

Appendix C

Anhang gemäß Promotionsordnung

Zusammenfassung

Die Markup-Sprache XML ist ein weithin akzeptierter Standard zur Darstellung von Daten. Je mehr Daten in Form von XML-Dokumenten vorliegen, desto wichtiger werden effektive und effiziente Verfahren zur Informationssuche. Ein für die Suche sehr wichtiges Merkmal von XML-Dokumenten ist deren selbstbeschreibende Struktur. Durch Ausnutzung dieser Struktur können sehr präzise Suchanfragen formuliert werden. Doch leider kann die Dokumentstruktur sehr komplex und heterogen sein. Dem Nutzer sind oftmals nicht alle strukturellen Details bekannt, so dass die Formulierung von treffenden Anfragen schwierig ist und nicht selten relevante, aber nicht exakt zur Anfrage passende Dokumente verfehlt werden.

In dieser Arbeit stellen wir einen innovativen Ansatz zur Suche in XML-Daten vor, der die selbstbeschreibende Struktur als Leitfaden zum Auffinden der gewünschten Information verwendet. Ein Nutzer benötigt lediglich partielle Kenntnisse über die Struktur, um Anfragen zu formulieren, die Selektionsbedingungen bezüglich Inhalt und Struktur der Dokumente spezifizieren. Eine Anfrage wird derart interpretiert, dass nicht nur exakt passende Ergebnisse gefunden werden, sondern auch solche, die als ähnlich zur Anfrage eingeschätzt werden. Um diese ähnlichen Ergebnisse zu finden, wird die Anfrage mit Hilfe von Transformationssequenzen an die Struktur eines jeden Dokuments in der Kollektion angepasst. Dabei hat jede Einzeltransformation bestimmte Kosten. Die Gesamtkosten einer Sequenz werden zur Bewertung der Ähnlichkeit zwischen der originalen Anfrage und den von der transformierten Anfrage selektierten Dokumenten verwendet. Diese Kosten werden den Dokumenten zugewiesen und bestimmen deren Positionen in der Ergebnisliste, die nach fallender Ähnlichkeit sortiert ist.

Wir stellen alle notwendigen Algorithmen und Datenstrukturen für die Realisierung eines Anfrage-Prozessors vor, der Suchanfragen in polynomieller – typischerweise sogar sublinearer – Zeit bezüglich der Kollektionsgröße beantwortet. Der Anfrage-Prozessor wertet Ausführungspläne bestehend aus Operatoren aus. Die Operatoren berechnen sukzessive die Transformationskosten. Wir beschreiben Techniken zur effektiven Optimierung von Ausführungsplänen, die Äquivalenzen zwischen den Operatoren ausnutzen. Um die Ausführungszeit für eine Anfrage weiter zu verringern, schlagen wir eine Methode zum effizienten Auffinden der besten n Ergebnisse vor, die eine strukturelle Zusammenfassung der Dokumente in der Kollektion verwendet, um die besten k transformierten Anfragen zu bestimmen. Diese Anfragen werden nacheinander ausgewertet, bis die besten n Ergebnisse gefunden sind.

Eine prototypische Implementierung bestätigt die Effizienz der entwickelten Methoden zur Anfrage-Auswertung. Wir beschreiben die Architektur des Prototyps und diskutieren die Ergebnisse von systematischen Tests zur Analyse von Anfrage-Auswertungszeiten für verschiedene Kollektionen mit realen und synthetischen XML-Dokumenten.

Lebenslauf

25. 3. 1969 geboren in Frauenstein/Erzgebirge
- 1975 – 1983 Polytechnische Oberschule in Cämmerswalde
- 1983 – 1985 Polytechnische Oberschule in Neuhausen/Erzgebirge
- 1985 – 1987 Ausbildung zum Maschinenbauer im VEB Fritz Heckert in Karl-Marx-Stadt
- 1987 – 1990 Reparaturschlosser für Werkzeugmaschinen beim VEB Fritz Heckert
- 1990 – 1992 Vogtlandkolleg Plauen
- 1992 Fachgebundene Hochschulreife Technik
- 1993 – 1994 Grundwehrdienst in Leipzig
- 1992 – 1999 Studium der Informatik an der Technischen Universität Dresden und der Universität Leipzig
- 1999 Diplom in Informatik
- 1999 – 2002 Promotionsstudent an der Freien Universität Berlin und im Berlin-Brandenburger Graduiertenkolleg “Verteilte Informationssysteme”

Selbstständigkeitserklärung

Ich versichere, dass ich die vorliegende Dissertation auf Grundlage der in der Arbeit angegebenen Hilfsmittel und Hilfen selbstständig angefertigt habe.

Berlin, Dezember 2002

(Torsten Schlieder)