Das XIRCUS Projekt

Lehrstuhl: Datenbanken und Informationssysteme, Universität Rostock
Vortragende: Lars Milewski
Ines Weber
Betreuer: Prof. Dr. rer. nat. habil. Andreas Heuer,
Dr.-Ing. Holger Meyer,
Dipl.-Inf. Ilvio Bruder

Es ist unbestritten, dass die Bedeutung und Verbreitung von XML, vor allem im Internet, in den letzten Jahren rasant zugenommen hat. Ein Ende dieses Trends ist noch nicht abzusehen. XML ist auf dem Weg, ein allgemeines Datenaustauschformat zu werden. Als Regelsammlung zur Definition von erweiterbaren Auszeichnungssprachen wird XML durch die Trennung des Layouts von dem Inhalt und der Struktur der Daten charakterisiert. Damit kann der Nutzer die Bestandteile des Dokumentes selbst festlegen. Dadurch hat sich XML weit über den Einsatz als HTML-Ersatz für die Präsentation von Web-Inhalten hinausentwickelt.
Als Teil dieser Entwicklung werden heute viele Informationen im Internet oder lokalen Netzwerken im XML Format zu Verfügung gestellt. Eine Suche über diesen Dokumenten ist zwar möglich, ist aber hinsichtlich der Struktur der Daten leider nicht in geeigneter Weise optimiert. Meist handelt es sich um reine Textsuchen, die die vom Nutzer definierten XML-Strukturen ignorieren.

Ziel des XIRCUS Projektes

Ziel des XIRCUS Projektes ist die Umsetzung einer Suchmaschine für semistrukturierte Daten (XML). Die Abkürzung XIRCUS steht für „XML-based Indexing, Ranking and Classification techniques for cUstomized Search Engines". Ein wichtiger Aspekt ist es, verschiedene Indexierungs- und Rankingmechanismen zu implementieren und zu testen. Dabei soll drei Punkten besondere Aufmerksamkeit geschenkt werden: Der Indexierung, dem Ranking und dem Testen der Suchmaschine.

Indexierung

Zur effizienten Speicherung von XML Dokumenten ist es notwendig, für verschiedene Aufgaben Zugriffsstrukturen zu unterstützen. Im einzelnen sind dies:

- Strukturindex, Pfadindex
- Werteindex
- Volltextindex

Ranking

Um ein möglichst abgestimmtes Ranking anbieten zu können, müssen verschiedene Heuristiken angewendet werden. Dazu sollen folgende Aspekte untersucht werden:

- Statistiken zu Dokumenten unter Beachtung von Stoppwortlisten: Termhäufigkeit, Term- und Dokumentanzahl, inverse Dokumenthäufigkeit
- Linguistika (Sprache): Stammwortreduktion, Synonymwörterbuch
- Strukturen: interne und externe: Titelheuristik, HITS, PageRank
- Dokumentklassen und -typen: Klassifizierung der Dokumente

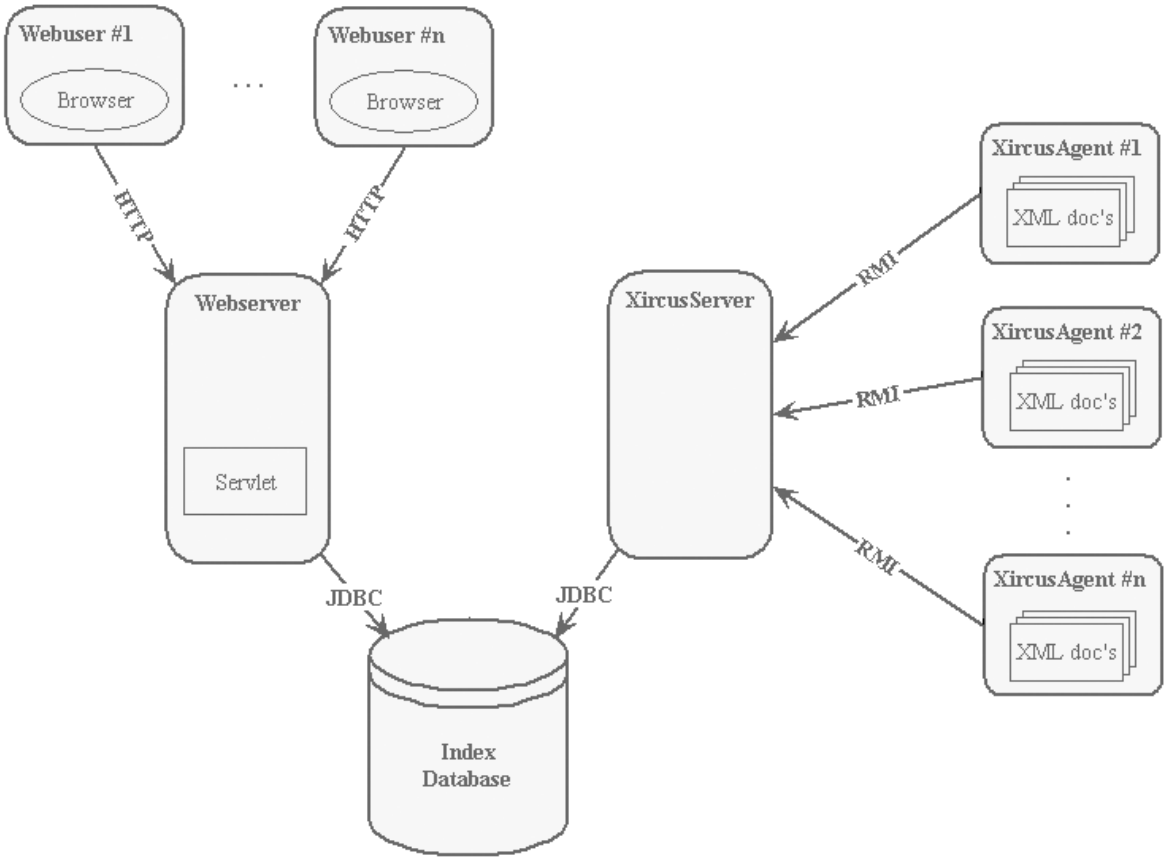
Für einzelne Techniken sollten teilweise bestehende Umsetzungen verwendet werden. Zur Anfragezeit sollen die einzelnen entwickelten Rankingfunktionen berechnet und kombiniert werden.

Testumgebung

Zum Testen steht eine XML-Kollektion zur Verfügung, die Kollektionen der XML Retrieval Evaluation Initiative.
Es soll ein Laufzeittest zur Beurteilung der Klassifikation und Bewertung der Indexstrukturen durchgeführt werden.

Umsetzung des XIRCUS Projektes

Das Projekt wurde in drei Komponenten aufgeteilt und jede Komponente von einer drei- bis vierköpfigen Gruppe umgesetzt. Diese Komponenten sind die Indexierung, der Index und die Suche inklusive Ranking. Abbildung 1 veranschaulicht das Zusammenspiel der einzelnen Komponenten.
Der Indexierung, also dem Einsammeln und Aufbereiten von XML-Inhalten, wurde ein Client-Server-Prinzip zugrunde gelegt. So gibt es mehrere Agenten, die die XML-Dokumente parsen und die gewonnen Inhalte weiter aufbereiten, in dem sie zum Beispiel Stoppwortreduzierung, Satzerkennung oder Stemming durchführen und Prüfsummen berechnen. Diese Dokumentinformationen werden dann an den Xircus-Server verschickt, der sie in den Index einfügt.



-Abbildung 1-

Der Index wird in einer relationalen Datenbank angelegt, auf die über die JDBC-API zugegriffen wird. Er stellt zwei Schnittstellen zur Verfügung: eine für die Indexierung, damit Daten eingefügt werden können und eine zweite, um auf dem Inhalt des Index suchen zu können.

Die Anfragen werden über eine Webseite gestellt. Ein Servlet bearbeitet die Anfragen, wertet die umfangreiche Anfragesprache aus und holt sich die benötigten Informationen über die Schnittstelle der Index-Komponente. Die Ergebnisse werden einer Ranking-Berechnung unterzogen, die auch auf die speziellen Wünsche des Nutzers eingeht.

Weiterentwicklung des XIRCUS-Prototypen

Basierend auf dem ersten Prototypen werden Arbeiten zur Umsetzung weiterer Indizierungs- und Rankings-Techniken von einer zweiten studentischen Gruppe ausgefuehrt. Darueber hinaus wird der Prototyp in laufende Forschungsarbeiten zur inhaltsbasierten Suche in Multimedia-Datenbanken und zur Integration von strukturierten und semi-strukturierten Anfragen und Information Retrieval eingebunden. Eine qualitative Bewertung von Xircus wird mit der INEX-Kollektion [inex02] erfolgen.

Quellen

- Fundamentals of RMI Short Course: http://developer.java.sun.com/developer/onlineTraining/rmi/RMI.html
- Apache Software Foundation: http://www.apache.org/
- Java API für Wordnet von John Didion, http://sourceforge.net/projects/jwordnet
- Jakarta Lucene: http://jakarta.apache.org/lucene/docs/index.html
- Definition des Porter Stemming Algorithmus': http://www.tartarus.org/~martin/PorterStemmer/def.txt
- RFC1321: The MD5 Message-Digest Algorithm: http://www.ietf.org/rfc/rfc1321.txt?number=1321
- Wordnet, Projekt der “University of Princeton” - “Online Lexical Reference System”: www.cogsci.princeton.edu/~wn/
- Apache Xerces 2 Parser: http://xml.apache.org/xerces2-j/index.html
- "Professional Java Server Programming J2EE Edition": Verlag WROX http://www.wrox.com
- [inex02] : http://qmir.dcs.qmw.ac.uk/INEX/