Fehlertolerantes Wrapping für Föderierte Informationssysteme*

Florian Jung und Dirk Rother
{fjung,drother}@cs.tu-berlin.de

Computergestützte Informationssysteme CIS, Technische Universität Berlin
Dezember 2002

1 Motivation

Integration von Informationen aus dem Internet bietet ein großes Potenzial an Anwendungsszenarien. Da ein Großteil der Informationsquellen keinen direkten Zugang zu den Daten (z.B. mittels Webservices) erlaubt, können die Daten nur durch Informationsextraktion unter Anwendung von Wrapperkonzepten gewonnen werden.

In diesem Kontext haben Wrapper die Aufgabe, die Informationen aus Dokumenten in semantische Strukturen zu überführen. Die Langlebigkeit solcher Wrapper hängt maßgeblich von der Beständigkeit der Struktur der informationstragenden Internetseiten ab.

2 Generischer Wrapper auf Basis von XML Technologien

Bei der Informationsgewinnung unterscheiden wir zwischen *aktivem* und *passivem* Wrappen von Internetseiten. Dem Wrappen liegen dynamisch erzeugte Internetseiten zugrunde. Beim passiven Wrappen sind die Seiten jedoch nicht Ergebnis einer Benutzeranfrage. Im Gegensatz zum passiven wird beim aktiven Wrappen eine Anfrage an ein Informationsangebot geschickt und die Ergebnisseite als Informationsquelle der Anfrage genutzt.

Unser Wissen über invariante Bereiche einer Internetseite erlaubt es, dem Wrapper die Informationen über die Angabe von Suchmarken aufzufinden.

Bei der Suche nach Informationen gehen wir mit folgender Strategie vor: Wir greifen auf eine Darstellung der Internetseiten als DOM-Bäume zurück. Da HTML-Seiten selten XML-konform sind, müssen diese erst in diesen Zustand überführt werden. Aus diesen XHTML-Dokumenten lassen sich dann die DOM-Bäume ableiten. Mittels XPath-Angaben[Bec02] kann zum Auffinden von Strukturen durch die DOM-Bäume navigiert werden.

Oft finden sich relevante Informationen semistrukturiert[ABS00,Suc97a,Suc97b] in Tabellen wieder. Das Kernstück unseres Wrappers bildet also der Teil, der den Umgang mit in Listboxen zugrunde liegender Tabellen beherrscht. Die Daten (content values) sind dann in den Tabellenzellen wiederzufinden. Filterangaben ermöglichen es, die Daten auf die gewünschte Information zu reduzieren.

Die Verteilung großer Mengen von Informationen auf mehrere Internetseiten die miteinander verlinkt sind, bereitet dahingehend keine Probleme, weil der Wrapper in der Lage ist solchen Links zu folgen.

Die Beschreibung der Informationen die mittels des Wrappers aus Dokumenten gewonnen werden sollen, werden in einer leicht verständlichen Syntax verfasst. Die Beschreibungssprache ist eine Anwendung des allgemeinen Austauschformats XML.

Die Charakteristika eines Wrappers für eine spezielle Dokumentengruppe werden direkt aus der Beschreibung abgeleitet. Aus diesem Grund ist es für den Benutzer nicht notwendig, dass er sich mit der Struktur und der Navigation in DOM-Bäumen auskennt.

* Die Arbeit ist eine laufende interne Projektarbeit (vergleichbar einer Studienarbeit), die wir unter Anleitung von Dr. Ralf-Detlef Kutsche am Lehrstuhl Computergestützte Informationssysteme CIS / TU Berlin (Prof. Dr.-Ing. Herbert Weber) durchführen.

3 Evolutionsfähigkeit und Fehlertoleranz

Internetseiten unterliegen in kurzen Zeiträumen häufigen Änderungen. Für Wrapper, die sich solchen Gegebenheiten nicht oder nur unzureichend anpassen, haben solche Änderungen ein undefiniertes Verhalten zur Folge. Dieser Vielfalt an Änderungsmöglichkeiten steht nur ein kleiner Teil an Reaktionen auf solche Änderungen gegenüber.

Alle relevanten Änderungen an den informationstragenden Dokumenten werden protokolliert. Mit Hilfe dieser Historisierung ist es dem Anwender möglich, Einsicht über den Verlauf von Änderungen und der daraus folgenden Anpassung des Wrappers zu erlangen.

Im Folgenden skizzieren wir Änderungsszenarien, die für evolutionäre Softwareentwicklung uns besonders relevant erscheinen.

3.1 Änderung von Suchmarken

Durch Umbenennung können Suchmarken, die für das Auffinden der Informationen nötig waren, aus den Dokumenten verschwinden. Ändert sich nur der Name der Suchmarke, ohne dass die Struktur des Dokumentes Änderungen aufweist, kann durch eine kleine Korrektur des Wrappers ein Fehlschlag leicht vermieden werden.

Ist die Suchmarke für die informationstragende Struktur in irgendeiner Weise charakteristisch, so kann mit Hilfe einer lokalen Synonymliste nach verwandten Begriffen gesucht werden.

Beim Auffinden einer solchen Alternativsuchmarke arbeitet der Wrapper problemlos weiter, sofern sich die Struktur des relevanten Bereichs des Dokumentes nicht geändert hat. Dabei betrachten wir zunächst nur den Fall, dass sich nur die Suchmarke und nicht der relevante Inhalt selbst geändert hat (ansonsten siehe 3.2ff).

Diese Synonymlisten basieren auf domänenspezifischen Thesauri und enthalten, je nach Anwendung, nicht nur Synonyme, sondern ggf. auch Ober- und Unterbegriffe. Durch Vertauschung der nicht mehr auffindbaren und Ergebnis liefernden Suchmarke wird bei der nächsten Verwendung des Wrappers nicht mehr die alte, sondern gleich die neue Suchmarke verwendet. Dabei wird die Reihenfolge der Begriffe in der Synonymliste dahingehend verändert, dass der das Ergebnis liefernde Begriff an den Anfang der Liste, der nicht mehr auffindbare Begriff um eine Position nach hinten verschoben wird.

Dadurch erreichen wir einen erhöhten Grad an Robustheit.

3.2 Änderungen in der Struktur

Da die die Information tragenden Tabellen selbst die Suchmarken, d.h. die invarianten Merkmale, enthalten, an denen sie aufzufinden sind, braucht der Wrapper auf Änderungen ausserhalb der Tabellen nicht zu reagieren. Änderungen die sich auf Formatierungsangaben beziehen werden ebenfalls ignoriert. Werden Spalten- und/oder Zeilenanzahl geändert, so ist dies auch nur dann ausschlaggebend, wenn der Inhalt der Spalte/Zeile sich ändert, also die uns interessierende Information „verschoben“ wurde.

3.3 Änderungen im Inhalt

Der Ansatz des aktiven Wrappens ermöglicht es uns, eine Erweiterung des Informationsangebots zu berücksichtigen, indem die Listboxen, die den Anfragen zugrunde liegen, dynamisch beim Wrapperaufruf ausgelesen werden.

3.4 Änderungen der URL

Ist die informationstragende Seite unter der bisherigen URL nicht mehr zu finden, so liegt eine schwerwiegende Änderung vor. Existiert unter der bisherigen URL eine Weiterleitung, also ein Link, bzw. wurde die Seite in ein Frame eingebettet, so besteht die Möglichkeit mit dem hier vorgestellten Wrapper auf der verlinkten Seite die Information zu finden.

4 Zusammenfassung und Ausblick

Die häufigen Veränderungen von Internetseiten erfordern die Entwicklung von evolutionsfähigen Wrappern. Die hier vorgestellten Wrapperkonzepte spiegeln den derzeitigen Entwicklungsstand unseres Wrappers in der Version 0.8 wieder.

Um unserem Wrapper diese Fähigkeit zu geben, ist die Verwendung globaler Thesauri zur Erzeugung von Synonymlisten ebenso in Arbeit, wie die Benutzung regulärer Ausdrücke zur Informationsidentifikation.

In der mittelfristigen Perspektive soll hier ein Werkzeug entstehen, das hilft das Wrapping autonomer, stark variierender Web-Quellen, ein Stück weit zu automatisieren (mit einer unvermeidbaren interaktiven Komponente) und darüber hinaus neue Quellen leicht in eine Föderation einzubinden.

5 Verwandte Arbeiten

Bei der Entwicklung unseres Wrappers haben wir folgende Quellen als Bezugspunkte herangezogen. Diese Arbeiten sind themenverwandt und bieten daher einen groben Überblick über die verschiedenen Ansätze des Wrappens und der Mediation von Informationen.

- World Wide Web Wrapper Factory (W4F), Tropea Inc., 1999
- The Standford IBM Manager of Multiple Information Sources (TSIMMIS), Univ. Stanford & IBM, 1995/97
- Java Extraction and Dissemination of Information (JEDI), Fraunhofer Institut Integrierte Publikations- und Informationssysteme

Literatur

[ABS00] S. Abiteboul, P. Buneman, D. Suciu: Data on the Web - From Relations to Semistructured Data and XML, Morgan Kaufmann, Publishers, San Francisco, California, 2000

[Bec02] O. Becker: XML Path Language (XPath), deutsche, kommentierte Übersetzung, http://www.obqo.de/w3c-trans/xpath-de-20020226, 2002

[Suc97a] D. Suciu: An Overview of Semistructured Data, In *Database Theory Column*, 1997

[Suc97b] D. Suciu: Semistructured Data and XML, In *Proceedings of International Conference on Foundations of Data Organization*, 1998

[GRVB] J.-R. Gruser, L. Raschid, M. E. Vidal, L. Bright: Wrapper Generation for Web Accessible Data Sources, Univ. of Maryland