Ontologie-basiertes Web Mining

Marc Ehrig[1], Jens Hartmann[1], Christoph Schmitz[2]

[1]Institut AIFB, Universität Karlsruhe, 76128 Karlsruhe
{ehrig,hartmann}@aifb.uni-karlsruhe.de

[2]FG Wissensverarbeitung, Universität Kassel, 34121 Kassel
schmitz@cs.uni-kassel.de

Zusammenfassung: Die Erkennung und Extraktion relevanter Daten im Internet wird zunehmend durch den rapiden Zuwachs an Dokumenten erschwert. Bestehende Ansätze, denen aktuelle Suchmaschinen in der Regel folgen, begegnen den anfallenden Datenmengen mit immer neuer Rechenleistung. Diese Vorgehensweise wird sich jedoch nicht beliebig fortsetzen lassen. In dieser Arbeit stellen wir ein fokussiertes Verfahren zur Identifikation und Extraktion kontextrelevanter Informationen aus dem Internet vor, welches Hintergrundwissen in Form von Ontologien verwendet.

Einführung - Web Mining

Die Anwendung von Data-Mining-Methoden zur Erkennung von Regularitäten in Daten auf das World Wide Web wird *Web Mining* genannt. Im Allgemeinen wird Web Mining in folgende drei Bereiche unterteilt:

- *Web Content Mining:* Die Erkennung von Regularitäten in Texten und Multi-Media Objekten (beispielsweise Grafiken) in Web Dokumenten.
- *Web Usage Mining:* Die Erkennung von Regularitäten in der Benutzung von Web Dokumenten.
- *Web Structure Mining:* Die Erkennung von Regularitäten in der Struktur von Web Dokumenten und ihrer Relationen.

Im Kontext der Erkennung neuer und nützlicher Ressourcen beispielsweise für ein semantisches Informationsportal [SEAL03], erscheint die kombinierte Verwendung von Content und Structure Mining Methoden sinnvoll. Erst die intelligente Kombination von Methoden zur inhaltlichen Analyse von Ressourcen sowie deren relationalen Betrachtung lässt bedarfsgerechte Schlüsse über die Zugehörigkeit und Relevanz einer Ressource für ein Informationsportal zu. Des Weiteren lässt sich vorhandenes Wissen für eine gezielte Suche verwenden, d.h. auf Basis von Hintergrundwissen über eine Domäne wird die Wertigkeit von Ressourcen in Bezug auf die Zugehörigkeit und Relevanz abgeschätzt. Dadurch kann die Suche nach potentiell bedeutsamen Ressourcen fokussiert werden.

Das Werkzeug welches zur Identifikation und Extraktion von Ressourcen aus dem Internet eingesetzt wird, wird im Allgemeinen *Web Crawler* genannt. Bestehende Web Crawler unterscheiden sich zum Teil in ihrer Konzeption und somit auch in ihrer Architektur stark, welches sich auf die jeweilige Verwendung der gewonnen Ressourcen bzw. der jeweiligen Anwendung zurückführen lässt [CHA02].

Es lassen sich dabei zwei grundlegende Arten von Crawlern unterscheiden. Zum einen Crawler, die die Dokumente aufgrund ihrer Verlinkung einsammeln, wie beispielsweise der Crawler von Google[1] Webbase [RAG99]. Dem Gegenüber stehen inhaltsbasierte Crawler, welche die enthaltenen Dokumenten-Texte in eine Relevanzbewertung einbeziehen und so den Crawling-Prozess fokussieren [EHR02]. Ein großer Nachteil bestehender Web Crawler ist deren proprietäre Ausrichtung auf eine bestimmte Anwendung hin. Eine Weiter-verwendung bzw. Erweiterung ist daher i.d.R. schwierig bis unmöglich. Die von uns entwickelte Konzeption stellt im Allgemeinen einen Ansatz für eine modulare und flexible Methodik zur intelligenten Erkennung und Extraktion von Wissen ab, die sich auf das Internet sowie auf Dokumente in Intranets bspw. in Unternehmen und Einrichtungen der öffentlichen Hand anwenden lässt.

METIS – Ein ontologie-basierter Web-Crawler

Allgemeiner Überblick

Die Konzeption unseres Web Crawlers **METIS** sieht die Verwendung von mehreren Modulen vor, welche flexibel austauschbar und erweiterbar sind. In Abbildung 1 wird die allgemeine System-architektur dargestellt. Generell ist das *Crawler-Modul* für das eigentliche Holen der Daten aus dem WWW zuständig, welche dann im *Preprocessing-Modul* weiter verarbeitet werden (Fehlerkorrektur, Datentransformation, etc.). Die verarbeiteten Seiten werden indiziert und bspw. in eine Datenbank gespeichert worauf das *Computation-Modul* mittels dem *Ontology-Modul* die jeweilige Relevanz einer Seite berechnet und dem *Crawling-Modul* die nächsten Webseiten angibt, die geholt werden sollen.

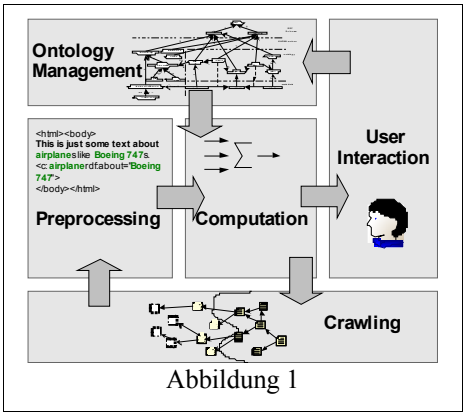

Abbildung 1

Wissensrepräsentation

Grundlage des entwickelten Crawlers für die gezielte Suche nach Ressourcen im Internet ist die Verwendung von *Hintergrundwissen* in Form von Ontologien, welche eine formale und interpretierbare Repräsentation von Wissen für Menschen und Maschinen erlauben. Dabei wird zwischen mehreren Arten von Ontologien unterschieden: Zum einen wird eine *Crawler-Ontologie* modelliert. Diese beschreibt die Struktur und Eigenschaften von Dokumenten im Internet, sowie deren Verknüpfungen mittels sog. Hyperlinks. In der *Domänen-Ontologie* wird die eigentliche Domäne beschrieben, also die Konzepte und Beziehungen, die in der betrachteten Anwendung eine Rolle spielen. Zur Handhabung der jeweiligen Ontologien wird die am AIFB entwickelte, frei verfügbare Software KAON [MOT02]eingesetzt.

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

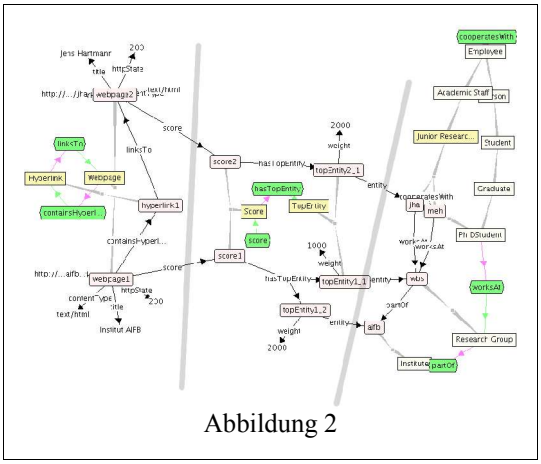

Abbildung 2

Über den Ontologie-Inklusionsmechanismus von KAON kann dann die Crawler-Ontologie sich die Domänenontologie zu Nutze machen, um zu beschreiben, welche Konzepte oder Instanzen (im Folgenden: Entitäten) in den jeweiligen Webseiten und Links relevant für die Belange des Nutzers sind. Exemplarisch zeigt die Abbildung 2 eine Ontologie. Auf der linken Seite ist die Crawler-Ontologie zu sehen mit zwei beispielhaften Instanzen, die Webseiten repräsentieren, sowie einem Hyperlink zwischen diesen beiden. Die rechte Seite zeigt einen Ausschnitt aus der Domänenontologie, in der Forscher, Institute, Publikationen und die Beziehungen zwischen diesen modelliert werden.

Der mittlere Teil zeigt, wie Web- und Domänenontologie verknüpft werden: über die inhaltliche Bewertung (s. nächster Abschnitt) wird bestimmt, welche Entitäten für eine Webseite eine Rolle spielen. Die wichtigsten davon werden mit ihrem Gewicht als „Score“ in die Ontologie aufgenommen. Diese Scores stellen dann den Bezug von Web-Entitäten wie Seiten und Links zu Domänen-Entitäten wie „Institut AIFB“ oder „Wissenschaftlicher Mitarbeiter Jens Hartmann“ her.

Bewertung von Inhalt und Linkstruktur

Die große Informationsdichte und -vielfalt, insbesondere im Internet, erfordert eine detaillierte Analyse und Beurteilung der gefundenen Ressourcen. Ressourcen weisen einen unterschiedlich hohen Informationsgehalt oder allgemein eine unterschiedliche

```
<html>
  <body>
       This page is about the Institute AIFB at the University of
       Karlsruhe.<br>
       <a href="http://www.aifb.de/index.html">AIFB</a><br>
       <a href="http://www.interest.com/institutes.html">
         Interesting Institutes</a><br>
       <a href="http://semanticweb.org/index.html">
         Semantic Web</a>
  </body>
</html>
```

Beispiel 3: Internetressource zum Thema AIFB

Relevanz zur gegeben Domäne auf.

Im Gegensatz zu Informationsextraktionsmechanismen, die eine Bewertung von Ressourcen erst nach der eigentlichen Extraktion zulassen, zielt der von uns entwickelte Ansatz auf eine Bewertung von Ressourcen während der eigentlichen Suche ab. Dies ermöglicht eine effektive und effiziente Nutzung vorhandener Kapazitäten.

Die fokussierte Suche nach Ressourcen basiert auf der Bestimmung der Relevanz einer Ressource zu einer bestimmten Domäne oder Teildomäne. Generell wird als Suchstrategie die Verfolgung von Relationen zu Ressourcen mit möglichst hoher Relevanz verwendet („fokussiertes Crawlen“).

Die inhaltliche Analyse von Dokumenten basiert zunächst auf klassischen Verfahren aus dem Bereich des Text Minings. Dabei wird der Text nach einer Vorverarbeitung, welche aus der Entfernung von Stoppwörtern und einer Rückbildung auf den Wortstamm besteht, analog zur gegeben Domänen-Ontologie untersucht. Hierbei wird für die Existenz von Worten im Text eine graduell verteilte Bewertung zum gesuchten Konzept (Ziel der Suche) in der Domänen-Ontologie vorgenommen. Worte und Textphrasen die einen nahen semantischen Bezug zur Domänen-Ontologie aufweisen werden demzufolge höher gewertet. Die Summe aller Einzelergebnisse ergibt ein Bewertungsmaß welches die inhaltliche Relevanz eines Dokuments abschätzt. Die Extraktion von Informationen aus Internetressourcen unter Berücksichtigung von Ontologien wurde bereits in [CRA00] beschrieben. Zu Veranschaulichung verweisen wir auf ein Beispiel: als Suchdomäne wird das Konzept „AIFB“ aus der Ontologie in Abbildung angenommen. In der heruntergeladenen Ressource sind dieser Term sowie der semantisch verwandte Term „Institute“ enthalten. Aggregiert ergibt sich ein Wert von 1,5; 1 für „AIFB“ plus 0,5 für „Institute“.

Die Bewertung der Verlinkung beruht auf der Analyse der eingehenden und ausgehenden Relationen eines Dokumentes. Eingehende Links sind dabei Links von anderen Dokumenten auf das zu untersuchende Dokument. Da dies ein Faktor ist, der nur schwer zu beeinflussen ist, bietet sich dessen Bewertung als Kriterium zur Relevanzbestimmung an. Ausgehende Links sind offensichtlich Links des zu untersuchenden Dokumentes auf andere Dokumente. Hierbei wird versucht von der Art der verlinkten Dokumente auf das Ursprungsdokument zu schließen. Bei der Bewertung eines Links wird dessen Text (sog. Ankertext) und Worte in der näheren Umgebung verwendet (bspw. der Abschnitt in dem der Link enthalten ist). Die Bewertung basiert dabei wiederum auf der Analyse der im Text enthaltenen Worte und in der Domänen-Ontologie, wobei hier graduell absteigend vom Linktext (Ankertext) zum umliegenden Text bewertet wird. Im Gegensatz zu klassischen Verfahren, dem HITS-Algorithmus [KLE99] oder dem PageRank-Algorithmus [PAG98] von Google, kann dieser Ansatz zur Laufzeit, also während der Suche angewendet werden und kann explizit semantisches Wissen zur fokussierten Suche einsetzen. Im gegebenen Beispiel würde die Ressource `http://www.aifb.de/index.html` eine Bewertung von 1, `http://www.interest.com/institutes.html` von 2 und `http://semantic-web.org/index.html` von 0 erhalten.

Die endgültige Berechnung der Relevanz setzt sich zusammen aus der inhaltlichen Analyse des Dokumenteninhalts und der Bewertung der Verlinkung der Ressource. Das Ergebnis ist ein numerischer Wert, welcher zur Erzeugung einer sortierten Menge an Ressourcen bzw. Links verwendet wird. Diese Menge dient zur weiteren Suche, wobei Ressourcen mit höherer Relevanz vorrangig verarbeitet werden. Aus unserem Beispiel ergäbe sich eine Reihenfolge von `http://www.aifb.de/index`, `http://www.interest.com/institutes.html`, `http://semanticweb.org/index.html`.

Zusammenfassung

Zusammenfassend werden Ontologien zur Modellierung der Umgebungswelt und zur Modellierung der eigentlichen Anwendung (Domäne) eingesetzt. Die Domänen-Ontologie beschreibt dabei Konzepte zu denen weitere Instanzen identifiziert werden sollen. Die Suche nach diesen Instanzen wird mittels einer semantischen Bewertung durchgeführt. Die Suchstrategie ist eine fokussierte Suche, da sie höherwertige Relevanzbewertungen vorrangig verfolgt.

Zur qualitativen Bewertung der Suchergebnisse können herkömmliche Kriterien nicht herangezogen werden, da sie in der Regel nur Aussagen über die Anzahl der Dokumente pro Zeiteinheit erlauben. Die Entwicklung solcher qualitativer Bewertungskriterien ist Bestandteil gegenwärtiger Forschungsarbeiten.

Durch die freie Verfügbarkeit von **METIS**[2] fördern wir die Anwendung und Erweiterung durch Interessierte.

Danksagungen

Wir danken allen Mitarbeitern des AIFB für hilfreiche Kommentare und Diskussionen. Teilbereiche dieser Arbeit wurden vom BMBF durch das Projekt SemIPort gefördert.

Literaturverzeichnis

[CHA02] Chakrabarti, S.: Mining the Web: Analysis of Hypertext and Semi Structured Data. Morgan Kaufmann; 1st edition (August 15, 2002).

[CRA00] Craven, M.; DiPasquo, D.; Freitag, D.; McCallum, A.; Mitchell, T.; Nigam, K.; Slattery, S.: Learning to Construct Knowledge Bases from the World Wide Web (2000)

[EHR03] Ehrig, M.; Maedche, A.: Ontology-Focused Crawling of Web Documents. In: Proc. of the Symposium on Applied Computing 2003 (SAC 2003), March 9-12, Melbourne, Florida, USA, 2003

[KLE99] Kleinberg, J. M.: Authoritative Sources in a Hyperlinked Environment. Journal of the ACM, Volume 46, 1999, Pages 604—632

[MOT02] Motik, B.; Oberle, D.; Staab, S.; Studer, S.; Volz, R.: KAON SERVER Architecture. In: Forschungsberichte „Rote Reihe“, Universität Karlsruhe, Bericht 421, Sep. 2002

[PAG98] Page, L.; Brin, S.; Motwani, R.; Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. Stanford Digital Library Technologies Project, 1998

[RAG99] Hirai, J.; Raghavan, S.; Garcia-Molina, H.; Paepcke, A.: WebBase : A repository of web pages. In: Computer Networks, Journal 33, Amsterdam – Netherlands, 1999

[SEAL02] Hartmann, J.; Sure, Y.: Scalable and Reliable Semantic Portals (SEAL) in Practice. In: In Proc. of International Conference on Ontologies, Databases and Applications of SEmantics (ODBASE 2003), 3-7 November 2003, Catania, Sicily (Italy).

[2] http://metis.ontoware.org