Parsing and Querying XML Documents in SML

Dissertation

Andreas Neumann
Fachbereich IV — Informatik
Universität Trier

Zusammenfassung

XML (Extensible Markup Language) ist ein sequentielles Format zur Speicherung und Übermittlung strukturierter Daten. Obwohl es ursprünglich für die Dokumentenverarbeitung entwickelt wurde, findet XML heute Verwendung in nahezu allen Bereichen der Datenverarbeitung, insbesondere aber im Internet.

Jede XML-Dokumentenverarbeitungs-Software basiert auf einem XML-Parser. Der Parser liest ein Dokument in XML-Syntax ein und stellt es als Dokumentbaum der eigentlichen Anwendung zur Verfügung. Dokumentenverarbeitung ist dann im wesentlichen die Manipulation von Bäumen. Moderne funktionale Programmiersprachen wie SML und HASKELL unterstützen Bäume als Basis-Datentypen und sind daher besonders gut für die Implementierung von Dokumentenverarbeitungs-Systemen geeignet. Um so erstaunlicher ist es, dass dieser Bereich zum größten Teil von JAVA-Software dominiert wird. Dies ist nicht zuletzt darauf zurückzuführen, dass noch keine vollständige Implementierung der XML-Syntax als Parser in einer funktionalen Programmiersprache vorliegt.

Eine der wichtigsten Aufgaben in der Dokumentenverarbeitung ist Querying, d.h. die Lokalisierung von Teildokumenten, die eine angegebene Strukturbedingung erfüllen und in einem bestimmten Kontext stehen. Die baumartige Auffassung von Dokumenten in XML erlaubt die Realisierung des Querying mithilfe von Techniken aus der Theorie der Baumsprachen und Baumautomaten. Allerdings müssen diese Techniken an die speziellen Anforderungen von XML angepasst werden. Eine dieser Anforderungen ist, dass auch extrem große Dokumente verarbeitet werden müssen. Deshalb sollte der Querying-Algorithmus in einem einzigen Durchlauf durch das Dokument ausführbar sein, ohne den Dokumentbaum explizit im Speicher aufbauen zu müssen.

Diese Arbeit besteht aus zwei Teilen. Der erste Teil beschreibt den XML-Parser *fxp*, der vollständig in SML programmiert wurde. Insbesondere werden die Erfahrungen mit SML diskutiert, die während der Implementierung von *fxp* gewonnen wurden. Es folgt eine Analyse des Laufzeit-Verhaltens von *fxp* und ein Vergleich mit anderen XML-Parsern, die in imperativen oder objektorientierten Programmiersprachen entwickelt wurden.

Im zweiten Teil beschreiben wir einen Algorithmus zum Querying von XML-Dokumenten, der auf der Theorie der Waldautomaten fundiert ist. Er findet alle Treffer einer Anfrage in höchstens zwei Durchläufen durch das Dokument. Für eine wichtige Teilklasse von Anfragen kann das Querying sogar in einem einzelnen Durchlauf realisiert werden. Außerdem wird die Implementierung des Algorithmus in SML mit Hilfe von *fxp* dargestellt.