Topiek - Dokumentenmanagementsystem trifft semantisches Web*

Demonstration

Daniel Klan #, Stefan Hagedorn *, Steffen Hirte #, Heiko Betz #, Firas Kassem #, Kai-Uwe Sattler #

#Department of Computer Science & Automation

TU Ilmenau, Germany

{*first.last*}@tu-ilmenau.de

* NT.AG, Erfurt, Germany

Hagedorn@nt.ag

Abstract

Die Verwaltung großer Datenmengen ist heute eine der wesentlichen Aufgaben im Unternehmensumfeld. Insbesondere das Wiederfinden relevanter und interessanter Informationen gestaltet sich dabei als Herausforderung. Im Folgenden soll mit Topiek ein System vorgestellt werden, welches zu diesem Zweck die Funktionalität herkömmlicher Dokumentenmanagementsysteme um Techniken aus dem Bereich des semantischen Webs erweitert.

1 Einleitung

Die elektronische Verwaltung von Dokumenten spielt eine immer größere Rolle im Unternehmensumfeld. Hunderttausend Dokumente und mehr im Monat sind heute keine Seltenheit in großen Unternehmen. Dokumentenmanagementsysteme (DMS), wie zum Beispiel Alfresco[1] oder Microsoft SharePoint[2], stellen entsprechende Lösungen zur Versionierung, Archivierung und zur Suche über diesen bereit. Der Suchaufwand steigt dabei mit der Menge der verwalteten Daten. Häufig kommen bei der Suche Techniken aus dem Bereich des Information Retrieval zum Einsatz [Ferber, 2003]. Hierbei werden zunächst Text-Indizes über den Dokumenten erstellt, welche anschließend für eine effiziente Suche genutzt werden können. Problematisch gestaltet sich dabei, dass Worte bzw. Wortgruppen nach denen gesucht werden sollen Bestandteil der Dokumentenbasis sein müssen. Wird nach Begriffen gesucht, welcher in dieser Form nicht in den Dokumenten vorkommen, so werden keine entsprechende Dokumente gefunden. Ähnliche Probleme finden sich bei der Informationssuche im Internet wieder. Es hat sich gezeigt, dass ein gezieltes Verschlagworten von Dokumenten zu einer deutlich verbesserten Suche führen kann. Ziel der Verschlagwortung (Annotation, Tagging) ist es, die Bedeutung eines Dokuments in einfachen wenigen Worten wiederzugeben. Einfache Begriffe erlauben somit Schlussfolgerungen auf den Inhalt eines Dokumentes. Neben den Möglichkeiten der erweiterten Suche steht zunehmend auch eine einfache Auswertbarkeit für Maschinen im Vordergrund. Die Vielzahl an Informationsquellen im Internet soll in Zukunft durch den Computer weitgehend selbstständig analysiert und wenn möglich in Beziehung zueinander gesetzt werden. Projekte wie Linked Open Data[3] führen bereits heute Informationen aus vielen hunderten Quellen zusammen.

Eines der größten Probleme bei diesem Ansatz bleibt die Annotation von Dokumenten. Internetplattformen wie Flickr oder Youtube setzen dabei gezielt auf die Mithilfe der Plattformnutzer. Jedoch führt ein manuelles Hinzufügen von Schlagwörtern häufig nicht zu den gewünschten Ergebnissen (zum Beispiel interpretieren Nutzer den Inhalt von Dokumenten unterschiedlich oder die große Datenmengen überfordern diese). Entsprechend sind Techniken zur vollständig autonomen Verschlagwortung von Dokumenten bzw. zur Unterstützung der Nutzer bei der manuellen Verschlagwortung ein aktueller Forschungsgegenstand [Reeve, 2005].

Im weiteren Verlauf dieser Arbeit wird das DMS Topiek vorgestellt, welches im Rahmen einer Demonstration präsentiert werden soll. Topiek erweitert die Möglichkeiten bisheriger DMS um die oben beschriebenen Konzepte aus dem Umfeld des semantischen Webs. So wurden u.a. neben der herkömmlich in DMS vorzufindenden Volltextsuche zusätzlich Techniken zur semantischen Suche integriert. Im Weiteren soll kurz auf die grobe Architektur des Systems sowie auf unsere geplante Demonstration des Prototypen eingegangen werden.

2 Architektur

Topiek folgt einer Drei-Schichten-Architektur (siehe Abbildung 1). Die Datenhaltungsschicht umfasst neben dem eigentlichen Dokumentenpool (*content repository*) zusätzlich noch eine inhaltsbezogene Ontologie, sowie die von den verschiedenen Data Mining Verfahren bestimmten Modelle. Zusätzlich findet sich in der Datenhaltungsschicht noch eine Sammlung aller über den Dokumenten erstellten Schlagwörter (*folksonomy*).

Die Logikebene von Topiek umfasst im Wesentlichen drei Komponenten: den Dokumentmanager, das Framework für die semantische Suche und das Empfehlungssystem. Der Dokumentenmanager übernimmt die Verwaltung von Dokumenten (Hinzufügen, Entfernen, Archivierung, Versionierung). Weiterhin indiziert er den Inhalt jedes einzelnen Dokumentes und bietet damit die Möglichkeit einer effizienten Volltextsuche. Der Dokumentenmanager von Topiek basiert auf dem Apache Framework Jackrabbit[4], welches eine Referenzimplementierung von JCR (*Content Repository for Java Technology*) darstellt.

Die semantische Suche basiert auf einer domainspezifischen Ontologie, welche als RDF-Graph in der Datenschicht abgelegt ist. Aktuell wird eine existierende Ontologie vorausgesetzt. In einer zukünftigen Version soll die

*Funded by the TAB under grant 2010FE9007

[1] http://www.alfresco.com/

[2] http://sharepoint.microsoft.com

[3] http://linkeddata.org/

[4] http://jackrabbit.apache.org/

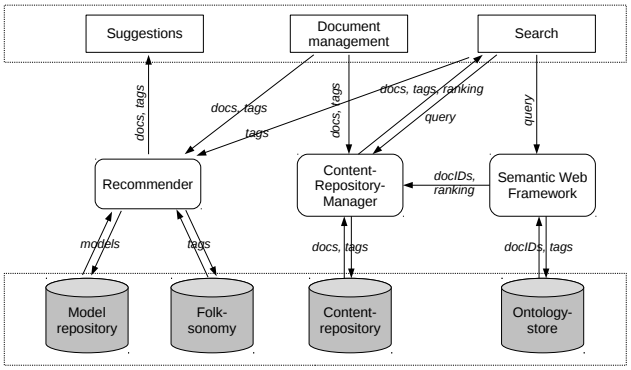

Abbildung 1: Topiek Architektur

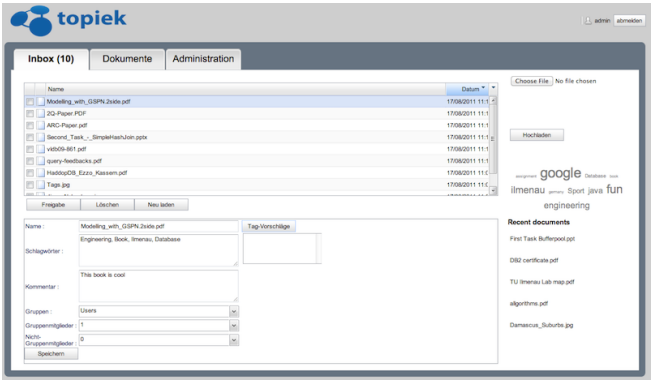

Abbildung 2: Webfrontend - Inbox

Ontologie auf Basis der Folksonomie automatisch erstellt werden. Der Zugriff auf die Ontologie erfolgt über das Semantic Web Framework Jena[5], welches Funktionen zum Laden, Speichern, sowie mit SPARQL eine Technik zum einfachen Anfragen an diese bereitstellt. Über zusätzliche Attribute im RDF-Graphen sind Schlagworte in der Ontologie mit Dokumenten im Dokumentenmanager verknüpft, so dass eine Suche über Schlagwörtern auf entsprechende Dokumente zurückführt.

Die wesentliche Komponente in der Logikebene ist das Empfehlungssystem. Die Komponente hat zwei grundlegende Aufgaben. Wird dem Dokumentenpool ein neues Dokument hinzugefügt, analysiert diese Komponente den Inhalt des Dokumentes und schlägt dem Anwender mögliche Begriffe zur Annotation vor. Die Menge der vorgeschlagenen Tags ist dabei eine Kombination aus zwei Strategien. Bei der ersten Strategie werden die Dokumente bezüglich ihres Inhaltes geclustert. Es wird davon ausgegangen, dass Dokumente, welche sich im gleichen Cluster befinden, auch analog annotiert werden. Entsprechend werden die Schlagworte des Clusters, in den ein neues Dokument eingefügt wird, als Vorschlag übernommen. Die zweite Strategie analysiert ausschließlich den Inhalt des neuen Dokumentes. Nach entsprechenden Vorverarbeitungsschritten werden die extrahierten Konzepte eines Dokumentes dem Anwender als mögliche Schlagworte präsentiert. Die zweite Aufgabe des Empfehlungssystems ist das Vorschlagen von, für den Anwender interessanten, Schlagworten und Dokumenten auf Basis seines bisherigen Such- und Klickverhaltens im DMS. Hierzu werden entsprechende Statistiken über allen Nutzern erfasst. Diese dienen im Weiteren zur Ableitung von Assoziationsregeln. Sucht ein Anwender nach einem oder mehreren Begriffen, zu denen mindestens eine häufige Regel existiert, dann werden ihm die Regelkörper als passende Vorschläge präsentiert. Für die Analyse wurde ein entsprechendes Framework entwickelt, welches unter anderem auf Rapidminer[6] für die Dokumentenvorverarbeitung und den Weka[7]-Tools für das Text Mining basiert.

Die Zugriffsschicht von Topiek stellt Schnittstellen zur Interaktion mit dem System bereit. Neben einer Web-Oberfläche zur Nutzerinteraktion stehen zusätzlich Webservices zur Verfügung, welche der REST-Architektur folgen. Über diese können Desktop-Clients sowie andere Systeme mit Topiek kommunizieren und Topiek so als Erweiterung der eigenen Funktionalität zu nutzen.

Das vollständige System wurde als Java Enterprise Application implementiert, die auf einem Java Application Server ausgeführt wird. Dadurch können Techniken wie Enterprise Java Beans (EJB) und Dependency Injection genutzt werden. Benötigte Objekte können so dynamisch vom Anwendungsserver erstellt und zerstört und Datenbankverbindungen verwaltet werden, was die Skalierbarkeit des Projektes gewährleistet.

3 Demonstration

In der Demonstration soll ein aktueller Prototyp des Topiek-Projektes präsentiert werden. Neben den grundlegenden Funktionen (Management von Dokumenten) eines DMS sollen insbesondere die erweiterten Funktionen von Topiek zur Dokumentenannotation und das Empfehlungssystem präsentiert werden. Der Nutzer kann Dokumente zum System hinzufügen und aus diesem Entfernen. Beim Hinzufügen werden neue Dokumente vom System in Echtzeit analysiert und dem Nutzer wird beim präsentieren der Dokumentinformationen umgehend eine Liste möglicher zum Dokument passender Begriffe vorgeschlagen, aus welchen er für das Dokument relevante auswählen kann. Weiterhin hat der Nutzer die Möglichkeit den aktuellen Dokumentenbestand zu durchsuchen. Hierbei wird er durch das Empfehlungssystem unterstützt, welches die letzten Suchen und Klicks des Nutzers analysiert und zu diesen passende Dokumente bzw. Suchbegriffe vorschlägt. Außerdem bekommt der Nutzer immer eine Liste der aktuell mit den Dokumenten im Bestand am häufigsten verknüpften Tags präsentiert. Zusätzlich besteht natürlich die Möglichkeit einer einfachen Volltext- und einer semantischen Suche.

Für die Nutzerinteraktion mit dem System stehen dem Nutzer zwei grafische Schnittstellen zur Verfügung. Ein Webfrontend (Abbildung 2), welches den vollen Funktionsumfang von Topiek inklusive des Empfehlungssystemes zur Verfügung stellt und eine Desktop-Anwendung, welche lediglich Grundlegende DMS Funktionalität bietet, dafür aber nahtlos in den Desktop integriert ist.

Literatur

[Ferber, 2003] Ferber, R. (2003). *Information Retrieval - Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web*. dpunkt.verlag.

[Reeve, 2005] Reeve, L. (2005). Survey of semantic annotation platforms. In *SAC '05*, pages 1634–1638. ACM Press.

[5] http://openjena.org/

[6] http://rapid-i.com/

[7] http://www.cs.waikato.ac.nz/ml/weka/