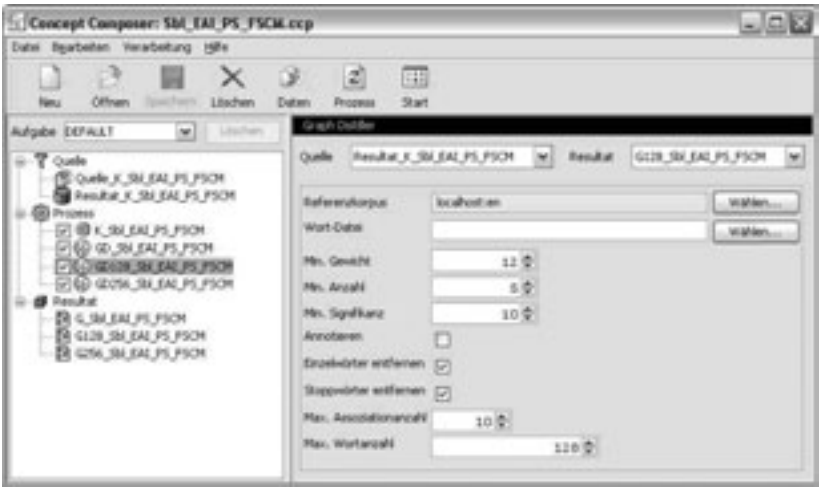

Abbildung 2: ConceptComposer – Parameter für den Graph Distiller

3 Visualisierung

Im nächsten Schritt wird die aus den Quelldokumenten generierte RDF-Struktur in Form eines semantischen Netzes visualisiert. Semantische Netze sind ein beliebtes Mittel der Wissensrepräsentation [ST05, Gr82]. Zentrale Begriffe der betrachteten Domäne werden als Knoten und Beziehungen zwischen diesen Begriffen als Kanten abgebildet. Die Knoten können um Attribute ergänzt werden. Diese netzförmige Verbindung von Begriffen entspricht dem menschlichen Assoziationsdenken und kann einen intuitiven Zugang zu umfangreichen und komplexen Themengebieten darstellen [So91]. Abbildung 3 zeigt ein

semantisches Netz in der Überblicksdarstellung, das aus einem Dokument des SoftWiki-Projektpartners Lecos GmbH generiert wurde, in dem ein spezifischer Anwendungsfall beschrieben wird.

Abbildung 3: Semantisches Netz einer Anwendungsfallbeschreibung

Die Visualisierung des semantischen Netzes auf Basis der zuvor erzeugten RDF-Struktur geschieht mittels *SemanticTalk*. Die extrahierten Wörter werden als Knoten und die ermittelten Kollokationen zwischen den Wörtern als Kanten dargestellt. Die Berechnung der Netzdarstellung basiert auf einem Simulated Annealing Algorithmus [KGV83]. Nach der initialen Generierung des Netzes lässt sich die Anordnung der Knoten manuell beliebig verändern. Bei jedem Knoten ist hinterlegt, an welchen Stellen innerhalb der analysierten Quelldokumente das jeweilige Wort vorkommt.

Abbildung 4 zeigt die Benutzeroberfläche von SemanticTalk. Im rechten Bereich wird das semantische Netz angezeigt, das in diesem Fall aus der Beschreibung eines Customer Relationship Management (CRM)-Systems generiert wurde. Der dargestellte Bereich lässt sich zoomen und verschieben, so dass verschiedene Teilausschnitte wie auch das gesamte Netz betrachtet werden können. Knoten und Kanten können editiert und entfernt werden. Außerdem lassen sich beliebige weitere Knoten, Kanten und Attributehinzufügen. Im linken Bereich der Benutzeroberfläche ist der Graph in einer Miniaturansicht abgebildet, die bei sehr großen Netzen die Übersicht und Orientierung erleichtert. Darüber befinden sich mehrere Regler, mit denen die Ansicht des Graphen angepasst werden kann. Neben dem Zoomfaktor lässt sich hierüber beispielsweise bestimmen, wie viele Kanten des Netzes dargestellt werden sollen. Wird die Anzahl mit dem Regler verringert, werden Kollokationen mit entsprechend geringeren Signifikanzwerten ausgeblendet. Darüber hinaus lassen sich die Knoten nach ihrem Typ filtern, zum Beispiel auf Grundlage der ermittelten Wortarten.

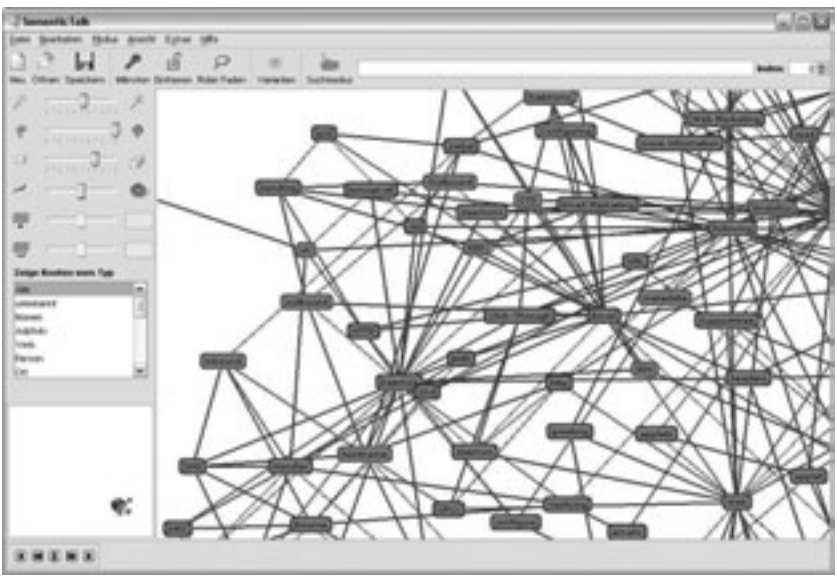

Abbildung 4: Benutzeroberfläche von SemanticTalk

4 Auswertung der semantisch aufbereiteten Inhalte

Die verschiedenen Möglichkeiten der Darstellungsanpassung erleichtern die Identifizierung anforderungsrelevanter Informationen. Insbesondere das attributspezifische Ein- und Ausblenden von Teilen des Graphen unterstützt vielfältige analytische Herangehensweisen. Da jeder Knoten des Graphen auf die entsprechenden Stellen in den Dokumentenbeständen verweist, kann auf spezifische Themen isoliert zugegriffen werden.

Abbildung 5 zeigt einen etwas größeren Ausschnitt des semantischen Netzes, das aus der CRM-Systembescheibung generiert wurde (vgl. Abbildung 4.). Die Visualisierung lässt mehrere Cluster erkennen: Links ein Cluster mit Begriffen, die sich der Konfiguration des E-Mail-Servers und dem Handling im Rahmen von E-Mail-Marketingkampagnen zuordnen lassen; in der Mitte ein Cluster mit Begriffen aus dem Online-Marketing, die häufig im Zusammenhang mit Newslettern und Kunden-Downloads auftreten; rechts unten ein Cluster zum Thema Benutzerschnittstelle und Nutzerinteraktion. Zwischen den beiden oberen Clustern fällt der Begriff „Web“ auf, von dem eine große Zahl Kanten in alle Richtungen abgeht. Solch sternförmig vernetzte Knoten finden sich an mehreren Stellen im Graphen – hierbei handelt es sich in der Regel um zentrale Begriffe eines Clusters oder Bindeglieder zwischen Clustern.

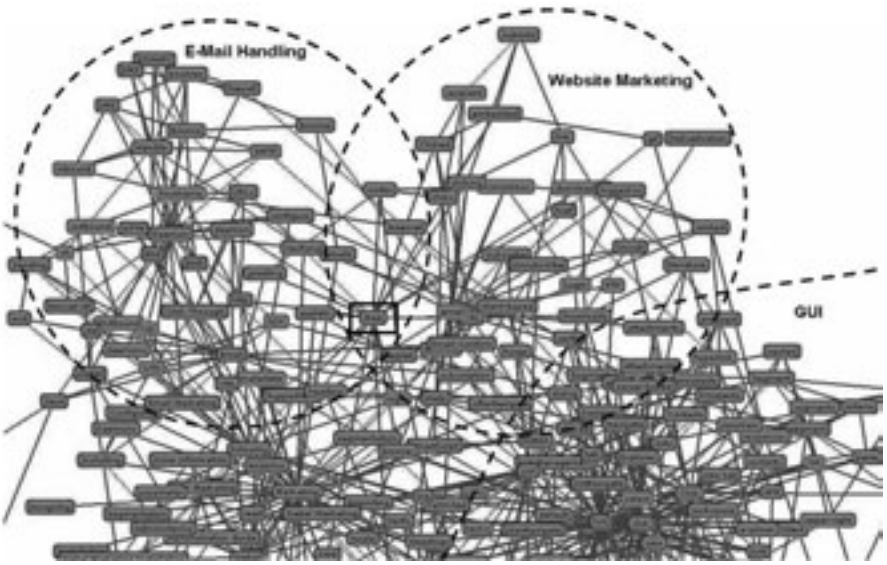

Abbildung 5: Ausschnitt eines semantischen Netzes einer CRM-Systembeschreibung

Die Visualisierung kann nützliche Anreize für domänenspezifische Kategorisierungen von Anforderungen liefern oder auf Begriffe aufmerksam machen, die eine Entsprechung in der Projektontologie (vgl. [LZ07, RL07]) bzw. dem Projektglossar haben sollten. Sofern geeignet, können bestimmte Teile des semantischen Netzes sogar direkt in die Projektontologie übernommen werden. Semantische Beschreibungen für die Begriffe lassen sich dann bestenfalls aus den referenzierten Stellen in den jeweiligen Dokumenten aufgreifen. Soll beispielsweise ein Re-Engineering im Zusammenhang mit dem genannten CRM-System durchgeführt werden, könnten Anforderungen unter anderem anhand der identifizierten Cluster „E-Mail Handling“, „Website Marketing“ und „GUI“ (vgl. Abbildung 5) klassifiziert werden. Fachbegriffe und Akronyme wie „Daemon“ oder „DMZ“ sollten in die Projektontologie bzw. das Projektglossar übernommen und semantisch eindeutig beschrieben werden.

Darüber hinaus kann die Visualisierung der semantischen Struktur kreative Prozesse und die Ideenfindung im Rahmen der Anforderungserhebung anregen. In anderem Zusammenhang haben wir mit dieser Form der Themenextraktion und -visualisierung semantische Kontexte für Gruppensitzungen erzeugt, die zu neuen Impulsen für den Diskussionsverlauf führten und die Gruppenkreativität anregten [Zi05].

5 Fazit und zukünftige Arbeiten

Wie in diesem Beitrag dargestellt wurde, kann die semantische Aufbereitung von schwach strukturierten, heterogenen Dokumentenbeständen zu einem Mehrwert für die Anforderungserhebung führen, der sich insbesondere in verringertem Auswertungsaufwand, verbessertem Überblick sowie themenbezogenem Zugriff auf die Inhalte manifestiert. Die Visualisierung der semantischen Struktur kann kreative Prozesse und die Ideenfindung anregen sowie die Identifizierung von domänenspezifischen Anforderungskategorien und Konzepten für die Projektontologie unterstützen. Anforderungsrelevante Teile des semantischen Netzes sollen in Zukunft unmittelbar in die kollaborative SoftWiki-Umgebung übernommen und dort erweitert und verfeinert werden können. Dies soll helfen, den häufig zu beobachtenden Bruch zwischen kreativen Vorstufen und formalen Modellierungsaktivitäten zu verringern. Da zu jedem Begriffsknoten die referenzierten Textstellen in den Dokumentenbeständen hinterlegt sind, bleibt eine hohe Traceability [RJ01] gewahrt: Es lässt sich leicht nachverfolgen, aus welchen Teilen der Dokumente Anforderungen hervorgegangen sind.

Zukünftige Arbeiten umfassen die Entwicklung gemeinsamer Schnittstellen mit der kollaborativen SoftWiki-Umgebung, um einen effizienten Transfer der semantischen Strukturen zu ermöglichen. Zum einen soll es in Zukunft möglich sein, auf komfortable Weise Teile des semantischen Netzes herauszulösen und in die kollaborative SoftWiki-Umgebung zu integrieren. Andersherum sollen auch in der kollaborativen Umgebung erzeugte RDF-Strukturen in Form von semantischen Netzen visualisiert und editiert werden können.

Danksagung

Wir bedanken uns bei Oliver Pape, Christian Räther und Sabine Köhler von der ISA Informationssysteme GmbH für ihren Input zum vorliegenden Beitrag.

Literaturverzeichnis

[Au06] Auer, S.; Riechert, T.; Fähnrich, K.-P.: SoftWiki – Agiles Requirements-Engineering für Softwareprojekte mit einer großen Anzahl verteilter Stakeholder. In (Meißner, K.; Engelien, M.; Hrsg.): Virtuelle Organisation und Neue Medien 2006: Workshop GeNeMe2006. Gemeinschaften in Neuen Medien. TUDpress, Dresden, 2006.

[FS06] Feldman, R.; Sanger, J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, 2006.

[Gr82] Griffith, R.L.: Three Principles of Representation for Semantic Networks. ACM Transactions on Database Systems, Vol. 7, Nr. 3, 1982; S. 417-442

[Ha07] Hagen, M.; Jungmann, B.; Lauenroth, K.: Ein Prozessmodell für ein agiles und wikibasiertes Requirements Engineering mit Unterstützung durch Semantic-Web-Technologien. In: Proceedings of 1st Conference on Social Semantic Web, LNI, Köllen-Verlag, Bonn, 2007.

[HQW06] Heyer, G.; Quasthoff, U.; Wittig, T.: Text Mining: Wissensrohstoff Text – Konzepte, Algorithmen, Ergebnisse. W3L-Verlag, Herdecke, Bochum, 2006.

[HQW02] Heyer, G.; Quasthoff, U.; Wolff, C.: Automatic Analysis of Large Text Corpora - A Contribution to Structuring WEB Communities. In: Proceedings of the 2nd International Workshop on Innovative Internet Computing Systems (IICS), Springer-Verlag, Berlin, Heidelberg, 2002; S. 15-26

[KGV83] Kirkpatrick, S.; Gelatt, C.D.; Vecchi, M.P.: Optimization by Simulated Annealing. Science, Vol. 220, Nr. 4598, 1983; S. 671-680

[Mc04] McBride, B.: RDF and its Vocabulary Description Language. In: Handbook on Ontologies, Springer-Verlag, Berlin, Heidelberg, 2004.

[LS99] Lassila, O.; Swick, R.R.: Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation, 22 Februar 1999. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222

[LZ07] Lohmann, S.; Ziegler, J.: Partizipationsformen und Entwicklung eines gemeinsamen Verständnisses bei der verteilten Anforderungserhebung. In: Proceedings of 1st Conference on Social Semantic Web, LNI, Köllen-Verlag, Bonn, 2007.

[RJ01] Ramesh B., Jarke M.: Towards Reference Models for Requirements Traceability. IEEE Transactions in Software Engineering, Vol. 27, Nr. 1, 2001; pp. 58-93

[RL07] Riechert, T.; Lohmann, S.: Mapping Cognitive Models to Social Spaces – Collaborative Development of Project Ontologies. In: Proceedings of 1st Conference on Social Semantic Web, LNI, Köllen-Verlag, Bonn, 2007.

[So91] Sowa, J.F. (Hrsg.): Principles of Semantic Networks: Explorations in the Representation of Knowledge, Morgan Kaufmann Publishers, San Mateo, CA, 1991.

[ST05] Steyvers, M., Tenenbaum, J.B.: The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. Cognitive Science, Vol. 29, Nr. 1, 2005; S. 41-78

[Zi04] Ziegler, J.; El Jerroudi, Z.; Böhm, C.; Beinhauer, W.; Busch, R.; Räther, C.: Automatische Themenextraktion aus gesprochener Sprache. In (Keil-Slawik, R.; Selke, H.; Szwillus, G., Hrsg.): Mensch & Computer 2004: Allgegenwärtige Interaktion. Oldenbourg Verlag, München, 2004; S. 281-290.