08261 Executive Summary

Structure-Based Compression of Complex Massive Data

— Dagstuhl Seminar —

Stefan Böttcher[1], Markus Lohrey[2], Sebastian Maneth[3], and Wojciech Rytter[4]

[1] Universität Paderborn, Germany
stb@uni-paderborn.de
[2] Universität Leipzig, Germany
lohrey@informatik.uni-leipzig.de
[3] NICTA & University of New South Wales, Australia
sebastian.maneth@nicta.com.au
[4] Warsaw University & Copernicus University, Poland
rytter@mimuw.edu.pl

Abstract. From 22nd June to 27th of June 2008, the Dagstuhl Seminar "08261 Structure-Based Compression of Complex Massive Data" took place at the Conference and Research Center (IBFI) in Dagstuhl. 22 researchers with interests in theory and application of compression and computation on compressed structures met to present their current work and to discuss future directions.

Keywords. Compression, Succinct Data Structure, Pattern Matching, Text Search, XML Query

1 Introduction

Compression of massive complex data is a promising technique to reduce stored and transferred data volumes. In industry, semi-structured data and XML in particular, have become the de facto data exchange format, and e.g. grammar-based compression offers efficient memory representations that can be used within XML databases. Therefore, we expect improvements on compression techniques for structured data and XML to have significant impact on a variety of industry applications where data exchange or limited data storage are bottlenecks.

The big advantage of structure and grammar-based compression over other compression techniques is that many algorithms can be performed directly on the compressed structure, without prior decompression. This idea of "computing on compressed structures" has given rise to efficient algorithms in several areas; e.g., term graph rewriting, model-checking using BDDs, and querying XML; all three areas profit from the use of dags (directed acyclic graphs) which allow to share common subtrees and which can be seen as a particular instance of grammar-based compression. In these areas and others, we expect efficiency improvements through the use of more sophisticated grammar-based compression techniques.

Dagstuhl Seminar Proceedings 08261
Structure-Based Compression of Complex Massive Data
http://drops.dagstuhl.de/opus/volltexte/2008/1681

2 Seminar Goal

The goal of the seminar was to bring together researchers working on various aspects of structure-based compression and to discuss new ideas and directions for doing research on compression and on computation over compressed structures. In particular, the sub-goals were to achieve a deeper understanding of the whole field of compressing structured data, to discover new interconnections between different research directions in compression, to interchange ideas between theory and application oriented research, to distinguish between solved and open questions in the field, to identify key problems for further research, and to initiate new collaborations.

3 Seminar Organization

Within the seminar, many different forms of cooperations took place. The seminar started with a short introduction of each participant. The introduction was followed by a first series of overview talks, namely:

(1) **"Grammar-Based Compression (Survey)"** by **Wojciech Rytter** and
(2) **"Pattern Matching on Compressed Strings"**, jointly given by **Shunsuke Inenaga** and **Ayumi Shinohara**.

During the plenum discussions covering the main research areas of the seminar, it was decided to split the group into two smaller working groups.

The **Working Group I** discussed **practical aspects** of XML-compression and string compression, the **Working Group II** was concerned with theoretical questions, in particular with the complexity of algorithmic questions on compressed strings and trees. Working group discussions were interleaved with further survey presentations and discussions in the whole group. In particular, the second series of overview talks consisted of the following two presentations,

(3) **"Practical Search on Compressed Text"** by **Gonzalo Navarro** and
(4) **"XML Compression Techniques"** by **Gregory Leighton**.

In addition to the meetings of the two working groups there were many discussions in smaller groups during the breaks, the excursion, and in the evenings, which influenced the discussions in the larger groups. At the end of the seminar, the results of the working groups were presented to all participants. Additionally we had one last survey presentation.

(5) **"Algorithmics on Compressed Strings"** by **Markus Lohrey**.

and a talk by **Stefan Böttcher** on **"practical aspects of XML (stream) compression"**. At the end of the seminar, the results of the working groups were presented to all participants, and short feed-back session completed this seminar.

Working Group I
Practical Aspects of XML and String Compression

This working group started with a comparison of XML compression research done in the community of string compression and in the community of XML structure compression. **Veli Mäkinen** talked about **"Storage and Retrieval of Individual Genomes"** and explained the idea of self-indexing in a collection of sequences. **Sebastian Maneth** talked about **"Grammar-Based Tree Compression"** and explained the BPLEX approximation algorithm for finding a small tree grammar for a given tree. The participants profited a great deal from exchanging the different views to key problems like efficient XPath query evaluation. As a consequence, a deeper insight into some of the major open problems in the field was achieved, for example, how to combine XML structure-based search with string ordering criteria on BWT compressed XML constants.

Additionally, the participants benefited from the exchange of complementary solutions provided by the different communities. For example, researchers from the field of string compression contributed efficient retrieval techniques for querying BWT compressed data and wavelets and for rank-based search in tree structures, whereas researchers from the field of XML structure compression contributed application requirements and optimization and compilation techniques for XPath queries.

As a result an integrated system architecture was developed that covers both aspects, efficient path queries plus efficient retrieval of XML text values. Finally, the participants planned and organized a **large distributed software project** involving researchers in Australia, Canada, Chile, and Europe. This software project plans to integrate the ideas discussed in the working group into a new prototype implementation for the evaluation of XPath and XQuery queries on top of compressed XML and text representations. In the end of the working group meeting, **Tomasz Müldner** presented **"XSAQCT. XML Schema-Aware Queryable Compression Technique"** and explained how schema information is used to derive a small compressed representation of XML structures.

Working Group II
Algorithmic Questions on Compressed Strings and Trees

Algorithms on compressed data allow the manipulation of compressed data without prior decompression. Such algorithms have potential applications wherever large amount of data are not only stored but where these data have to be analyzed as well (e.g. in bioinformatics). Moreover, algorithms on compressed data can be used in order to avoid the explicit generation of large intermediate data structures. Using this idea, several algorithmic problems (e.g. in verification, program analysis, computational group theory) can be solved efficiently (in polynomial time).

Several algorithmic problems on compressed data were discussed in the working group. **Wojciech Fraczak** gave a presentation on **"Matching Integer Intervals by Minimal Sets of Binary Words with 'don't cares' "**

in which he discussed the problem of coding integer intervals for fast lookup tables. **Hiroshi Sakamoto** talked about a **Space-Saving Compression and Its Application to Fast Pattern Matching**, which improves the memory requirements of previous algorithms for pattern matching on compressed strings. Moreover, **Wojciech Rytter** gave a presentation on the **Complexity of the Compressed Membership Problem for Context-Free Languages**, where he sharpened known complexity results.

Among others let us emphasize in more detail the **evaluation problem of tree automata on compressed trees** which was presented by **Markus Lohrey**. Here trees are represented succinctly with context-free tree grammars and it is asked whether the decompressed tree is accepted by a given tree automaton. It was open so far, whether this problem can be solved in polynomial time. Due to the interaction of several seminar participants, we could finally obtain a polynomial time solution for this problem. Moreover, it turned out that this algorithm can be even extended to a larger class of tree automata, where subtrees can be tested for equality. A publication is in preparation. The participants plan to implement their algorithm and to provide an interface to the software project of working group I. This is of great importance, since tree automata are a fundamental concept in XML processing. For instance, large fragments of the XPath query language can be efficiently compiled into tree automata.

Other relevant problems that were discussed in the working group include the complexity of compressed membership problems for context-free languages, algorithms for updating compressed strings, and the relationship between string compression and some concepts from complexity theory (e.g. leaf languages).

4 Conclusions

The seminar brought together researchers from different fields of data compression. Many fruitful discussions provided a unique opportunity to discover interconnections between different research directions in compression, to identify open key problems for further research, and to get a much broader understanding of the field. Furthermore, a creative interchange of ideas between theory and application oriented research led to many deep insights for further research in the field. Finally, Dagstuhl initiated new collaborations which have been fruitful already (one open problem has been resolved, as mentioned in the section on Working Group II), and a large software project software project has been initiated (see the section on Working Group I). As always, Dagstuhl provided a highly enjoyable atmosphere to work in.