Prof. Dr. Markus Lohrey
Universität Siegen
Department für Elektrotechnik und Informatik
Hölderlinstrasse 3
D-57076 Siegen
Germany
e-mail: lohrey@eti.uni-siegen.de

29. August 2014

Beschreibung der Projektgruppe
Grammatik-basierte Baumkompression

Die Idee der *Grammatik-basierten Kompression* besteht darin, ein Wort (Text) durch eine kontextfreie Grammatik, welche genau das Ausgangswort erzeugt, zu kodieren. Solch eine Grammatik ist in vielen Fällen wesentlich kleiner als das Wort, da sich wiederholende Teilmuster nur einmal abgespeichert werden müssen. In der Litertatur finden sich eine Reihe von Grammatik-basierten Kompressoren (z.B. RePair, LZ78, BISECTION, SEQUITUR, etc.), siehe z.B. [3].

Grammatik-basierte Kompression kann von Wörter auf Bäume erweitert werden. Kontextfreie Grammatiken werden dann durch *kontextfreie Baumgrammatiken* ersetzt. Diese Erweiterung wurde in [2] erstmals betrachtet. In [6] wurde *TreeRePair* eingeführt, welcher den Grammatik-basierten Wortkompressor RePair von Wörter auf Bäume verallgemeinert. Es hat sich gezeigt, dass TreeRePair (implementiert in C++, siehe `https://code.google.com/p/treerepair/`) hervorragende Resultate sowohl bezüglich Kompressionsrate, Laufzeit, und Speicherplatz erzielt.

In der Zwischenzeit wurden eine Reihe von weiteren Grammatik-basierten Baumkompressoren entwickelt. In [5] wurde *TtoG* entwickelt, und gezeigt, dass für einen Baum der Größe n TtoG eine kontextfreie Baumgrammatik berechnet, welche im schlechtesten Fall nur um einen Faktor $O(\log n)$ größer als eine optimale Baumgrammatik ist (man sagt auch, dass die Approximationsrate von ToG $O(\log n)$ beträgt). Da jedoch keine Implementierung von TtoG vorliegt, ist nicht klar, ob TtoG auch in der Praxis gute Ergebnisse erzielt und insbesondere mit TreeRePair Schritt halten kann.

In der Arbeit [4] wurde der Kompressor BISECTION von Wörter auf Bäume verallgemeinert, der resultierende Kompressor soll im folgenden mit *TreeBISECTION* bezeichnet werden. Es konnte gezeigt werden, dass TreeBISECTION für jeden Baum der Größe n eine kontextfreie Baumgrammatik der Größe $O(n/\log n)$ berechnet. Wiederum existiert bisher jedoch keine Implementierung von TreeBISECTION.

In [1] finden sich schließlich mehrere weitere Grammatik-basierter Baumkompressoren, welche auf verschiedenen Varianten von gerichteten azyklischen Graphen (sogenannten DAGs) beruhen. Implementierungen dieser Kompressoren liegen vor.

Die Aufgaben der Projektgruppe sind folgende:

- Einarbeitung in das Thema der Grammatik-basierten Baumkompression anhand der Arbeiten [1, 4, 5, 6].
- Implementierung von TtoG in C++.
- Implementierung von TreeBISECTION in C++.
- Definition und Sammeln von diversen Baumklassen mit spezifischen Charakteristika, z.B. Bäume die sich aus verschiedenen Anwendungsdomänen ergeben (XML, Termersetzungssysteme, Theorembeweiser, etc.), oder Bäume die künstlich generiert wurden (z.B. volle Binärbäume, zufällig erzeugte Bäume).
- Vergleichende Analyse der Baumkompressoren TreeRePair, TtoG, TreeBISECTION und der DAG-Varianten aus [1] mittels der im vorherigen Punkt definierten Baumklassen.
- Anfertigung einer Ausarbeitung zu den experimentellen Resultaten.

Literatur

[1] M. Bousquet-Mélou, M. Lohrey, S. Maneth, and E. Noeth. XML compression via DAGs. *Theory of Computing Systems*, 2014. DOI 10.1007/s00224-014-9544-x.

[2] G. Busatto, M. Lohrey, and S. Maneth. Efficient memory representation of XML document trees. *Information Systems*, 33(4–5):456–474, 2008.

[3] M. Charikar, E. Lehman, A. Lehman, D. Liu, R. Panigrahy, M. Prabhakaran, A. Sahai, and A. Shelat. The smallest grammar problem. *IEEE Transactions on Information Theory*, 51(7):2554–2576, 2005.

[4] D. Hucke, M. Lohrey, and E. Noeth. Constructing small tree grammars and small circuits for formulas. Technical report, arXiv.org, 2014. `http://arxiv.org/abs/1407.4286`.

[5] A. Jèz and M. Lohrey. Approximation of smallest linear tree grammars. In *Proceedings of STACS 2014*, volume 25 of *Dagstuhl Seminar Proceedings*, pages 445–457. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany, 2014. Langversion erhältlich unter `http://arxiv.org/abs/1309.4958`.

[6] M. Lohrey, S. Maneth, and R. Mennicke. XML tree structure compression using RePair. *Inf. Syst.*, 38(8):1150–1167, 2013.