Veli Mäkinen, Greger Lindén, and Hannu Toivonen (Eds.)

# Summer School on Algorithmic Data Analysis (SADA 2007)

# and

# Annual Hecse Poster Session

**Abstract proceedings**

**Helsinki, Finland, May 28 - June 1, 2007**

## Contact Information

Postal address:

Department of Computer Science
P.O.Box 68 (Gustaf Hällströminkatu 2b)
FIN-00014 University of Helsinki
Finland

URL: http://www.cs.helsinki.fi

Telephone: +358 9 1911
Telefax: +358 9 191 51120

Series of Publications B, Report B-2007-4
ISSN 1458-4786
ISBN 978-952-10-3945-4
Computing Reviews (1998) Classification: A.1, E.1, E.2, E.4, F.2, H.3, I.7
Helsinki 2007
Helsinki University Printing House

# Preface

This book contains abstracts of lectures and posters presented at the Summer School on Algorithmic Data Analysis, held in Helsinki (Hotel Rantapuisto), Finland from May 28 to June 1, 2007. The summer school was the third in a series of Nordic summer schools, previously held in Norway (2005) and in Denmark (2006), organized by *NoNA* (Nordic Network on Algorithms). This year the school was organized by *Hecse* (Helsinki Graduate School in Computer Science and Engineering) in collaboration with NoNA and the *Network of Finnish Graduate Schools in Information Technology.* The Annual Hecse Poster Session was co-located at the venue.

The school was targeted at graduate students interested in algorithmic challenges raised by the demand for efficient management and analysis of various large-scale data masses. The focus was on fundamental algorithmic principles and techniques to cope with feasibility issues such as inefficiency of exact or online computation, and lack of main memory or even permanent storage space, when working on very large data sets. The school covered different algorithmic approaches motivated by these feasibility issues, including I/O-efficient algorithms and space-efficient index structures to deal with limited main memory, data stream algorithms to cope with time series data whose complete storage is not possible, and approximation algorithms for high-dimensional data. The aim was to give the participants a good overview on the current topics in algorithmic data analysis research and an in-depth introduction to a few selected subareas.

The school attracted 112 applications. Due to space limitations of the summer school venue, only 74 students could be admitted. The selections were based on the algorithmic background of the applicants with some emphasis on data analysis algorithms background and affiliation at Hecse, NoNA, and the Network of Finnish Graduate Schools in Information Technology. The Organizing Committee also decided to give preference to starting PhD students.

The school agenda consisted of minicourses/tutorials given by the 6 distinguished lecturers: Nir Ailon, Lars Arge, Paolo Ferragina, Aristides Gionis, Piotr Indyk, and S. Muthukrishnan. The agenda was supplemented with an invited talk by Juha Kärkkäinen. In addition, there were two poster sessions with overall 92 posters: one presented the posters of the summer school participants, and the other presented the posters of Hecse students in the Annual Hecse Poster Session.

We would like to thank Academy of Finland and Nordforsk (via NoNA) for the funding to make this summer school possible. The funding enabled us to offer the summer school free of charge to all students and to subsidize the accommodation cost of participants from NoNA and the Finnish network.

Helsinki, May, 2007

Veli Mäkinen
Greger Lindén
Hannu Toivonen

## Organizing Committee

Tapio Elomaa (Tampere University of Technology)

Jyrki Kivinen (University of Helsinki)

Greger Lindén (University of Helsinki), *Local Chair*

Heikki Mannila (Helsinki University of Technology)

Veli Mäkinen (University of Helsinki), *Program Chair*

Pekka Orponen (Helsinki University of Technology)

Jorma Tarhio (Helsinki University of Technology)

Hannu Toivonen (University of Helsinki), *Chair*

Esko Ukkonen (University of Helsinki)

## Lecturers and Topics

Nir Ailon (Institute of Advanced Study, Princeton):
”New algorithms for high dimensional data”

Lars Arge (University of Aarhus):
“I/O-efficient algorithms and data structures”

Paolo Ferragina (University of Pisa):
“Compressed data structures for strings”

Piotr Indyk (MIT):
“Efficient nearest neighbor search algorithms”

Aristides Gionis (Yahoo! Research, Barcelona and University of Helsinki):
“Mining the graph structures of the web”

Juha Kärkkäinen (University of Helsinki):
“Suffix array construction algorithms”

S. Muthukrishnan (Google Inc.):
“Data stream algorithms”

## Agenda

| | Monday | Tuesday | Wednesday | Thursday | Friday |
|---|---|---|---|---|---|
| 09.00-10.30 | Arge | Muthukrishnan | Ferragina | Gionis | Ferragina |
| 10.30-11.00 | *Coffee break* | *Coffee break* | *Coffee break* | *Coffee break* | *Coffee break* |
| 11.00-12.30 | Muthu. | Arge | Muthu. | Ferragina | Gionis |
| 12.30-14.00 | *Lunch break* | *Lunch break* | *Lunch break* | *Lunch break* | *Lunch break* |
| 14.00-15.30 | Indyk | Indyk | Ailon | Poster session | Kärkkäinen |
| 15.30-16.30 | *Coffee break* | *Coffee break* | *Coffee break* | *Coffee break* | |
| 16.00-17.30 | Ailon | Annual Hecse Poster Session | Excursion | Poster session | |

# Content

# Lecture abstracts

# The Fast Johnson-Lindenstrauss Transform and Applications

Nir Ailon

School of Mathematics, Institute for Advanced Study, Princeton NJ, USA
nailon@math.ias.edu
www.math.ias.edu/~nailon

**Abstract**

Dimension reduction is a highly useful tool in algorithm design, with applications in nearest neighbor searching, clustering, streaming, sketching, learning, approximation algorithms, vision and others. It removes redundancy from data and can be plugged into algorithms suffering from a "curse of dimensionality".

In my talk, I will describe a novel technique for reducing the dimension of points in Euclidean space, improving a now classic algorithm by Johnson and Lindenstrauss from the mid 80's. Our technique is the first to offer an asymptotic improvement, and has already been used in design of efficient algorithms for nearest neighbor searching and high dimensional linear algebraic numerical computations.

I will present our algorithm, its proof, applications, and interesting open questions. Joint work with Bernard Chazelle.

## 1 Introduction

Most of the talk will be based on a paper by the speaker with Bernard Chazelle [1]. Other related work [2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13].

## References

[1] N. Ailon and B. Chazelle Approximate Nearest Neighbors and the Fast Johnson-Lindenstrauss Transform *In Proceedings ofSTOC'06; To appear in SICOMP*

[2] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003.

[3] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Knowledge Discovery and Data Mining*, pages 245–250, 2001.

[4] S. DasGupta and A. Gupta. An elementary proof of the Johnson-Lindenstrauss lemma. *Technical Report, UC Berkeley*, 99-006, 1999.

[5] P. Frankl and H. Maehara. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory Series A*, 44:355–362, 1987.

[6] P. Indyk. Dimensionality reduction techniques for proximity problems. In *Proceedings of the 11th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 371–378, 2000.

[7] P. Indyk. Stable distributions, pseudorandom generators, embeddings and data stream computation. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, pages 189–197, 2000.

[8] P. Indyk. Uncertainty principles, extractors, and explicit embeddings of L2 into L1. Technical Report TR06-126, ECCC, 2006.

[9] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the 30th Annual ACM Symposium on Theory of Computing (STOC)*, pages 604–613, 1998.

[10] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[11] E. Kushilevitz, R. Ostrovsky, and Y. Rabani. Efficient search for approximate nearest neighbor in high dimensional spaces. *SIAM Journal on Computing*, 30(2):457–474, 2000.

[12] S. Muthukrishnan and S. C. Sahinalp. Simple and practical sequence nearest neighbors with block operations. In *Proceedings of the 13th Annual Symposium on Combinatorial Pattern Matching (CPM)*, pages 262–278, 2002.

[13] T. Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, Berkeley, CA, 2006.

# Aggregating Discrete Information from Inconsistent Sources

Nir Ailon

School of Mathematics, Institute for Advanced Study, Princeton NJ, USA
nailon@math.ias.edu
www.math.ias.edu/˜nailon

**Abstract**

An output of a meta-search engine is to be computed from outputs of several search engines. Microarray data is to be clustered based on outputs of several clustering heuristics. A ranking of hotels is to be compiled from qualitative information (”number of stars” rating) collected from several hotel critics.

More generally: How do we combine possibly contradicting discrete information from different sources into one consistent output? These questions, lying in the intersection of combinatorial optimization and social choice theory, have enjoyed a recent flurry of activity in theoretical and experimental computer science research. I will survey recent work and present new and improved approximation algorithms.

Based on joint work with Moses Charikar and Alantha Newman.

## 1 Introduction

The talk will be mostly based on the papers [1, 2]. Other related papers are [3, 4, 5, 6, 7, 8].

## References

[1] N. Ailon. Aggregation of partial rankings, p-ratings and top-m lists. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2007.

[2] N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: ranking and clustering. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing (STOC)*, pages 684–693, 2005.

[3] J. Aslam and M. Montague. Condorcet fusion for improved retrieval. In *Proceedings of the 11th International Conference on Information and Knowledge Management*, pages 538–548, 2002.

[4] D. Coppersmith, L. Fleischer, and A. Rudra. Ordering by weighted number of wins gives a good ranking for weighted tournamnets. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2006.

[5] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation revisited. *Manuscript*, 2001.

[6] R. Fagin, R. Kumar, M. Mahdian, D. Sivakumar, and E. Vee. Comparing and aggregating rankings with ties. In *Proceedings of the ACM Symposium on Principles of Database Systems (PODS)*, pages 47–58, 2004.

[7] V. Filkov and S. Skiena. Integrating microarray data by consensus clustering. In *Proceedings of International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 418–425, Sacramento, 2003.

[8] A. Gionis, H. Mannila, and P. Tsaparas. Clustering aggregation. In *Proceedings of the 21st International Conference on Data Engineering (ICDE)*, Tokyo, 2005.

# I/O-efficient Algorithms and Data Structures

Lars Arge

MADALGO[*]
Department of Computer Science
University of Aarhus
Aarhus, Denmark

large@madalgo.au.dk
www.madalgo.au.dk/∼large

**Abstract**

In many modern applications that deal with massive data sets, communication between internal and external memory, and not actual computation time, is the bottleneck in the computation. This is due to the huge difference in access time of fast internal memory and slower external memory such as disks. In order to amortize this time over a large amount of data, disks typically read or write large blocks of contiguous data at once. This means that it is important to design algorithms with a high degree of locality in their disk access pattern, that is, algorithms where data accessed close in time is also stored close on disk. Such algorithms take advantage of block transfers by amortizing the large access time over a large number of accesses. In the area of I/O-efficient algorithms the main goal is to develop algorithms that minimize the number of block transfers (I/Os) used to solve a given problem.

These lectures will cover I/O-efficient algorithms and data structures for a few fundamental problems in computational geometry and graph theory, with focus on the techniques used to design such algorithms. The problems include sorting and fundamental searching problems, as well as the list ranking graph problem. Most of the covered material is described in [1, 2, 3].

## References

[1] A. Aggarwal and J. S. Vitter. The Input/Output complexity of sorting and related problems. *Communications of the ACM*, 31(9):1116–1127, 1988.

[2] L. Arge. External-memory geometric data structures. Lecture notes (2005) available at http://www.daimi.au.dk/∼large/ioS06/ionotes.pdf

[3] N. Zeh. I/O-efficient graph algorithms. Lecture notes (2002) from EEF Summer School on Massive Datasets available at
http://users.cs.dal.ca/∼nzeh/Teaching/Summer2007/6104/Notes/zeh02.pdf

[*] Center for Massive Data Algorithmics — a center of the Danish National Research Foundation.

# Compressed data structures for strings

Paolo Ferragina

Dipartimento di Informatica, Università di Pisa, Italy
ferragina@di.unipi.it
www.di.unipi.it/~ferragin

### Abstract

I have seen too many papers that, when dealing with large sets of long strings, argue that one must choose between the *time* efficiency of some (more or less sophisticated) search operations and the *space* succinctness of their solutions. In my lectures I address this *"old" dichotomy* by discussing the most recent advances in compressing and indexing strings. The moral is that: theoretically, it is no longer the case that such a dichotomy does exist; practically, do exist some engineered implementations that ease and stimulate the use of these sophisticated and effective tools in real applications.

## On the Lectures

The string matching field has grown to such a complicated stage that various issues come into play when studying it: data structure and algorithmic design, compression techniques, architectural features, database principles, algorithmic engineering and experimentation. My lectures concentrate on the currently well-studied interplay that does exist between two research fields: indexing data structures and compressor design. This connection, at a first glance, might appear paradoxical because these "tools" have antithetical goals. In fact, index design aims at augmenting data with routing information (i.e. data structures) that allow the efficient retrieval of patterns or the extraction of some information. Conversely, compressors aim at removing the repetitiveness present in the data to squeeze them in a reduced space occupancy. Recent results have shed new light on these two fascinating topics by surprisingly showing that compressed indexes and strong compressors do exist, and they can be designed by carefully orchestrating known, and novel, ideas born in both these two research fields. This is actually an active area of research that, apart of interesting solutions to many individual problems, lead to a foundational contribution: several indexing and compression problems *can be reduced to* the design of some surprisingly simple basic tools; improving these tools immediately leads to *guaranteed* time and space improvements for the more sophisticated problems. This algorithmic framework has a twofold advantage: theoretically, it allows researchers to study the simpler problems in order to design efficient solutions and/or derive computational limitations for them; practically, it allows programmers to build efficient compressed indexes starting from engineered implementations of those basic blocks (see e.g. the Pizza&Chili site [5]).

To highlight these interesting algorithmic issues, I deal with strings of various types— binary or from a general alphabet, raw or with some structure (e.g. XML)— and with the design of several kinds of query operations— from the classical substring/prefix/suffix searches to more sophisticated operations which involve string content and structure. The following bibliography provides few seeds to start digging into this fascinating topic.

## References

[1] J. Barbay and I. Munro. Succinct Encoding of Permutations and its Applications to Text Indexing. *Encyclopedia of Algorithms*, Editor in Chief Ming-Yang Kao, Springer, 2007 (to appear).

[2] P. Ferragina, R. Giancarlo, G. Manzini, M. Sciortino. Compression boosting in optimal linear time. *Journal of the ACM*, 52(4):688-713, 2005.

[3] P. Ferragina, R. Giancarlo, G. Manzini. The myriad virtues of wavelet trees. *International Colloquium on Automata, Languages and Programming*, LNCS vol. 4051, 561–572, 2006.

[4] P. Ferragina, F. Luccio, G. Manzini, S. Muthukrishnan. Structuring labeled trees for optimal succinctness, and beyond. *IEEE Symposium on Foundations of Computer Science*, 184–196, 2005. Extended version downloadable from `http://roquefort.di.unipi.it/∼ferrax/xml_ferra.pdf`

[5] P. Ferragina and G. Navarro. The Pizza&Chili site. Two mirrors at `http://pizzachili.di.unipi.it` and `http://pizzachili.dcc.uchile.cl`.

[6] P. Ferragina, R. Venturini. A simple storage scheme for strings achieving entropy bounds. *Theoretical Computer Science*, 372(1): 115-121, 2007.

[7] P. Ferragina, R. Venturini. Compressed permuterm index. *ACM SIGIR Conference*, 2007 (to appear).

[8] A. Gupta, W.K. Hon, R. Shah, J. Vitter. Compressed data structures: dictionaries and data-aware measures. *IEEE Data Compression Conference*, 213-222, 2006.

[9] G. Navarro and V. Mäkinen. Compressed full text indexes. *ACM Computing Surveys*, 39(1), 2007.

[10] N. Raman and R. Raman. Rank and select operations on binary strings. *Encyclopedia of Algorithms*, Editor in Chief Ming-Yang Kao, Springer, 2007 (to appear).

# Mining the graph structures of the web

Aristides Gionis

Yahoo! Research, Barcelona, Spain, and
University of Helsinki, Finland
gionis@yahoo-inc.com

**Abstract**

Graph structures is a general way of modeling entities and their relationships and they are used to describe a wide variety of data including the Internet, the Web, social networks, metabolic networks, protein-interaction networks, food webs, networks of citations among papers, and many more. In the recent years there has been an increasing amount of literature on studying properties, models, and algorithms for graph data. The first part of the seminar gives a brief overview of graph-generation models and graph-mining algorithms. The set of topics includes algorithms for discovering communities, models for characterizing the evolution of graphs over time, as well as discussion on their ubiquitous scale-free properties. In the second part we discuss applications where exploiting the graph structure is beneficial for certain data-mining tasks and we present challenges of graph mining in the context of problems appearing in a search engine.

## 1 Background

One of the most pervasive properties of real-world graphs is the emergence of power laws that seems to characterize many of their statistical properties [1, 5]. Power laws have intrigued the interest of researchers and many models that attempt to explain their presence in real graphs have been proposed, e.g., see [1, 4, 11]. In the first part of the seminar we discuss properties of power-law distributions and describe underlying processes that generate such distributions [14, 15, 17].

We then discuss the problem of finding communities in graphs, which is related to the problem of graph clustering. We give an overview of objective functions used for the task of finding communities and we review a number of combinatorial and spectral algorithms [7, 8, 16]. Many of the clustering methods are prohibitively expensive for applying them on large-scale graphs, so we also discuss scalable algorithms that have been designed for finding communities on the Web [12].

Finally we review studies on statistical properties of graphs that evolve over time [10, 13].

## 2 Applications

In the second part of the seminar we present specific applications of graph mining in the context of problems appearing in search engines.

The first application is spam detection [2]. A common approach to detecting spam is to extract a set of content-based and link-based features from Web pages and treat the spam-detection problem as a classification problem. In addition to extracting discriminative features, one can exploit the observation that linked hosts tend to belong to the same class: either both are spam or both are non-spam. We discuss different algorithms that attempt to leverage this observation and exploit the topology of the web graph in order to improve the accuracy of a baseline feature-based spam-detection system.

Then we discuss the problem of predicting the popularity of items in a dynamic environment in which authors post new items and provide feedback on existing ones [3]. The basic setting can be applied to predict popularity of blog posts, rank photographs in a photo-sharing system, or predict the citations of a scientific article using author information and monitoring the item of interest for a short period of time after its creation. One of the components of the system is the eigenrumor algorithm [6], an adaptation of the HITS algorithm [9].

We conclude by describing complex graph structures that emerge in problems related to search engines and we discuss challenges on mining those graphs.

## References

[1] A.-L. Barabasi, R Albert. *Emergence of Scaling in Random Networks.* Science, 286, 1999.

[2] C. Castillo, D. Donato, A. Gionis, V. Murdock, F. Silvestri. *Know your Neighbors: Web Spam Detection using the Web Topology.* 30th Annual International ACM SIGIR Conference, 2007.

[3] C. Castillo, D. Donato, A. Gionis. *Estimating the number of citations of a paper using author reputation.* Submitted for publication.

[4] A. Fabrikant, E. Koutsoupias, C. Papadimitriou. *Heuristically Optimized Trade-offs: A New Paradigm for Power Laws in the Internet.* 29th International Colloquium on Automata, Languages and Programming (ICALP), 2002.

[5] M. Faloutsos, P. Faloutsos, C. Faloutsos. *On Power-Law Relationships of the Internet Topology.* ACM SIGCOMM, 1999.

[6] K. Fujimura, N. Tanimoto. *The EigenRumor algorithm for calculating contributions in cyberspace communities.* Trusting Agents for Trusting Electronic Societies, 2005.

[7] J. Hopcroft, O. Khan, B. Kulis, B. Selman. *Natural communities in large linked networks.* 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.

[8] G. Karypis, V. Kumar. *A fast and high quality multilevel scheme for partitioning irregular graphs.* SIAM Journal on Scientific Computing, 20(1), 1999.

[9] J. Kleinberg. *Authoritative sources in a hyperlinked environment.* Journal of the ACM, 46, 1999.

[10] R. Kumar, J. Novak, P. Raghavan, A. Tomkins. *On the bursty evolution of Blog Space.* 12th International World Wide Web Conference, 2003.

[11] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, E. Upfal. *Stochastic models for the Web graph.* 41th IEEE Symposium on Foundations of Computer Science, 2000.

[12] R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins. *Trawling the web for emerging cyber-communities.* 8th International World Wide Web Conference, 1999.

[13] J. Leskovec, J. Kleinberg, C. Faloutsos, *Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations.* International Conference on Knowledge Discovery and Data Mining, 2005.

[14] L. Li, D. Alderson, J. Doyle, W. Willinger. *Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications.* Internet Mathematics, 2006.

[15] M. Mitzenmacher. *A Brief History of Generative Models for Power Law and Lognormal Distributions.* Internet Mathematics, 2004.

[16] M. E. J. Newman. *Power laws, Pareto distributions and Zipf's law.* Contemporary Physics, 46(5), 2005.

[17] M. E. J. Newman, M. Girvan. *Finding and evaluating community structure in networks.* Physical Review E, 2004.

# Data Stream Algorithms

S. Muthukrishnan

Google Inc
muthu@google.com
www.cs.rutgers.edu/~muthu

**Abstract**

How does one deal with massive data sets that is available for analyses? We will describe the classical *data stream model* in which we make one pass over the data and with sublinear resources perform much of the data analyses we care about, such as frequent items, summaries, compressed sensing, clustering and others. We will present the basic algorithmic techniques used to build the sophisticated analyses above. In addition, we will present extensions to other problems (graph, matrix, statistics) and to other models (probabilistic models, parallel models such as Google's MapReduce). One of the reasons this area of research thrives is its immediate application to a number of scenarios, which we will describe.

## 1 Introduction

See [1].

## References

[1] S. Muthukrishnan. Data Streams: Algorithms and Applications. *Foundations and Trends in Theoretical Computer Science*, Vol 1, Issue 2, August 2005.

# Poster abstracts of SADA 2007 participants

# Tuning adaptive binary search trees for working sets

Timo Aho

Institute of Software Systems, Tampere University of Technology, Finland
timo.aho@tut.fi
www.cs.tut.fi/~aho26

## ABSTRACT

With binary search trees we are traditionally interested in the length of the longest possible search path in the tree. Also the path to specific elements may be optimized for their assumed access frequency. However, in practice we do not often know all the characteristics of the access sequence or the characteristics may vary between different parts of it. Thus a single static binary search tree may not be an optimal solution. If we notice this we are considering binary search trees as online algorithms [1]. For a solution to this online problem adaptive binary search trees like splay tree [4] have been designed. Splay tree always raises the accessed element near the root. Thus the last accessed elements are always near the root and efficiently accessible.

The problem with the splay trees is the amount of executed expensive splay operations. The number of splay operations has been reduced by probabilistic approaches [2] with some success. However, there is also another alternative: Access requests to keys stored into a data structure often exhibit locality of reference in practice. Such a regularity can be modeled, e.g., by working sets [3]. Optimal adaptive data structures should trace the current working set and keep it efficiently accessible. However, splay tree raises current working set near the root without any monitoring. We have studied how the existence of working sets could benefit us on splay trees. We are especially interested in how much we can win if we use methods of machine learning to monitor the current working set.

The work is done jointly with Tapio Elomaa and Jussi Kujala.

## References

[1] Susanne Albers. Online algorithms: A survey. *Mathematical Programming*, 97(1-2):3–26, 2003.

[2] Susanne Albers and Marek Karpinski. Randomized splay trees: Theoretical and experimental results. *Information Processing Letters*, 81(4), 213–221, 2002.

[3] Peter J. Denning. Working Sets Past and Present. *IEEE Transactions on Software Engineering*, 6(1):64–84, 1980.

[4] Daniel Dominic Sleator and Robert Endre Tarjan. Self-adjusting binary search trees. *Journal of the ACM*, 32(3):652–686, 1985.

# Breadth first search on massive graphs

Deepak Ajwani

Max-Planck-Institut für Informatik, Saarbrücken, Germany
ajwani@mpi-inf.mpg.de
www.mpi-inf.mpg.de/~ajwani

## ABSTRACT

Breadth first search is a fundamental graph traversal strategy and an archetype for many important graph problems. Despite the existence of simple $O(n+m)$ algorithms in the standard RAM model, it was considered non-viable till recently because of the large number of I/Os it incurs. Munagala and Ranade [1] proposed an algorithm for computing a BFS level decomposition of undirected graphs in external memory incurring $O(n+\mathrm{sort}(m))$ I/Os. Mehlhorn and Meyer [2] gave the first $o(n+m)$ I/O algorithm, thereby improving the upper bound for sparse graphs. With our STXXL based implementations of these algorithms exploiting pipelining and disk parallelism, we were able to compute the BFS level decomposition of a web-crawl based graph of around 130 million nodes and 1.4 billion edges in less than 4 hours using a single disk and 2.3 hours using 4 disks. Coupled with a heuristic proposed in [4], it reduces the time for BFS on many different families of undirected graphs from *months* to *hours*.

The implementation was first presented in [3] and the later improvements were shown in [4].

This is a joint work with Roman Dementiev, Ulrich Meyer and Vitaly Osipov.

## References

[1] K. Munagala and A. Ranade. I/O-complexity of graph algorithms. Symposium On Discrete Algorithms (SODA), 687–694, 1999.

[2] K. Mehlhorn and U. Meyer. External-Memory Breadth-First Search with Sublinear I/O. European Symposium on Algorithms (ESA), LNCS - 2461, 723–735, 2002.

[3] D. Ajwani, R. Dementiev and U. Meyer. A Computational Study of External-Memory BFS Algorithms. Symposium On Discrete Algorithms (SODA), 601–610, 2006.

[4] D. Ajwani, U. Meyer and V. Osipov. Improved external memory BFS implementations. Workshop on Algorithm engineering and experiments (ALENEX), 2007.

# Efficient Time-Travel Search over Web Archives

Klaus Berberich

Max-Planck Institute for Informatics, Saarbrücken, Germany
kberberi@mpi-inf.mpg.de
www.mpi-inf.mpg.de/~kberberi

## ABSTRACT

The Web's evolution is a mirror of our history and thus an important information source that is worth preserving. Web archives have taken responsibility for the preservation part; the access part, however, has been largely neglected. In particular, a time-travel search functionality, to evaluate a keyword query ”as of” a user-specified temporal context, is missing. Besides providing convenient end-user access to archived contents, time-travel search can support or even enable text-mining tasks along the time axis.

Time-travel search is a challenging research topic, as the data volumes of even small web archives are easily at the order of terabytes. Naïve approaches that simply apply existing techniques from web search, without paying special attention to temporal aspects of the collection, do not scale well and therefore fail.

This poster illustrates our approach to time-travel search consisting of two major building blocks: i) Score synopses [1] as an efficient solution for the management of time-varying collection statistics (e.g., PageRank and IDF scores). In this component, time-varying collection statistics are viewed as time series, and a more compact representation is obtained through their piecewise approximation. ii) TTIX, our Time-travel Text IndeX, that provides a scalable index for temporally versioned document collections. TTIX initially achieves a compact representation of the indexed collection without sacrificing result quality (see [2] for details). In a second step, it allows to trade off space consumption and query performance and thus to tune the index with regard to predefined space constraints or performance guarantees. The two presented components are implemented in our prototype system FluxCapacitor that we employed to enable time-travel search across two collections, namely the revision history of the English Wikipedia and a subset of the Internet Archive (a well-known web archive). Experiments using this implementation demonstrated the effectiveness and efficiency of our overall approach.

## References

[1] Klaus Berberich, Srikanta J. Bedathur, and Gerhard Weikum. Rank Synopses for Efficient Time Travel on the Web Graph. In *CIKM*, pages 864–865, 2006.

[2] Klaus Berberich, Srikanta J. Bedathur, and Gerhard Weikum. Efficient Time-Travel on Versioned Text Collections. In *BTW*, pages 44–63, 2007.

# Towards Constraint-Based Subgraph Mining

Michele Berlingerio

IMT Lucca, ISTI-CNR Pisa, Italy
michele.berlingerio@isti.cnr.it

## ABSTRACT

The explosive growth of automated collected data has opened the possibility of extracting useful information and knowledge from the data. Data mining helps in discovering non-trivial patterns on a stored data and hence more research is being done in this field to extract useful information from large amounts of collected data. Most often the data of interest is very complex. It is common to model complex data with the help of graphs consisting of nodes and edges that are often labeled to store additional information. Having a graph database, it is interesting to find common graphs in it, connections between different graphs and graphs that are subgraphs of a certain number of entries. Thus, many algorithms are being developed for graph mining to discover interesting patterns on data having inherent structural relationship. Representation of complex relationships in data can be readily done using graphs and since graph mining makes use of structural relationship to discover interesting patterns, graph mining has emerged as an appropriate solution to mine over data that can be represented as graphs. This graph-based data mining has become more and more popular in the last few years. It has a broad range of applications. Examples are the analysis of XML documents, CAD circuits, weblogs, the mining for social networks, interesting structures in molecular biology and the workflow mining. In the last few years, there have been many studies on efficient and effective frequent graph mining. Algorithms such as AGM , FSG [1], gSpan [2], Gaston and ADI-Mine [3] have been presented for improving scalability on mining subgraphs one after one. However, the extraction is not always tractable for the user defied constraints. The motivations supporting the need for constraint-based pattern discovery (i.e., effiency and focus on interesting knowledge) are strong when dealing with graphs. In fact, it is well known that subgraph isomorphism (i.e., deciding whether a graph is subgraph of another one) is a NP-complete problem: as a consequence, subgraph miners are exponential in runtime and memory consumption. The constraint-based paradigm could play an important role in enhancing frequent subgraph pattern discovery effectiveness and usability.

## References

[1] M.Kuramochi, G.Karypis. Frequent Subgraph Discovery *Proc. of the 1st IEEE Int. Conf. on Data Mining* (ICDM'01)

[2] X.Yan, J.Han. gSpan: Graph-based substructure pattern mining *Proc. of the 2nd IEEE Int. Conf. on Data Mining* (ICDM'02)

[3] C.Wang, W.Wang, J.Pei, Y.Zhu, B.Shi. Scalable mining of large disk-based graph databases *Proc. of the 10th ACM Int. Conf. on Knowledge Discovery and Data Mining* (SIGKDD'04)

# New Algorithms for Regular Expression Matching

Philip Bille

IT University of Copenhagen, Denmark
beetle@itu.dk
http://www.itu.dk/people/beetle/

## ABSTRACT

We revisit the classical regular expression matching problem, namely, given a regular expression $R$ and a string $Q$, decide if $Q$ matches one of the strings specified by $R$. Let $m$ and $n$ be the length of $R$ and $Q$, respectively. On a standard unit-cost RAM with word length $w \geq \log n$, we show that the problem can be solved in $O(m)$ space with the following running times:

$$\begin{cases} O(n\frac{m \log w}{w} + m \log w) & \text{if } m > w \\ O(n \log m + m \log m) & \text{if } \sqrt{w} < m \leq w \\ O(\min(n + m^2, n \log m + m \log m)) & \text{if } m \leq \sqrt{w}. \end{cases}$$

This improves the best known time bound among algorithms using $O(m)$ space. Whenever $w \geq \log^2 n$ it improves all known time bounds regardless of how much space is used.

An extended abstract of this work appeared in Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP 2006). A draft of the full version is availiable as arxiv preprint cs.DS/0606116. Both can be found at the above homepage.

# Adaptive thinning of weather observations

Vladimir Bondarenko

Department of Computer and Information Science, University of Constance, Germany
vladimir.bondarenko@inf.uni-konstanz.de
www.inf.uni-konstanz.de/gk/people/member/bondarenko.html

## ABSTRACT

Numerical weather prediction (NWP) is akin to solving an initial value problem: a weather forecast is computed as a time evolution of an estimated initial weather state. The forecast error is to a large extent due to the miss-specification of the initial state estimate. To reduce the estimation error, a precise information about the atmospheric state is required. The modern meteorological observation systems, in particular satellite instruments, produce ever increasing amounts of weather measurements with high spatial and temporal density. Nowadays, the extremely large size of the measurement sets becomes prohibitive for their complete incorporation into a weather forecast model. Therefore, a preprocessing routine, called observation thinning, is commonly applied at many of the NWP centers to reduce the size of the observational data sets prior to the state estimation.

Observation thinning is a procedure that maps a complete observation set onto one of its subsets. Most of the NWP centers nowadays apply a simple thinning strategy resulting in uniform spatial distribution of thinned observations regardless of a given meteorological situation. In contrast to this non-adaptive method, the adaptive observation-thinning algorithms attempt to identify regions characterized by a sensitive meteorological state and adapt the spatio-temporal distribution of thinned observations accordingly. Two such methods are presented here: a cluster-based thinning scheme that is basically a multi-dimensional clustering algorithm and an Estimation-Error-Analysis algorithm that makes use of the leave-one-out cross-validation technique. A test study of these algorithms in the operational NWP model of the National German Weather Service yielded controversial results [2]. Therefore a simplified experimental NWP-framework was developed to gain a deeper insight into the properties of the thinning algorithms with respect to both estimation and forecast errors [1].

## References

[1] Bondarenko, V., Ochotta, T., Saupe, D., Wergen, W. The interaction between model resolution, observation resolution and observation density in data assimilation: a two-dimensional study. *Proceedings 11th Symposium on Integrated Observing and Assimilation Systems for the Atmosphere, Oceans, and Land Surface*, San Antonio, Texas, 2007.

[2] Ochotta, T., Gebhardt, C., Saupe, D., Wergen, W. Adaptive thinning of atmospheric observations in data assimilation with vector quantization and filtering methods. *Quarterly Journal of the Royal Meteorological Society*, Vol. 131, pp. 3427-3437, Oct. 2005.

# Data stream algorithms for mining frequent graph minors in large scale complex networks

Ilaria Bordino

Department of Computer and Systems Science
*La Sapienza*University of Rome,Italy
bordino@dis.uniroma1.it

## ABSTRACT

We present a class of random sampling algorithms that with probability at least $(1-\delta)$ compute a $(1 \pm \epsilon)$ approximation of the number of several small subgraphs in a graph given as the stream of its edges.We have algorithms for the 13 connected subgraphs of three nodes and some minors of four nodes in a directed graph, and for the 6 connected subgraphs of four nodes in an undirected graph.For each considered minor we have tried different sampling strategies in order to figure out the one that in each particular case let us gain the best quality in approximation.Our algorithms use space that is inversely related to the ratio between the number of occorrunces of the specific pattern we look for and the number of structures that we choose to sample, that can be the paths of length two or the couples of edges or the $K_{13}$s occurring in the graph.Since the space complexity depends only on the structure of the input graph and not on the number of nodes, our algorithms scale very well with increasing graph size,so they provide a basic tool to analyze the structure of large graphs. We have implemented the algorithms and evaluated their performance on networks from several different application domains:the sizes of the considered input graphs varied from about 5,000 nodes and 50,000 edges to about 40 million nodes and 1 billion edges. For all the algorithms we run experiments with a sample set size equal to 10,000 100,000 1,000,000 , to evaluate running time and approximation guarantee. The algorithms appear to be time efficient for these sample set sizes.The problem of counting minors has many applications, for example in discoverying web communities, analyzing the structure of large networks and mining the most frequent interconnection patterns in a graph.We have used our algorithms to classify the structure of complex networks by observing the frequency of occurrence of graph minors of small size : we have taken into consideration both real networks coming from several different domains and synthetic graphs generated according to some of the classic network models proposed for the Webgraph (copying model,evolving network model).

Full version of this paper appears in [1]. Joint work with Luciana Salete Buriol,Debora Donato and Stefano Leonardi.

## References

[1] Ilaria Bordino . Data stream algorithms for mining frquent graph minors in large scale complex networks . *Journal of Algorithmic Data Analysis*, 1(1):1–20, 2007.

# Predictions in Graph Databases

Björn Bringmann

Department of Computer Science, Katholieke Universiteit Leuven, Belgium
Bjoern.Bringmann@cs.kuleuven.be
http://www.kuleuven.be/cv/u0053117.htm

## ABSTRACT

Prediction problems occur in many applications in chemistry, biology, social sciences and so on where data is usually represented as graphs or network. Algorithms searching interesting subgraphs useful for prediction make extensive use of subgraph isomorphism. Unfortunately subgraph isomorphism is known to be NP-complete. Hence, there is a need to optimize the search strategy in order to reduce the number of subgraph isomorphism problems.

We will present an approach to exploit relationships among pattern languages in order to find patterns with high predictive power efficiently. The resulting features can be used for most machine learning algorithms or directly with approaches such as [2] or [3].

This work is partially presented in [1],

## References

[1] Björn Bringmann, Albrecht Zimmermann, Luc De Raedt, and Siegfried Nijssen Don't Be Afraid of Simpler Patterns *Proceedings of the 15th European Conference on Machine Learning (ECML 2006)*, 2006.

[2] Mohammed J. Zaki, Charu C. Aggarwal XRules: An Effective Structural Classifier for XML Data *Machine Learning Journal special issue on Statistical Relational Learning and Multi-Relational Data Mining* Vol. 62, No. 1-2, pp. 137-170, Feb. 2006.

[3] Bjrn Bringmann and Albrecht Zimmermann. Tree$^2$ - Decision Trees for Tree Structured Data. *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2005)*, 2005

# Stochastic Joint Optimization of Multiple Images to 3D Model Registration

Ioan Cleju

Department of Computer and Information Science, University of Konstanz, Germany
cleju@inf.uni-konstanz.de
www.inf.uni-konstanz.de/~cleju

## ABSTRACT

We study the problem of registering multiple images to a 3D model by estimating the camera parameters for each image. In [1] we introduced a framework based on mutual information [2] for joint registration of several images to a 3D model. We chose the mutual information as the objective function because it is robust and it does not require segmentation and feature extraction. Since many images are needed to cover the whole surface of the object, we extended the registration framework from [2] to consider not only the geometry, but also the color of the model. We define the mutual information between images by using the color information of the images that contain common parts of the object. The assumption is that the optimal camera parameters maximize all image-to-model and image-to-image mutual information functions.

Stochastic gradients of the objective functions are estimated from small subsamples of data. The update directions are computed from the stochastic gradients, and the parameters are updated iteratively, similarly to the gradient descent method. As for each set of camera parameters we obtain several gradients corresponding to the optimization functions, we propose a probabilistic solution and an heuristic implementation for the gradients fusion.

We showed in several experiments with synthetic models that our algorithm is robust to illumination conditions and to varying surface characteristics. Restricted to the same computational load, the joint optimization algorithm was always better than the independent optimization of image-to-model alignment. We applied the algorithm with good results for an object with highly specular surface and photographs acquired in moderate lighting environment.

## References

[1] I. Cleju and D. Saupe. Stochastic Optimization of Multiple Texture Registration Using Mutual Information. *Submitted*, 2007.

[2] P. Viola and W. M. Wells III. Alignment by Maximization of Mutual Information. *International Journal of Computer Vision*, 24(2):137–154, 1997.

# Dynamic Multi-level Overlay Graphs for Shortest Paths

Gianlorenzo D'Angelo

Department of Electrical and Information Engineering, University of L'Aquila, Italy
gdangelo@ing.univaq.it
www.diel.univaq.it/people/dangelo/

## ABSTRACT

Multi-level overlay graphs represent a speed-up technique for shortest paths computation which is based on a hierarchical decomposition of a weighted directed graph $G$. They have been introduced in [1] and shown to be experimentally efficient, especially when applied to real-world large graphs. Given a weighted directed graph $G = (V, E)$, with $n$ nodes and $m$ edges, and a sequence $S_0, S_1, \ldots, S_l$ of subsets of $V$ such that $V \equiv S_0 \supset S_1 \supset S_2 \supset \ldots \supset S_l$, a *multi-level overlay graph* is defined as $\mathcal{M} = (V, E \cup E_1 \cup E_2 \cup \ldots \cup E_l)$, where $E_i$, $1 \leq i \leq l$, is a set containing the so called *i-level edges*, which are additional edges determined by shortest paths among nodes in $S_i$. When a $s$-$t$ distance query is asked, $\mathcal{M}$ allows to build a graph $\mathcal{M}_{st} = (V_{st}, E_{st})$ which is much smaller than $G$, and such that the distance from $s$ to $t$ in $\mathcal{M}_{st}$ is the same in $G$. Thus, an $s$-$t$-query can be answered faster in $\mathcal{M}_{st}$ than in $G$. However, in the literature, no theoretical result on the cost of constructing, maintaining and querying multi-level overlay graphs in a dynamic environment is known. We propose a theoretical study that lead us to the definition of a new data structure for the computation of a multi-level overlay graph of a given graph. This new data structure can be also dynamized for the maintenance of multi-level overlay graph while *weight decrease* or *weight increase* operations are performed on the original graph. In particular, the contribution is twofold: (1) we show theoretical properties of the multi-level overlay graphs that allow us to: store the information on $\mathcal{M}$ in a data structure requiring $O(n+m+|\bigcup_{i=1}^{l} E_i|)$ space; compute $\mathcal{M}$ in $O(|S_1|(m+n\log n))$ worst case time; answer $s$-$t$ distance queries in $O(m+|V_{st}|\log|V_{st}|)$, $|V_{st}| < n$. (2) We propose a new data structure for the dynamic maintenance of $\mathcal{M}$ requiring the additional storage of $|S_1|$ shortest paths trees. We show that, if either a weight decrease or a weight increase operation occurs on an edge of $G$, to update $\mathcal{M}$, it is sufficient to update the stored $|S_1|$ shortest paths trees. We propose a dynamic algorithm that requires $O((n+m)|S_1|)$ space, $O(|S_1|(m+n)\log n)$ preprocessing time, and $O(|S_1|n+m+\Delta\sqrt{m}\log n)$ worst case time to deal with a modification, by using the fully dynamic algorithm in [2]. Here, $\Delta$ is the number of pairs in $S_1 \times V$ that change either the distance or the shortest path as a consequence of a modification.
Joint work with Serafino Cicerone, Gabriele Di Stefano, and Daniele Frigioni.

## References

[1] M. Holzer, F. Schulz, and D. Wagner. Engineering multi-level overlay graphs for shortest-path queries. *Proceedings in Applied Mathematics*, 129:156–170. SIAM, 2006.

[2] D. Frigioni, A. Marchetti-Spaccamela, and U. Nanni. Fully dynamic algorithms for maintaining shortest paths trees. *Journal of Algorithms*, 34(2):251–281, 2000.

# Floating Buffer R-Trees

Craig Dillabaugh

Department of Computer Science, Carleton University, Ottawa, Canada
cdillaba@connect.carleton.ca
www.scs.carleton.ca/~cdillaba

## ABSTRACT

The use of web based Geographic Information Systems (GIS) has grown significantly in recent years, with examples being services such as Google Maps and Mapquest. Such databases often store data that is static or is changed only infrequently as such the most common operations performed on the databases are queries, in particular window queries where the set of database objects stored in rectangular query window must be returned to the client. For databases that are under heavy query loads answering queries in batches provides an opportunity to improve the I/O efficiency of answering all queries. In [1] a buffered version of the R-Tree [2] data structure is presented which provides an efficient technique for answering batched insertions, deletions and query operations on R-Trees. One disadvantage of this buffered R-Tree is that operations are stored in buffers until those buffers become full at which time the buffers are flushed. For a web based GIS is that clients are expecting queries to be answered more or less immediately. The buffered R-Tree can be used to answer such queries immediately if all buffers are flushed following each batch of queries, but doing so negates some of the benefits of this data structure. This research presents a variation of the buffered R-Tree that is designed specifically to quickly answer batches of queries. The buffered R-Tree uses buffers associated with nodes at fixed levels in the R-Tree while the technique presented here, the *Floating Buffer R-Tree*, allows buffers to be assigned dynamically to nodes in the tree as needed. The basic algorithm involves continuously routing queries in an input set to the leaf level until available main memory is exhausted. If a query is routed to a node not in main memory it is stored in the buffer that is assigned to that node's parent. Buffers are assigned only to nodes where queries must be stored in such a manner. All queries are reported by continuously flushing existing buffers until all buffers are empty. To evaluate the effectiveness of the floating buffer tree the technique has been implemented, along with the buffered R-Tree, and regular R-Tree to compare the I/O efficiency of the three techniques for batched queries.

## References

[1] Arge L, KH Hinrichs, J Vahrenhold, JS Vitter Efficient Bulk Operations on Dynamic R-Trees *Algorithmica*, 33:104–128, 2002.

[2] Guttman A. R-Trees: A Dynamic Index Structure for Spatial Searching *SIGMOD'84, Proceedings of Annual Meeting, Boston, Massachusetts, June 18-21*, 47-57, 1984.

# Mining logical constraints from a data stream

Anton Dries

Department of Computer Science,
Katholieke Universiteit Leuven, Belgium
anton.dries@cs.kuleuven.be
www.cs.kuleuven.be/~anton

## ABSTRACT

In recent years, there is an increasing interest in the development of mining algorithms for data streams. Dealing with concept drift is one of the major challenges researchers are faced with when designing these algorithms [2].

We propose a framework for mining k-CNF-expressions from a data stream. A k-CNF expression is a logical formula in conjunctive normal form, i.e., a conjuction of disjunctive clauses.

k-CNF expressions play an important role in Computational Learning Theory where it has been shown that they are PAC-learnable [1]. We can thus derive theoretical bounds on the number of examples needed to achieve a given accuracy (with a certain confidence). We believe that these results can be used to detect and quantify concept drift and help us in finding the optimal settings for our window-based approach (related to [3]).

Another benefit of CNF-expressions is that they are easy to understand. We can see them as logical constraints that hold on the data. Because they are logical formulas, we can reason with them in order to understand the dynamics of the concept drift and predict these dynamics on a higher level.

This poster presents an incremental algorithm for mining k-CNF expressions from a data stream and upgrade it to a window-based approach. We will mainly focus on the technical aspects of this algorithm and discuss the techniques used to traverse the search space and the data structures involved.

Joint work with prof. Luc De Raedt (Department of Computer Science, KULeuven).

## References

[1] L. Valiant. A theory of the learnable. In *Proceedings of the ACM Symposium on Theory of Computing (STOC)*, pages 436–445. ACM Press New York, NY, USA, 1984.

[2] K. Wang, S. Zhou, C. Fu, and J. Yu. Mining Changes of Classification by Correspondence Tracing. In *Proceedings of the SIAM Conference on Data Mining*. SIAM Publishing, 2003.

[3] G. Widmer and M. Kubat. Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 23(1):69–101, 1996.

# Succinct $O(|P| \log \sigma)$-Time Pattern-Search in Suffix Arrays

Johannes Fischer

Inst. für Informatik, Ludwig-Maximilians-Universität München
Johannes.Fischer@bio.ifi.lmu.de

## ABSTRACT

We propose a novel indexing scheme for locating all *occ* occurrences of a pattern $P$ of length $m$ in a static text $T$ of length $n$ in $O(m \log \sigma + occ)$ time, where $\sigma$ is the size of the alphabet. Our indexing scheme is based on a generalization of a preprocessing scheme for range minimum queries and uses only about $2.54\,n + o(n)$ bits of space. The key technique is the replacement of Cartesian Trees with *Schröder Trees*, which implies some other generalizations of well-known combinatorial objects and numbers, such as Super-Catalan Numbers. Our preprocessing scheme can be combined with compressed representations of suffix arrays, yielding pattern matching times that, for large alphabets, surpass all previous results on compressed indexes. For example, combining our search strategy with the Compressed Suffix Array (Thm. 2 in [1]), locating all *occ* occurrences takes $O((m \log \sigma + occ) \log_\sigma^\alpha n)$ time ($0 < \alpha \leq 1$), for *any* alphabet size $\sigma$, while needing only $\alpha^{-1} H_0 n + O(n)$ bits in total ($H_0$ being the empirical order-0 entropy of the input text). If *occ* is not too small, this improves the currently fastest locating time in compressed indexes, such as [2].

Some bibliographic remarks are in order. It is well known that with suffix *trees* one can achieve $O(m \log \sigma)$ deterministic searching time ($\sigma$ being the size of the alphabet) if the outgoing edges of each node are implemented as a height-balanced binary tree. Concerning the suffix *array*, it has been shown in [3] how to achieve $O(m \log \sigma)$ searching time in the ESA by using a different version of the child-table using $5n + o(n)$ bits of space. Apart from occupying only half of the space, an additional advantage of our preprocessing scheme is that it is also much simpler that of [3]. We also emphasize the fact that parts of our methods are necessary for the claimed $O(m \log \sigma)$-search in compressed suffix trees, which the author of [4] claims to achieve, without giving any details how.

Joint work with Volker Heun.

## References

[1] R. Grossi and J. S. Vitter. Compressed Suffix Arrays and Suffix Trees with Applications to Text Indexing and String Matching. *SIAM J. Comput.*, 35(2):378–407, 2005.

[2] P. Ferragina, G. Manzini, V. Mäkinen, and G. Navarro. Compressed Representations of Sequences and Full-Text Indexes. *ACM Transactions on Algorithms*, to apppear 2007.

[3] D. K. Kim and H. Park. A New Compressed Suffix Tree Supporting Fast Search and Its Construction Algorithm Using Optimal Working Space. *Proc. CPM'04*, 33–44, 2004.

[4] K. Sadakane. Compressed Suffix Trees with Full Functionality. *Theory of Computing Systems*, to apppear 2007.

# The Traveling Beams
# Optical Solutions for Bounded NP-Complete Problems

Hen Fitoussi

Department of Computer Science, Ben-Gurion University of the Negev, Israel
henf@cs.bgu.ac.il
www.cs.bgu.ac.il/~henf

## ABSTRACT

Architectures for optical processors designed to solve bounded instances of NP-Complete problems are suggested.

One approach mimics the traveling salesman by traveling beams that simultaneously examine the different possible paths.

The other approach uses a pre-processing stage in which $O(n^2)$ masks are constructed, and uses them as a nested black-boxes with holes. Each mask represents a different edge in the graph. The mask is a binary matrix encoding all possible Hamiltonian paths, where zero and one are represented by opaque and transparent screen respectively. The construction of the masks is done by an iterative process of optical copying. Given an input graph, the choice and combination of the appropriate (small) subset of these masks yields the solution. The solution is rejected in cases where the combination of these masks totally blocks the light and accepted otherwise.

We present detailed designs for basic primitives of the optical processor. We propose designs for solving Hamiltonian path, Traveling Salesman, Clique, Independent Set, Vertex Cover, Partition, 3-SAT, and 3D-matching. Extended abstract of this paper will appear in [1].
Joint work with Shlomi Dolev

## References

[1] S. Dolev and H. Fitoussi. The Traveling Beams: Optical Solutions for Bounded NP-Complete Problems. *Fourth International Conference on Fun with Algorithms*, LNCS, vol. 4475, pp. 120–134, 2007.

# Efficient Data Structure for Particle Repulsive Force Calculation in Graph Drawing

Karlis Freivalds

Institute of Mathematics and Computer Science, University of Latvia, Latvia
Karlis.Freivalds@mii.lu.lv
www.gradetools.com/karlisf

## ABSTRACT

We present a simple multi-dimensional data search structure similar to R-tree [2] that allows efficient range and nearest neighbour queries. We consider a set of points on the plane and R-tree data structure splits them into hierarchically nested, non-overlapping minimum bounding rectangles. Each non-leaf node of the tree stores the bounding rectangle of the subtree and pointers to two child nodes. Each leaf node stores a list of points.

The tree is built by recursive splitting points of the given rectangle into two sets by the largest dimension of the bounding rectangle. The split is performed at the median value of the point coordinates in the selected dimension. The process stops when the number of points in each node drops below a given threshold.

This data structure can be used to perform efficient nearest neighbor and rangle queries.

We use this data structure for approximating repulsive forces in a particle system required in force-directed type graph drawing algorithms[3]. Previously, a quadtree based approximation algorithm by Barnes and Hut [1] was used. But this algorithm is rather slow and the calculated force is not symmetric (with zero mean). Symmetric force is necessary in interactive layout to achieve layout stability.

To calculate the forces, the tree is traversed starting from the root. At each level, node pairs are identified which are far enough and the repulsive force is calculated and accumulated for these pairs. Other pairs are expanded into the next level.

Practical experiments show that repulsive force calculation using the new data structure is about 3.5 times faster than Barnes-Hut data structure. Repulsive force calculation is the most time consuming part of force-directed layout. The performance of the whole graph layout algorithm improved more than 2 times by using this data structure. Also we achieve stable layouts since the force is symmetric.

## References

[1] J. Barnes, P. Hut. A hierarchical O($n$ log $n$) force calculation algorithm. *Nature*, Vol. 324, pp. 446-449. 1986.

[2] A. Guttman. R-Trees: A Dynamic Index Structure for Spatial Searching. *Proc. 1984 ACM SIGMOD International Conference on Management of Data*, pp. 47-57. ISBN 0-89791-128-8.

[3] S. Hachul, M. Jnger An Experimental Comparison of Fast Algorithms for Drawing General Large Graphs. *Proc. Graph Drawing*, LNCS vol. 3843, pp. 235-250, 2006.

# Improved Exact Algorithms for Counting 3- and 4-Colorings

Serge Gaspers

Department of Informatics, University of Bergen, Norway
serge@ii.uib.no
http://www.ii.uib.no/~serge/

## ABSTRACT

We introduce a generic algorithmic technique and apply it on decision and counting versions of graph coloring, one of the oldest and most intensively studied problems in Combinatorics and Algorithms.

Our approach is based on the following idea: either a graph has nice (from the algorithmic point of view) properties which allow a simple recursive procedure to find the solution fast, or the pathwidth of the graph is small, which in turn can be used to find the solution by dynamic programming. We design a general framework combining algorithms enumerating independent sets or maximal independent sets, and pathwidth based algorithms. A direct application of this framework gives the fastest known exact algorithms

- running in time $\mathcal{O}(1.7272^n)$ for deciding if a graph is 4-colorable and
- running in time $\mathcal{O}(1.6262^n)$ and $\mathcal{O}(1.9464^n)$ for counting the number of $k$-colorings for $k = 3$ and 4 respectively.

The full version of this paper, which is joint work with Fedor V. Fomin and Saket Saurabh, appears in [1].

## References

[1] F. V. Fomin, S. Gaspers, and S. Saurabh. Improved Exact Algorithms for Counting 3- and 4-Colorings. *Proceedings of the 13th Annual International Computing and Combinatorics Conference (COCOON 2007)*, Springer LNCS, to appear.

# Dynamic FM-Index for a Collection of Texts with Application to Space-efficient Construction of the Compressed Suffix Array

Wolfgang Gerlach

AG Genominformatics, Technical Faculty, Bielefeld University, Germany
Wolfgang.Gerlach@CeBiTec.Uni-Bielefeld.DE
www.wolfgang-gerlach.com

## ABSTRACT

We have implemented a dynamic FM-index for a collection of texts, a self-index whose space usage is bound by the 0-order entropy of the text to be encoded. The index allows to insert or delete texts of the collection, which avoids reconstruction of the whole index as it is necessary for static indices even for small modifications. We have shown that the insertion of small texts into our dynamic FM-index can be faster than the construction of the static FM-index (v.2) by Ferragina and Manzini [2]. Furthermore, we have shown that an immediate result of the dynamic FM-index is the space-efficient construction of a compressed suffix array which consists of the Burrows-Wheeler Transformation, sampled suffix array, and its inverse.

## References

[1] M. Burrows and D. J. Wheeler. A Block-Sorting Lossless Data Compression Algorithm. *Digital System Research Center*, 124: 1, 1994.

[2] P. Ferragina and G. Manzini. Indexing compressed texts. *Journal of the ACM*, 52(4): 552–581, 2005.

[3] R. Grossi and J. S. Vitter. Compressed Suffix Arrays and Suffix Trees with Applications to Text Indexing and String Matching. *SIAM Journal on Computing*, 35(2): 378–407, 2005.

[4] V. Mäkinen and G. Navarro. Dynamic Entropy-Compressed Sequences and Full-Text Indexes. *Proceedings of the 17th Annual Symposium on Combinatorial Pattern Matching (CPM 2006)*, LNCS 4009, 307–318, 2006.

# Constructing and Pruning Spanners in the Cache-Oblivious Model

Fabian Gieseke

Department of Computer Science, University of Dortmund, Germany
fabian.gieseke@cs.uni-dortmund.de
http://ls11-www.cs.uni-dortmund.de/people/gieseke/

## ABSTRACT

A $t$-spanner for a finite point set $S \subset \mathbb{R}^d$ is a euclidean graph $G = (S, E)$ in which the length of the shortest path between two arbitrary points $p, q \in S$ is at most $t$-times the euclidean distance between them. The minimum value $t'$ for which a euclidean graph $G = (S, E)$ is a $t'$-spanner for its vertex set $S$ is called the *dilation* of $G$. Many algorithms are known that compute $t$-spanners for a given point set $S \subset \mathbb{R}^d$ with $\mathcal{O}(|S|)$ edges and additional properties such as bounded degree, small spanner diameter and low weight; see, e.g., the survey [3]. In addition to the construction of $t$-spanners, the problem of *pruning* $t$-spanners received attention in recent works [2]. That is, given a euclidean graph $G = (S, E)$ with constant dilation $t$ and a positive constant $\varepsilon$, the aim is to compute a subgraph $G'$ of $G$, which is a $t(1 + \varepsilon)$-spanner with $\mathcal{O}(|S|)$ edges.

The cache-oblivious model of Frigo *et al.* [1] is a computation model which captures the characteristics of working with large data sets without having detailed information about the underlying memory hierarchy. This poster presents the results of my master's thesis, namely constructing and pruning $t$-spanners efficiently in the cache-oblivious model of computation.

## References

[1] M. Frigo, C. E. Leiserson, H. Prokop and S. Ramachandran: Cache-Oblivious Algorithms. *Proceedings of the 40th Annual Symposium on the Foundations of Computer Science*, 285–299, 1999.

[2] J. Gudmundsson, C. Levcopoulos, G. Narasimha and M. Smid: Approximate Distance Oracles for Geometric Graphs. *Proceedings of the 13th annual ACM-SIAM symposium on Discrete algorithms*, 828–837, 2002.

[3] G. Narasimhan and M. Smid: Geometric Spanner Networks. Cambridge University Press, 2007.

# Faster Path Summaries

Nils Grimsmo

Department of Computer and Information Science,
Norwegian University of Science and Technology
nils.grimsmo@idi.ntnu.no
http://www.idi.ntnu.no/~nilsgri/

## ABSTRACT

Query languages for semi-structured data, such as XPath for searching in XML, has predicates both for structure and content. Many early systems viewed the XML data as a string of XML elements and term elements, and matching was performed by searching for all query elements separately and mering the results [3]. This can be very inefficient if any of the XML elements in the query have poor selectivity, giving costly merges.

Since XML data often follows a schema, the set of all XML element paths seen in a document collection is often small compared to the total data size. Data structures for indexing the structure of the XML is usually called *path summaries* of *path indexes*. The DataGuide [1] from the Lore database system for semi-structured data pioneered the idea of path summaries.

This work describes how to implement efficient memory resident path summaries. The set of all XML paths seen in the data is indexed as a set of strings from the alphabet of all XML elements. Two techniques are introduced, and they are shown to be significantly faster than previous methods when facing path queries using the descendant axis and wildcards. The first is conceptually simple and combines inverted lists, selectivity estimation, hit expansion and brute force search. The second builds on work by Zuopeng et al. [4], which uses a generalised suffix tree [2] for indexing the set of path strings. The variant used here has additional statistics in the tree, and uses multiple entry points into the query. The entry points are partially evaluated in an order based on estimated cost until one of them is complete.

## References

[1] R. Goldman and J. Widom. DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases. In *Proc. VLDB*, pages 436–445, 1997.

[2] D. Gusfield. *Algorithms on strings, trees, and sequences: Computer science and computational biology.* 1997.

[3] C. Zhang, J. Naughton, D. DeWitt, Q. Luo, and G. Lohman. On Supporting Containment Queries in Relational Database Management Systems. *SIGMOD Rec.*, 30(2):425–436, 2001.

[4] L. Zuopeng, H. Kongfa, Y. Ning, and D. Yisheng. An Efficient Index Structure for XML Based on Generalized Suffix Tree. *Information Systems*, 32:283–294, 2007.

# Comparing Segmentations by Applying Randomization Techniques

Niina Haiminen

Department of Computer Science, University of Helsinki, Finland
niina.haiminen@cs.helsinki.fi
www.cs.helsinki/u/haiminen

## ABSTRACT

Segmentation is a widely used and general method for analyzing sequential data. Numerous segmentation techniques have been introduced in various application domains. Furthermore, a single technique can be applied on different features of the same sequence, thus yielding alternative segmental descriptions.

We investigate the problem of evaluating the quality of a proposed segmentation, when the underlying true segmental structure is known. We apply randomization techniques to estimate the probability of obtaining an equally good segmentation by chance. Our randomization tests can be applied when deciding if some segmentation technique, or feature of the data, yields results significantly close to previously verified segmental structures in the sequence.

We apply a distance measure based on conditional entropies to characterize the similarities between segmentations. Our aim is to transform the process of quality evaluation from simply viewing the segmentations and looking for similarities, to obtaining $p$-values denoting significance of segmentation similarity.

We demonstrate the use of our randomization techniques on two examples of biological segmental structures. Previous work on segmentation similarities in biological context has been done in the case where the underlying data sequence is also known. In our approach, we focus on comparing the segment boundaries only, so that we can easily compare results arising from different types of data.

Our example applications include the detection of isochores (large homogeneous DNA blocks) and the discovery of coding-noncoding structure. We obtain segmentations of relevant sequences by applying different techniques, and use alternative features to segment on. We show that some of the obtained segmentations are very similar to the underlying true segmentations, and this similarity is statistically significant. For some other segmentations, we show that equally good results are likely to arise by chance.

Joint work with Heikki Mannila and Evimaria Terzi.

# Learning Sparse Networks From Poor Data

Goele Hollanders

Department of Mathematics, Physics and Computer Science, Hasselt University
and Transnational University of Limburg, Hasselt, Belgium
goele.hollanders@uhasselt.be
www.uhasselt.be/goele.hollanders

## ABSTRACT

This paper is concerned with the learning process of a sparse interaction network, for example, a gene-protein interaction network. The advantage of the process we propose is that there will always be a student $S$ that fits the teacher $T$ very well with a relatively small data set and a high number of unknown components, i.e., when the number of measurements $M$ is significantly smaller than the system size $N$.

To measure the efficiency of this learning process, we use the generalization error, $\epsilon_{gen}$, which represents the probability that the student is a good fit to the teacher. From our experiments it follows that the quality of the fit depends on several factors: First, the ratio $\alpha = M/N$ of the number of measurements to the system size has a strong impact. Surprisingly, we find that a sudden identification transition occurs for value $\alpha \approx \alpha_{gen}$ which corresponds to $\epsilon_{gen} = 1/2$. From this sample size onwards the student will be a good fit to the teacher. Interestingly, the generalization threshold $\alpha_{gen}$, will always be significantly smaller than 1. Second, the quality of the fit depends on the sparsity of the network. If the number of non-zero components increases, as sparsity disappears, the efficiency of the process will gradually increase. Finally there is an impact of the noise level. The learning process is robust to noise upto a certain threshold. We see that, at this level, the impact on the noise suddenly and dramatically increases as a consequence of which the student will no longer be a good fit to the teacher.

Full version of this paper appears in [1] and is a joint work with Geert Jan Bex, Marc Gyssens, Karl Tuyls and Ronald L. Westra.

## References

[1] G. Hollanders, G. Bex, M. Gyssens, R.L.Westra, K. Tuyls Learning Sparse Networks From Poor Data. *Belgian-Dutch Benelearn Conference*, 2007.

[2] R. Westra, G. Hollanders, G. Bex, M. Gyssens, K. Tuyls Piecewise Linear Modeling of Gene-Protein Interaction Networks. *LNBI*, 4366/2007:157–170, 2007.

[3] R. Westra, G. Hollanders, G. Bex, M. Gyssens, K. Tuyls Reconstruction of Flexible Gene-Protein Interaction Networks using Piecewise Linear Modeling and Robust Regression. *Workshop on Adaptation in Artificial and Biological Systems (AISB)*, 2006.

# Comparison of $s$-gram Proximity Measures in CLIR

Antti Järvelin

Department of Computer Sciences, University of Tampere, Finland
`antti.jarvelin@cs.uta.fi`
`www.uta.fi/~antti.jarvelin`

## ABSTRACT

Cross-Language Information Retrieval (CLIR) refers to retrieval of documents written in a language other than that of the user's request. A typical approach to CLIR is automatically translating the query into the target language(s). Due to the terminology missing from dictionaries, out-of-vocabulary (OOV) words constitute a major problem in query translation in CLIR and in machine translation. Many typical OOV words, like proper names and technical terms, are often important query keys [1]. Therefore their successful translation is essential for query performance. In European languages, technical terms often share a common Greek or Latin root but are rendered with different spelling of the underlying sounds. This provides a good basis for the use of approximate string matching in translation.

Classified $s$-grams have been successfully used in CLIR as an approximate string matching technique [2]. $s$-grams have consistently outperformed other approximate string matching techniques, such as edit distance or traditional $n$-grams. In $s$-gram matching, the Jaccard coefficient has traditionally been used as an $s$-gram based string proximity measure. However, other proximity measures for $s$-gram matching have not been tested. The current study addresses this question by evaluating the performances of seven proximity measures for classified $s$-grams in CLIR context. The performance was evaluated based on the proximity measures' capability to find correct translations to source words from large database indices between eleven language pairs. The results showed that the binary proximity measures (e.g. Jaccard coefficient) were better suited for the $s$-gram matching than the non-binary. For $n$-grams the non-binary proximity measures performed slightly better than the binary measures, though the type of proximity measure used is of more importance.

Full version of this paper is submitted to the SPIRE 2007 conference and is co-written with MSc Anni Järvelin, Department of Information Studies, University of Tampere, Finland.

## References

[1] A. Pirkola. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. *SIGIR '98: Proceedings of the 21st ACM SIGIR Conference*, 55–63, 1998.

[2] H. Keskustalo, A. Pirkola, K. Visala, E. Leppänen, K. Järvelin. Non-adjacent Digrams Improve Matching of Cross-Lingual Spelling Variants. *SPIRE 2003: Proceedings of the 10th SPIRE Conference*, 252–265, 2003.

# Industrial-Strength SAT Solving and Restricted Branching

Matti Järvisalo

Helsinki University of Technology (TKK), Finland
firstname.lastname@tkk.fi
http://www.tcs.tkk.fi/~mjj/

## ABSTRACT

Considerable recent advances in fundamental and implementational techniques for constraint-based declarative problem solving have established constraint-based methods as competitive and even dominant compared to more specific algorithmic approaches for solving computationally difficult problems in a wide range of applications. Propositional satisfiability (SAT) solving procedures, in particular, have been found to be extremely efficient as back-end search engines in solving large industrial-scale combinatorial problems in domains such as computer-aided verification (for example, bounded model checking of finite state systems) and automated planning. In such domains, huge search spaces need to be covered in order to solve the resulting problem instances. Hence techniques that exploit structural properties of problem instances become vital in developing efficient solving methods.

Since irrelevant decisions may have an exponential effect on the running times of the solver, techniques for making decisions, i.e., branching, play a central role in complete SAT methods. In many typical real-world problem domains, the constraint encoding is often derived from a transition relation, where the behaviour of the underlying system is dependent only on the *input*—initial state, nondeterministic choices, et cetera—of the system. Since the system behaviour is determined by its input, SAT solving remains complete when branching is restricted to *input variables* of the propositional encoding, representing the system input. Intuitively, this drops the search space size from $2^N$ to $2^I$ with $I << N$, where $I$ and $N$ are the number of input variables and all variables in the encoding, respectively.

A natural question to ask is *whether the power of the underlying inference systems of SAT solvers is affected by the input-restriction.* The poster presents recent answers to this question through *proof complexity* theoretical characterizations of SAT solvers. Namely, our results imply that all implementations of the popular DPLL SAT procedure (*with or without* the widely applied efficiency-boosting technique of *clause learning*), even with optimal search heuristics, have the potential of suffering a notable efficiency decrease if branching is restricted to input variables. We also present experimental results aiming at explaining why branching restrictions are difficult to apply with typical SAT solver techniques.

The results appear in [1, 2], and are joint work with Tommi Junttila and Ilkka Niemelä.

## References

[1] M. Järvisalo and T. Junttila and I. Niemelä. Unrestricted vs Restricted Cut in a Tableau Method for Boolean Circuits. AMAI **44**(4) (2005) 373–399

[2] M. Järvisalo and T. Junttila. Limitations of Restricted Branching in Clause Learning, 2007. Submitted manuscript.

# An IO-efficient Approach to Decomposition of Geospatial Data

Anders Hessellund Jensen

Department of Computer Science, University of Aarhus, Denmark
ahj@daimi.au.dk
http://www.daimi.au.dk/~ahj

## ABSTRACT

Today it is possible to acquire higly detailed digital terrain models (DTM) using LIDAR and other mapping technologies. Such detailed terrain models are often far too large to fit in internal memory.

Most GIS software has highly efficient algorithms for processing geospatial data in internal memory. An often used approach is therefore to decompose the input into a set of smaller, non-overlapping segments $S$. The data can now be processed by running the internal memory algorithm each segment $s$ in $S$.

Typically this will cause problems around the edges of each segment. For example, if a Grid is generated from a point cloud using this approach, there will be visible inaccuracies around the edges of each segment.

One solution to this problem is to compute, for each segment $s$, the set of neighboring segments and run the internal memory algorithm on the union of $s$ and its neighbors. An IO efficient solution to this problem is presented in [2]. The decomposition used is an IO-efficient quadtree. I present a solution based on morton codes [1] and an IO-efficient priority queue [3]. I use the same decomposition, but by avoiding to construct the quadtree explicitly, I get a slightly better worst-case performance, namely $O\left(\frac{N}{B}\log_{\frac{N}{B}}\frac{N}{B}\right)$, compared to $O\left(\frac{N}{B}\frac{h}{\log\frac{M}{B}}\right)$, where $h$ is the height of the quadtree. Furthermore, due to the locality of the Morton-order, I expect my algorithm to perform better in practice.

## References

[1] G. R. Hjaltason and H. Samet Speeding up construction of quadtrees for spatial indexing. *VLDB*, 11(2):109-137, 2002.

[2] P. K. Agarwal, L. Arge and A. Danner From Point Cloud to Grid DEM: A Scalable Approach. *Proc. Intl. Sympos. Spatial Data Handling*, 2006.

[3] G. S. Brodal and J. Katajainen. Worst-case efficient external-memroy priority queues. *Proc. Scandinavian Workshop on Algorithms Theory*, pages 107-118, 1998.

# Resilient Algorithms in the Faulty-Memory Model

Allan Grønlund Jørgensen, PhD Student

BRICS, Department of Computer Science, University of Aarhus, Denmark
jallan@daimi.au.dk
www.daimi.au.dk/~jallan

## ABSTRACT

Finocchi and Italiano [1] introduced the *faulty-memory RAM* where cells can get corrupted at any time execution of the algorithm. Motivated by the fact that registers in the processor are considered uncorruptible, $O(1)$ safe memory locations are provided. The model is parametrized by an upper bound, $\delta$, on the number of corruptions occurring during an algorithm. Finally, moving values is considered an atomic operation. An algorithm is resilient if it works correctly, at least on the set of uncorrupted cells in the input.

Several problems have been addressed in the faulty-memory RAM. In the original paper [1], algorithms for sorting and searching were introduced. Matching upper and lower bounds for sorting and randomized searching, were given in [2]. Recently, resilient search trees that support searches, insertions, and deletions in $O(\log n + \delta^2)$ amortized time [3] were introduced.

Our work in this model includes priority queues[1], and dictionaries[2]. We design the first priority queue in the faulty-memory RAM model. It uses linear space and performs both Insert and Deletemin operations in $O(\log n + \delta)$ time amortized. We introduce a resilient randomized static dictionary that support searches in $O(\log n + \delta)$ time. This dictionary is simple and uses only $O(\log \delta)$ worst case random bits. We give the first optimal resilient deterministic dictionary. It supports searches in a sorted array in $O(\log n + \delta)$ time in the worst case, matching the lower bounds from [1]. We introduce a dynamic dictionary supporting searches in $O(\log n + \delta)$ in the worst case, and insertions and deletions in $O(\log n + \delta)$ time amortized. Range queries are supported in $O(\log n + \delta + k)$ time, where $k$ is the output size.

## References

[1] Irene Finocchi and Giuseppe F. Italiano Sorting and searching in the presence of memory faults (without redundancy) *Proc. 36th Annual ACM Symposium on Theory of Computing*: 101–110, 2004

[2] Irene Finocchi and Fabrizio Grandoni and Giuseppe F. Italiano Optimal Resilient Sorting and Searching in the Presence of Memory Faults *Proc. 33rd International Colloquium on Automata, Languages and Programming*: 286–298, 2006

[3] Irene Finocchi and Fabrizio Grandoni and Giuseppe F. Italiano Resilient search trees *Proc. 18th ACM-SIAM Symposium on Discrete Algorithms*: 547–554, 2007

[1] Joint work with, Thomas Mølhave and Gabriel Moruz

[2] Joint work with, Thomas Mølhave, Gabriel Moruz, Gerth S. Brodal and Rolf Fagerberg

# How Branch Mispredictions Affect Quicksort

Kanela Kaligosi

Max-Planck-Institut für Informatik, Saarbrücken , Germany
kaligosi@mpi-inf.mpg.de
www.mpi-inf.mpg.de/~kaligosi

## ABSTRACT

Sorting is one of the most important algorithmic problems both practically and theoretically. Quicksort is perhaps the most frequently used sorting algorithm since it is very fast in practice, needs almost no additional memory, and makes no assumptions on the distribution of the input. Hence, quicksort, its analysis and efficient implementation is discussed in most basic courses on algorithms. When we take a random pivot, the expected number of comparisons is $2n \ln n \approx 1.4n \lg n$. One of the most well known optimizations is that taking the median of three elements reduces the expected number of comparisons to $\frac{12}{7} n \ln n \approx 1.2n \lg n$. Indeed, by using the median of a larger random sample, the expected number of comparisons can be made as close to $n \lg n$ as we want. At first glance, counting comparisons makes a lot of practical sense since in quicksort, the number of executed instructions and cache faults grow proportionally with this figure.

However, in comparison based sorting algorithms like quicksort or mergesort, neither the executed instructions nor the cache faults dominate execution time. Comparisons are much more important, but only indirectly since they cause the direction of branch instructions depending on them to be mispredicted. In modern processors with long execution pipelines and superscalar execution, dozens of subsequent instructions are executed in parallel to achieve a high peak throughput. When a branch is mispredicted, much of the work already done on the instructions following the predicted branch direction turns out to be wasted. Therefore, ingenious and very successful schemes have been devised to accurately *predict* the direction a branch takes.

Our main theoretical contribution is an analysis of quicksort in the context of branch mispredictions. For simplicity we assume that the elements are distinct. We look at two variants of quicksort: random and skewed pivot, and three branch prediction methods: static, 1-bit predictor and 2-bit predictor. The theoretical results are complemented by experiments. In particular, we also look at the classical median-of-three pivot selection. It turns out that this frequently used improvement only gives a negligible advantage over random pivot. Its advantages wrt. instruction count basically cancel with its disadvantages wrt. branch prediction. Somewhat surprisingly, taking a pivot with rank around $n/10$ can lead to a better performance.

This work appears in [1] and is joint with Peter Sanders.

## References

[1] K. Kaligosi and P. Sanders. How Branch Mispredictions Affect Quicksort. *In Proc. 14th Annual European Symposium on Algorithms*, 780–791, 2006.

# Towards Autonomous Clustering: Voronoi Clustering

Heidi Koivistoinen

Department of Information Technology, Tampere University of Technology, Finland
heidi.koivistoinen@tut.fi
http://www.cs.tut.fi/~hkoivist

20.4.2007

## ABSTRACT

Clustering is a basic tool in unsupervised machine learning and data mining. The idea of clustering is to define natural groupings in the data by learning the relevant information of the unlabeled instances. To obtain a clustering, algorithms can use two basic approaches: either a bottom-up *agglomerative* or a top-down *divisive* construction. The former begins with each instance in same cluster and successively merges clusters together until it meets a stopping criterion. In the latter approach, one begins with all instances in a single cluster and partitions the clusters until satisfaction of a stopping criterion. The results of basic clustering methods suffer from using random numbers as initial values and being dependent on given exactly the right number of clusters. Random initial cluster centers can be far too incorrect and the centers cannot be moved enough during the iterations. Wrong number of clusters can divide points of all clusters incorrectly causing totally useless results.

Autonomous algorithms learn all the information needed, so data-descriptive parameters, such as the number of clusters, are not needed, and through redefinition of the original clustering problem use of random values can be avoided. Autonomous clustering typically includes a basic clustering algorithm and a complex parameter-learning phase that increases time and memory requirements considerably. Distance-based algorithms rarely have the means to autonomously come up with the correct number of clusters from the data [2]. Our recent approach [1] to identify the natural clusters autonomously is to compare the point densities in different parts of the sample space. We have put forward an agglomerative algorithm which accesses density information by constructing a *Voronoi diagram* for the input sample. The volumes of the point cells directly reflect the point density in the respective parts of the instance space. Scanning through the input points and their Voronoi cells once, we combine the densest parts of the instance space into clusters. The number of clusters is now a by-product of the algorithm, and the new more learnable parameter is the maximum size of a dense area. Our empirical experiments demonstrate the proposed algorithm is able to come up with a high-accuracy clustering for many different types of data.

## References

[1] Koivistoinen, Ruuska, and Elomaa, A Voronoi Diagram Approach to Autonomous Clustering. 9th International Conference on Discovery Science, Spain, 137–148, 2006.

[2] Elomaa and Koivistoinen, On Autonomous k-Means Clustering. 15th International Symposium on Methodologies for Intelligent Systems, USA, 228–236, 2005.

# Imitating the Outcome of Ascertainment Process for Coalescent Simulator Output

Jussi Kollin

Helsinki Institute for Information Technology
Department of Computer Science, University of Helsinki, Finland
jussi.kollin@cs.helsinki.fi
www.cs.helsinki.fi/~jkollin

## ABSTRACT

The genetic diversity within a population can be examined in a relatively cheap and efficient way by using single nucleotide polymorphisms (SNPs). Recently, comprehensive human SNP data sets of over 1.5 million SNPs have been released to the public, e.g., [1], creating new opportunities for genome-wide population genetic analysis.

Computationally intensive methods for hypothesis testing require the ability produce synthetic data that matches the null hypothesis. Coalescent simulation is frequently the method of choice to generate these synthetic haplotypes. The simulation requires a number of parameters to be set. Recently, Schaffner et al. [2] presented simulation parameters calibrated to generate haplotype data close to real haplotype data in terms of a number of statistics, such as mean $r^2$ as a function of distance and allele frequency spectrum.

The sampling bias in the selection of SNPs for the data set is called ascertainment bias. This sampling or ascertainment process is pivotal in the selection of SNPs to create realistic synthetic data with statistics matching those of real data. This bias can be corrected e.g. by modeling the process or by weighting the SNPs. The ascertainment process is not always easy to model, possibly to the extent of making this approach infeasible in some cases.

We examine a method of filtering the SNPs with the intention of refining the raw synthetic data, generated by the simulator by Schaffner et al. [2], and the Perlegen data [1] to have same characteristics, independent of the ascertainment process originally used for the real data.

This is joint work with Mikko Koivisto and Heikki Mannila.

## References

[1] David. A. Hinds, Laura L. Stuve, Geoffrey B. Nilsen, Eran Halperin, Eleazar Eskin, Dennis G. Ballinger, Kelly A. Frazer, and David R. Cox. Whole-genome patterns of common DNA variation in three human populations. *Science*, 307:1072–1079, 2005.

[2] Stephen F. Schaffner, Catherine Foo, Stacey Gabriel, David Reich, Mark J. Daly, and David Altshuler. Calibrating a coalescent simulation of human genome sequence variation. *Genome Research*, 15:1576–1583, 2005.

# Analysis of tree growth using tree ring data and weather patterns

Mikko Korpela

Laboratory of Computer and Information Science
Helsinki University of Technology (TKK), Finland
mvkorpel@cis.hut.fi
www.cis.hut.fi/mvkorpel/

## ABSTRACT

The forest industry continues to be a significant source of income in the Finnish economy. We want to model the effects of weather on the yearly growth of trees. Existing applications of these interactions include the reconstruction of past climate based on tree growth rings. It is also interesting how the prospective climate change will affect trees.

We have growth data of hundreds of fir trees from several locations in Finland. The accompanying weather data contains daily measurements of temperature, precipitation, and other variables, dating as far back as the 1950s. We look at the relationship between weather and growth as a regression problem. The first step is to model growth based on only the age of the tree. Using this approach, we may arrive at smoothed versions of the growth patterns of individual trees. In the second step, we use previous growth values as base inputs. We aim to see whether adding some parts of the weather data to the model improves the accuracy of growth predictions. The goal is to model either the growth as such, or the residual growth after removing the smoothed growth. Domain-knowledge of forest researchers is used in setting the goals and parameters of the smoothing. With different parameters, the focus shifts between long-term and short-term growth phenomena. A prospective solution for the regression problem is Gaussian processes [1], a Bayesian approach. Due to the abundance of weather data, variable selection is also an issue to consider. Wrapper methods [2] can be used for variable selection, although their time requirements may be a problem.

This is joint work with Jaakko Hollmén and Mika Sulkava from Laboratory of Computer and Information Science at TKK and Harri Mäkinen from Finnish Forest Research Institute (Metla).

## References

[1] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning.* MIT Press, 2006.

[2] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1–2):273–324, 1997.

# Property-Based Collapsing in Hierarchical Visualisation

Darja Krushevskaja Meelis Kull

## ABSTRACT

Hierarchical visualisation techniques are extensively used due to their capabilities to display the similarities of data objects in different granularity. In our work we concentrate on improving the visual outcome of hierarchical clustering.

For a large data set the tree with all objects cannot be presented in one picture. Therefore, collapsed nodes must be used, each giving a summarised view of the corresponding subtree. Usually, collapsing is performed on subtrees that are either at fixed depth or of fixed size.

We propose an algorithm that works on data sets where each object is described by two sets of properties. The sets are used separately at different stages of algorithm – the first set is used for hierarchical clustering whereas the other set, which consists of binary features only, is used for collapsing.The algorithm looks for subtrees with significant overrepresentation of some binary feature and favours these in the collapsing process.

We have built a microarray gene expression data visualisation system implementing this technique. The genes are clustered by their expression patterns using our fast approximate hierarchical clustering algorithm and Gene Ontology annotation enrichments are found using our tool g:Profiler and used to guide collapsing.

# Efficient learning, and limits of, in online problems, with focus on binary search trees

Jussi Kujala

Department of Computer Science, Tampere Institute of Technology, Finland
jussi.kujala@tut.fi
www.cs.tut.fi/~kujala2

## ABSTRACT

In online learning we continuously make decisions under uncertainty about the future. The uncertainty is modelled adversially in contrast to a probabilistic approach. Then we measure the algorithm under study relative to an algorithm or solution that is the best against the adversary, in contrast to f.e. the worst-case cost over all situations. We are interested in methods in online learning, both general ones [2] and ones applied to data structures [4], like binary search trees, with focus on efficient implementation and use of different cost models, like cache-obliviousness, to derive useful results.

We have also studied how to discretize continuous features for use in classifiers such as Naive Bayes [1, 3].

## References

[1] Tapio Elomaa and Jussi Kujala and Juho Rousu Approximation Algorithms for Minimizing Empirical Error by Axis-Parallel Hyperplanes. *16th European Conference on Machine Learning (ECML05)*, Lecture Notes in Computer Science 3720:547–555, 2005.

[2] Jussi Kujala and Tapio Elomaa On Following the Perturbed Leader in the Bandit Setting. *16th International Conference of Algorithmic Learning Theory (ALT05)*, Lecture Notes in Computer Science 3734:371–385, 2005.

[3] Tapio Elomaa and Jussi Kujala and Juho Rousu Practical Approximation of Optimal Multivariate Discretization. *16th International Symposium on Methodologies for Intelligent Systems (ISMIS06)*, Lecture Notes in Computer Science 612-621, 2006.

[4] Jussi Kujala and Tapio Elomaa Poketree: A Dynamically Competitive Data Structure with Good Worst-Case Performance. *The 17th International Symposium on Algorithms and Computation (ISAAC06)*, Lecture Notes in Computer Science 4317:277-288, 2006.

# Clustering Algorithms for Cellular Transition Data

Kari Laasonen

Department of Computer Science, University of Helsinki, Finland
kari.laasonen@cs.helsinki.fi
www.cs.helsinki.fi/kari.laasonen

## ABSTRACT

Determining the current location of a mobile device is a key feature in many pervasive computing scenarios. Representative applications range from location-enabled recommendation and presence services to logistic tracking and planning systems. We use the current serving cell tower id ("cell") as a simple location indicator. Inferring location from cells naturally means that physical coordinates and relationships are unavailable, but it also means that such a service can run on almost any mobile device.

The data we observe is a timed sequence of cell transitions; the cell identifiers are opaque. An important property of the data is that there is not necessarily a one-to-one correspondence between cells and physical locations, as most urban locations are covered by several cells. Furthermore, in certain situations the mobile device can alternate between any available cells, generating transitions that are not associated with user movement.

We use cell clustering to group related cells into clusters, with the idea that a cluster does have a reliable corresponding physical location. This reduces uninteresting transitions and other noise, and increases the information content of the event stream, simplifying the design of higher-level algorithms. All of our clustering algorithms are based on a set of common requirements: clusters corresponding to locations must be small (no pair of cells may be separated by more than one other cell), the algorithm must be able to make local decisions, and preferably evolve the clustering in an online fashion.

We define a number of conditions that a set of cells needs to satisfy to be a cluster. It is not difficult to design an offline algorithm that finds such clusters. However, the search space is large, taking exponential time in the worst case. Checking the cluster conditions furthermore requires the entire transition sequence to be available. Both of these are serious problems for a system constrained by the limited resources of mobile devices.

Our contribution is an online cell clustering algorithm, which tracks cell transitions in two stages. The first stage is executed for each transition: it uses simple heuristics to choose promising sets for the second stage, where chosen sets are tracked more thoroughly. Actual clusters are constructed on demand (e.g., once per day) from these stage two sets. We have also experimented with a completely different similarity-based clustering method, which utilizes the cell transition distribution.

We evaluate the proposed methods with real cell transition data, collected from almost one hundred different users, using a scoring system that combines several measurable cluster properties.

# On the complexity of computing treelength

Daniel Lokshtanov

Department of Informatics, University of Bergen, N-5020 Bergen, Norway.
daniello@ii.uib.no
www.ii.uib.no/~daniello

## ABSTRACT

We resolve the computational complexity of determining the *treelength* of a graph, thereby solving an open problem of Dourusboure and Gavoille, who introduced this parameter, and asked to determine the complexity of recognizing graphs of bounded treelength [1]. The treelength of a graph attempts to measure how far a graph is from being chordal. While recognizing graphs with treelength 1 is easily seen as equivalent to recognizing chordal graphs, which can be done in linear time, the computational complexity of recognizing graphs with treelength 2 was unknown until this result. We show that the problem of determining whether a given graph has treelength at most $k$ is NP-complete for every fixed $k \geq 2$, and use this result to show that treelength in weighted graphs is hard to approximate within a factor smaller than $\frac{3}{2}$. Additionally, we give an $O(1.8899^n)$ algorithm for the Chordal Sandwich, and show that this algorithm can be applied to compute the treelength of $G$ within the same time bound.

## References

[1] Y. Dourisboure and C. Gavoille. Tree-decompositions with bags of small diameter. *To appear in Discrete Mathematics.*

# Maximum independent sets in bounded-degree hypergraphs

Elena Losievskaja

Department of Computer Science, University of Iceland, Iceland
elenal@hi.is

## ABSTRACT

A *hypergraph* $H$ is a pair $(V, E)$, where $V = \{v_1, \ldots, v_n\}$ is a set of vertices and $E = \{e_1, \ldots, e_m\}$ is a collection of subsets of $V$, or (hyper)edges. An independent set in $H$ is a subset of $V$ that doesn't properly contain any edge of $H$. Let MIS denote the problem of finding a maximum independent set in hypergraphs. Given that MIS generalizes the independent set problem in graphs, the problem is NP-hard to approximate within a factor $\Delta/2^{O(\sqrt{\log(\Delta)}}$ unless $P = NP$ [4]. MIS is of fundamental interest, both in practical and theoretical aspects, since it is intimately related with classical covering problems, such as Hitting Set and Set Cover, and arises in various applications in data mining, image processing, database design, parallel computing and many others.

The focus of this work is on hypergraphs with maximum degree at most $\Delta$. We present a general technique that reduces the worst case analysis of certain algorithms to their performance in the case of ordinary graphs. This technique, called *shrinkage reduction*, can be applied to a wide class of algorithms and problems on hypergraphs. In particular, using this technique we show that the greedy algorithm for MIS that corresponds to the classical greedy set cover algorithm has a performance ratio of $(\Delta + 1)/2$, improving the bounds obtained by Bazgan et al. [1]. It also allows us to extend local search algorithms of Berman and Furer [2, 3] on graphs to obtain a $(\Delta + 1)/2$ approximation for weighted MIS and $(\Delta+3)/5+\epsilon$ approximation for unweighted MIS in hypergraphs. We improve the bound in the weighted case to $\lceil(\Delta + 1)/3\rceil$ using a simple partitioning algorithm. Finally, we show that another natural greedy algorithm for MIS, that adds vertices of minimum degree, achieves only a ratio of $\Delta - 1$, significantly worse than on ordinary graphs.

Joint work with Magnús M. Halldórsson, Department of Computer Science, University of Iceland.

## References

[1] C. Bazgan, J. Monnot, V. Paschos, F. Serrière, On the differential approximation of MIN SET COVER, *Theoretical Computer Science*, 332:497-513, 2005.

[2] P. Berman, A $d/2$ approximation for Maximum Weight Independent Set in $d$-claw free graphs, *In. Proc. Seventh Ann. Scandinavian Workshop on Algorithm Theory* LNCS 1851:214–219, 2000.

[3] P. Berman, M. Fürer, Approximating maximum independent set in bounded degree graphs, *In. Proc. Fifth Ann. ACM-SIAM Symp. on Discrete Algorithms* 365–371, 1994.

[4] L. Trevisan, Non-approximability results for optimization problems on bounded degree instances, *In Proc. of the 33rd ACM STOC*, 2001.

# Mining Itemsets in the Presence of Missing Values

Michael Mampaey

Department of Mathematics and Computer Science
University of Antwerp, Belgium
michael.mampaey@ua.ac.be
www.adrem.ua.ac.be/~mmampaey/

## ABSTRACT

Missing values make up an important and unavoidable problem in data management and analysis. In frequent itemset mining, however, this issue never received much attention. Nevertheless, the well known measures of support and confidence are misleading when missing values occur, and more suitable definitions typically don't have the desired monotonicity property of support. In our work, we overcome this problem and provide an efficient algorithm, XMiner, for mining frequent itemsets in databases with missing values.

Several methods exist to deal with missing values - deleting transactions, imputing estimates, using probabilistic measures ... - all with their respective advantages and drawbacks, the latter usually includes loss of information, or a skewing of the data distribution. We use backward compatible definitions of support and confidence, introduced by Ragel & Crémillieux [1], that can be used for different kinds of missingness. These measures apply in specific samples of the database in stead of the complete dataset. More specifically, the support of an itemset is defined as the number of supporting transactions divided by the number of transactions without missing values for the individual items in that itemset. Consequently this sample can be different for each different itemset. Since intuitively frequent sets should be in samples that are large enough, the representativity measure is introduced (the size of a corresponding sample). Unfortunately, the adapted support measure is not monotone.

Our algorithm overcomes this by defining extensible itemsets - sets that have at least one frequent (representative) superset, although the itemset itself might not be frequent. All frequent itemsets are contained in the extensible ones, and furthermore, their support can be derived immediately when checking extensibility, which in turn can be done efficiently because of monotonicity. Our trick is to use representativity to determine part of the extensibility formula, in a fashion similar to maximal itemset mining. Using this technique, we can find both the representativity and extensibility (and thus also support) simultaneously, i.e. in a single lattice traversal. This effectively allows XMiner to mine all frequent itemsets, without generating too much candidates [2].

## References

[1] A. Ragel and B. Crémillieux. Treatment of missing values for association rules. *In Research and Development in Knowledge Discovery and Data Mining, LNAI*, 1998.

[2] T. Calders, B. Goethals, M. Mampaey. Mining Itemsets in the Presence of Missing Values. *In Proceedings of the ACM Symposium on Applied Computing*, 2007.

# Computing and extracting minimal cograph completions in linear time

Daniel Lokshtanov, Federico Mancini and Charis Papadopoulos
Department of Informatics, University of Bergen, N-5020 Bergen, Norway
{daniello, federico, charis}@ii.uib.no
www.ii.uib.no/~{daniello, federico, charis}

## ABSTRACT

The class of cographs is a well studied graph class that has been discovered in various fields independently, and a large number of papers have been published on it [1]. In this paper we investigate the problem of making an arbitrary graph into a cograph, adding edges to it. Any graph can be embedded into a cograph by adding edges to the original graph and the resulting graph is called a *cograph completion*, whereas the added edges are called *fill edges*. A cograph completion with the minimum number of edges is called *minimum*, and a cograph completion is called *minimal* if no proper subset of the fill edges result in a cograph when added to the original graph.

Computing a minimum completion of an arbitrary graph into a specific graph class is an important and well studied problem with applications in molecular biology, numerical algebra, and more generally areas involving graph modelling with missing edges due to lacking data[2]. Unfortunately minimum completions into most interesting graph classes, *including cographs*, are NP-hard to compute. However, as the set of minimum completions is a subset of the set of minimal completions, it is an interesting problem to study for which graph classes minimal completions can be computed in polynomial time. One can in fact search for a minimum among the set of the minimal ones.

What we prove is that the problem of computing a minimal cograph completion is polynomial time solvable. In particular we give linear time algorithms for the following two problems: 1. Computing a minimal cograph completion of an arbitrary input graph, and 2. Extracting a minimal cograph completion from an arbitrary cograph completion of the input graph. First we characterize minimal cograph completions, and then by exploiting this characterization and the structural properties of the unique tree representation (known as the *cotree*) of cographs, we give a linear-time algorithm for extracting a minimal cograph completion from any given one. Furthermore we give a vertex incremental algorithm to compute a minimal cograph completion directly from the input graph in time linear in the size of the computed graph (problem 1).

## References

[1] A. Brandstädt, V.B. Le, and J.P. Spinrad. *Graph Classes: A Survey.* SIAM Monographs on Discrete Mathematics and Applications, 1999.

[2] A. Natanzon, R. Shamir, and R. Sharan. Complexity classification of some edge modification problems. *Disc. Appl. Math.*, 113:109–128, 2001.

# Visualizing Source-Destination Relationships of Network Traffic in the Internet

Florian Mansmann

Department of Computer and Information Science, University of Konstanz, Germany
mansmannn@inf.uni-konstanz.de
http://infovis.uni-konstanz.de/~mansmann

## ABSTRACT

Network communication has become indispensable in business, education, and government. With the pervasive role of the Internet as a means of sharing information across networks, its misuse for destructive purposes, such as spreading malicious code, compromising remote hosts or damaging data through unauthorized access, has grown immensely in the recent years. The vast number of security incidents and other anomalies overwhelms attempts at manual analysis, especially when monitoring activity on service provider backbone links.

The classical way of monitoring the operation of large network systems is by analyzing the system logs for detecting anomalies. In this work, we present *Hierarchical Network Map*, an interactive visualization technique for analyzing network flow behavior by means of user-driven visual exploration. Our approach is meant as an enhancement to conventional analysis methods based on statistics or machine learning.

We superimpose a hierarchy on IP address space, and study the suitability of Treemap variants for each hierarchy level. Because viewing the whole IP hierarchy at once is not effective in most analysis tasks, we evaluate layout stability when eliding large parts of the hierarchy, while maintaining the visibility and ordering of the data of interest. A case study demonstrates how interactive visualization can be applied to gain deeper insight into large network traffic data sets.

The interdisciplinary approach integrating data warehouse technology, information visualization, and decision support, brings about the benefit of efficiently collecting the input data and aggregating over very large data sets, visualizing the results, and providing interactivity to facilitate analytical reasoning.

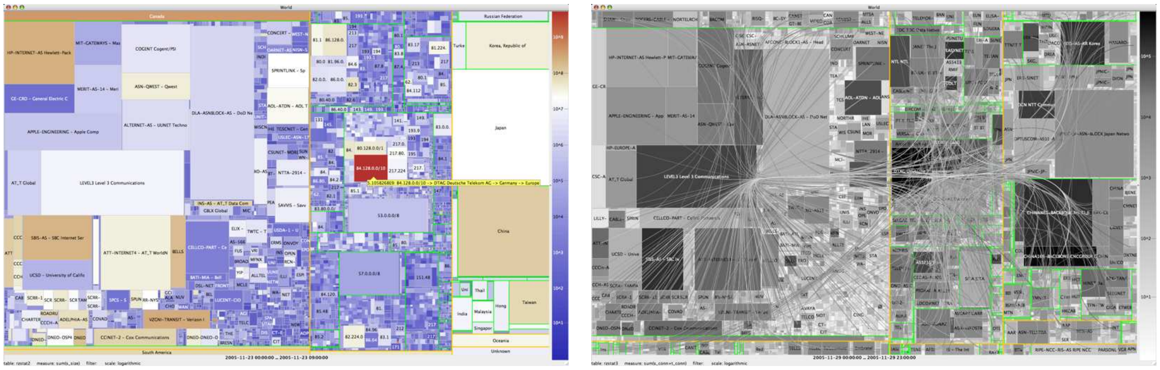


Figure 1: Multi-resolution HNMap approach (left) extended through edge bundles (right)

# A characterisation of the minimal triangulations of permutation graphs

Daniel Meister

Institutt for Informatikk, Universitetet i Bergen, Norway
Daniel.Meister@ii.uib.no
www.ii.uib.no/~danielm

## ABSTRACT

A minimal triangulation of a graph is a chordal graph obtained from adding an inclusion-minimal set of edges to the graph. For permutation graphs, i.e., graphs that are both comparability and cocomparability graphs, it is known that minimal triangulations are interval graphs. We (negatively) answer the question whether every interval graph is a minimal triangulation of a permutation graph. We give a non-trivial characterisation of the class of interval graphs that are minimal triangulations of permutation graphs and obtain as a surprising result that only "a few" interval graphs are minimal triangulations of permutation graphs.

A full version is available as a technical report [1].

## References

[1] Daniel Meister. A characterisation of the minimal triangulations of permutation graphs. *Reports in Informatics from Department of Informatics, University of Bergen*, report no. 350, 2007.

# Nonnegative Column and Column-Row Matrix Decompositions

Pauli Miettinen

Helsinki Institute for Information Technology
Department of Computer Science, University of Helsinki, Finland
pamietti@cs.helsinki.fi
www.cs.helsinki.fi/u/pamietti

## ABSTRACT

Matrix decompositions are important tools in data mining. In [1] we study two special types of matrix decompositions, viz. the nonnegative column and column-row decompositions, and the corresponding problems of finding such decompositions. We call these problems nonnegative CX and CUR problem, respectively.

Assume that we are given a nonnegative $n \times m$ matrix $\boldsymbol{A}$. In the *nonnegative CX problem* we are also given an integer $c$ and our goal is to find an $n \times c$ matrix $\boldsymbol{C}$ and a nonnegative $c \times m$ matrix $\boldsymbol{X}$ so that $\boldsymbol{C}$ contains $c$ columns of $\boldsymbol{A}$ and $\|\boldsymbol{A} - \boldsymbol{CX}\|_F$ is minimized. The *nonnegative CUR problem* resembles the CX problem, but we are also given another integer, $r$, and in addition to $\boldsymbol{C}$, the goal is to find an $r \times m$ matrix $\boldsymbol{R}$ containing rows of $\boldsymbol{A}$ and a nonnegative matrix $\boldsymbol{U}$ so as to minimize $\|\boldsymbol{A} - \boldsymbol{CUR}\|_F$. The hardness of the problems lies on selecting good $\boldsymbol{C}$ and $\boldsymbol{R}$; given them, matrices $\boldsymbol{X}$ and $\boldsymbol{U}$ are fairly easy to compute. Similar problems, but without the nonnegativity constraint, have been studied previously; see, e.g., [2]. We say that these are the *classical* CX and CUR problems.

The problems are motivated by the ease of interpretation of the decomposition: the columns of $\boldsymbol{C}$ and rows of $\boldsymbol{R}$ are as easy to interpret as is the original data, and imposing the nonnegativity constraint on $\boldsymbol{X}$ and $\boldsymbol{U}$ should increase the intrepretability of the results. Indeed, the presence of negative values in decompositions is often a major source of problems when one wants to interpret the factor matrices.

We propose two algorithms for solving the problems. Both algorithms can solve both problems, but they are based on different methods: one uses NMF while the other is a local swap algorithm. We compare the results given by our algorithms against the results given by algorithms designed for the classical CX and CUR problems. We show that our algorithms give good results, and that the results are also intuitively appealing.

This is a joint work with Saara Hyvönen and Evimaria Terzi.

## References

[1] S. Hyvönen, P. Miettinen, and E. Terzi. Nonnegative column and column-row matrix decompositions. To be submitted to PKDD'07.

[2] P. Drineas, M.W. Mahoney, and S. Muthukrishnan. Subspace sampling and relative-error matrix approximation: Column-row-based methods. In: ESA, 304–314, 2006.

# Mixed search number and linear-width of interval graphs and split graphs

Rodica Mihai

Department of Informatics, University of Bergen, Norway
rodica.mihai@ii.uib.no
www.ii.uib.no/~rodica

## ABSTRACT

We show that the mixed search number and the linear-width of interval graphs and of split graphs can be computed in linear time and in polynomial time, respectively.

Joint work with Fedor V. Fomin and Pinar Heggernes.

## References

[1] S. Alpern and S. Gal, *The theory of search games and rendezvous*, International Series in Operations Research & Management Science, 55, Kluwer Academic Publishers, Boston, MA, 2003.

[2] D. Bienstock, *Graph searching, path-width, tree-width and related problems (a survey)*, in Reliability of computer and communication networks (New Brunswick, NJ, 1989), vol. 5 of DIMACS Ser. Discrete Math. Theoret. Comput. Sci., Amer. Math. Soc., Providence, RI, 1991, pp. 33–49.

[3] D. Bienstock and P. Seymour, *Monotonicity in graph searching*, J. Algorithms, 12 (1991), pp. 239–245.

[4] K. S. Booth and G. S. Lueker, *Testing for the consecutive ones property, interval graphs, and graph planarity using pq-tree algorithms*, J. Comp. Syst. Sc., 13 (1976), pp. 335–379.

[5] J. Cohen, F. V. Fomin, P. Heggernes, D. Kratsch, and G. Kucherov, *Optimal linear arrangement of interval graphs.*, in Proceedings of MFCS 2006, Springer LNCS 4162, 2006, pp. 267–279.

[6] J. Díaz, J. Petit, and M. Serna, *A survey of graph layout problems*, ACM Computing Surveys, 34 (2002), pp. 313–356.

[7] F. V. Fomin, *A generalization of the graph bandwidth*, Vestnik St. Petersburg Univ. Math., 34 (2001), pp. 15–19 (2002).

[8] F. V. Fomin and D. M. Thilikos, *A 3-approximation for the pathwidth of halin graphs.*, J. Discrete Algorithms, 4 (2006), pp. 499–510.

# Similarity between Trajectories by using a Qualitative Feature Extraction Approach

Bart Moelans

Theoretical Computer Science
Hasselt University & Transnational University of Limburg, Belgium
bart.moelans@uhasselt.be
http://alpha.uhasselt.be/bart.moelans

## ABSTRACT

This poster will present an algorithm for extracting features from polylines (polygons) based on a qualitative cross formalism, the double-cross calculus. We will use these qualitative features to determine the degree of similarity between two polylines.

The algorithm first computes their generalized polygons, that consist of almost equally long line segments and that approximate the length of the given polylines within an $\varepsilon$-error margin. Next, the algorithm determines the double-cross matrices of the generalized polylines and the difference between these matrices is used as a measure of dissimilarity between the given polylines. We apply our method to query-by-sketch, indexing of polyline databases, and classification of terrain features and show experimental results for each of these applications.

A full version of this poster appeared in [1]. In this paper, we optimize the approach used in [2] and give a theoretical analyzis and complexity considerations of this idea. We also prove termination of our algorithm and show that its sequential time complexity is bounded by $O\left((\frac{max(N_1,N_2)}{\varepsilon})^2\right)$, where $N_1$ and $N_2$ are the number of vertices of the given polylines.

Joint work with Bart Kuijpers (Theoretical Computer Science, Hasselt University & Transnational University of Limburg, Belgium) and Nico Van de Weghe (Geography Department, Ghent University, Belgium)

## References

[1] Bart Kuijpers, Bart Moelans, and Nico Van de Weghe. Qualitative polyline similarity testing with applications to query-by-sketch, indexing and classification. In *GIS '06: Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems*, pages 11–18, New York, NY, USA, 2006. ACM Press.

[2] N. Van de Weghe, G. De Tré, B. Kuijpers, and Ph. De Maeyer. The double-cross and the generalization concept as a basis for representing and comparing shapes of polylines. In R. Meersman et al., editor, *Proceedings of the 1st International Workshop on Semantic-based Geographical Information Systems (SeBGIS'05)*, volume 3762 of *Lecture Notes in Computer Science*, pages 1087–1096. Springer, 2005.

# Perceptron vs. Winnow in the Boolean Domain

Krishnan Narayanan

Department of Computer Science, P.O Box 68, 00014 University of Helsinki, Finland
krishnan.narayanan@cs.helsinki.fi

## ABSTRACT

Learning an unknown linear threshold function is one of the major recurring themes in computational learning theory and practice. For learning linear threshold functions, the classical perceptron algorithm and the more recent winnow algorithm are two classes of online algorithms based on additive and multiplicative updates respectively. The perceptron and the winnow algorithms can be regarded as belonging to a generic family of algorithms based on variable norm. A unified analysis of their worst case mistake bounds can be given using amortized analysis with the notion of Bregman divergences as suitable potential functions.

A remarkable feature of the winnow algorithm is that the dependence of its mistake bound is logarithmic in the number of irrelevant attributes in contrast to the linear dependence in the case of the perceptron algorithm [1]. This makes the winnow algorithm attractive in learning environments in which the concept to be learned depends only on a relatively small number of the total attributes that are present. Such algorithms are called attribute efficient and one of the important goals in computational learning theory is to find attribute efficient algorithms for various concept classes.

The expressiveness of linear threshold functions can be increased by extending the feature space either explicitly or implicitly via kernels. As boolean disjunctions can be viewed as linear threshold functions it suggests the possibility to use perceptron and winnow algorithms to learn interesting boolean function classes. The contrasting trade-off between generalization error and efficiency of the perceptron and the winnow algorithms is known in the literature [2].

We attempt to answer the interesting question whether maximizing the 1-norm margin instead of the 2-norm margin by an online learning algorithm can yield an efficient algorithm for learning sparse disjunctions in the statistical learning framework.

## References

[1] J. Kivinen, M.K. Warmuth and P. Auer. The Perceptron algorithm vs. Winnow. linear vs. logarithmic mistake bounds when few input variables are relevant. *Special issue of Artificial Intelligence, on Relevance*, 97(1-2):325–343, December 1997.

[2] R. Khardon, D. Roth and R.A. Servedio. Efficiency versus Convergence of Boolean Kernels for On-Line Learning Algorithms. *Journal of Artificial Intelligence Research*, 24: 341–356, 2005.

# Visualization of Spatial Data Algorithms

Jussi Nikander

Department of Computer Science, Helsinki University of Technology, Finland
jtn@cs.hut.fi
http://www.cs.hut.fi/u/jtn/

## ABSTRACT

Handling multidimensional data requires specialized data structures and algorithms. In the context of geoinformatics, a branch of science that applies information technology methods to cartography and other geosciences, these are known as *spatial data structures* and *algorithms*. A natural way of representing spatial data are maps and diagrams, and therefore visualization plays an important part in the field. Geographic information systems (GIS), for example, are an important tool in geoinformatics that use visualization for representing information.

The visualizations used in GIS are typically concerned with representing the data in an understandable and useful manner, and the details of the underlying data structures and algorithms are uniteresting and therefore omitted. However, for students learning the field, or for developers trying to implement new problem solving methods, data structure visualizations and algorithm animations can be a powerful tool. Data structures and algorithms are abstract and complicated concepts, and often hard to understand. Visualization can make them more concrete and help a student to learn how an algorithm works, how a data structure is organized in a given situation.

Typically, algorithm visualization is used to show how the internal connections and hierarchy between the parts of a data structure are organized. Most of the time this is accomplished using widely–accepted *canonical views* immediately recognizable for anyone familiar with the field [1]. However, when the data being visualized is multidimensional, such visualizations are not adequate for showing relationships between data items. If, for example, two–dimensional points are stored in a tree, the typical tree view is not suited for showing the distances between the points.

In order to show how multidimensional data items are related to one other, more abstract *representation level views* of data structures are required [1]. For two–dimensional data a good representation is an *area*. An area covers a relevant part of the two–dimensional plane sufficient for showing all the data items in the structure being visualized. The visualization also includes some information about the data structure. For example, if each node of a data structure covers some subarea of the visualization, these subareas can visualized as polygons.

## References

[1] J. Nikander, A. Korhonen, E. Valanto, K. Virrantaus Visualization of Spatial Data Structures on Different Levels of Abstraction *Proceedings of the Fourth Program Visualization Workshop*, 60–66, 2006.

# Succinct Oracles for Exact Distances in Undirected Unweighted Graphs

Igor Nitto

Department of Computer Science, University of Pisa, Italy
nitto@di.unipi.it
www.di.unipi.it/~nitto

## ABSTRACT

Various authors have investigated the problem of designing succinct graph encodings for supporting the retrieval of either the *adjacency list* of a node (see [2] and references therein), or the *approximate distance* between node pairs in various types of graphs (see [1] and references therein). Specifically, [1] introduced a class of data structures called *approximate distance oracles* for approximating shortest-path distances in a general weighted graph $G$ up to a multiplicative *stretch factor*. For any fixed $k$, they present a distance oracle answering approximated distance in $O(k)$ time within $2k-1$ stretch factor. Their data structure takes $O(kn^{1+1/k}\log n)$ bits of storage, where $n$ is the number of nodes in $G$. This has been later improved for planar digraph.

When *exact* distances are needed, it is still open whether it is possible to obtain a distance oracle using substantially less space than the plain encoding of the distance matrix $\mathbf{D}$ of the graph $G$.

In our paper we solve this problem for unweighted graphs by providing a succinct oracle for unweighted and undirected graphs which requires $(\log_2 3)\,\frac{n^2}{2}+o(n^2)$ bits of storage and allows calculation of any node-pair distance in $O(1)$ time. A simple counting argument shows a lower bound of $\frac{n^2}{2}$ bits therefore our solution is *asymptotically optimal* and asymptotically better than the plain encoding of the distance matrix $\mathbf{D}$, requiring $O(n^2)$ memory words or $O(n^2\log n)$ bits.

Stated in other terms we propose a succinct storage scheme for $\mathbf{D}$ which takes a constant number of bits per entry and still allows to retrieve an element of $\mathbf{D}$ in constant time. Joint work with Paolo Ferragina and Rossano Venturini, (ferragin@di.unipi.it,rventurini@di.unipi.it, University of Pisa).

## References

[1] M. Thorup, U. Zwick. Approximate distance oracles. In *STOC*, pages 183-192, 2001. In STOC, pages 183-192, 2001

[2] I. Munro and V. Raman. Succinct representation of balanced parentheses, static trees and planar graphs. In *Proc. of the 38th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 118-126, 1997.

# Narrow proofs may be spacious: Separating space and width in resolution

Jakob Nordström

Royal Institute of Technology (KTH)
SE-100 44 Stockholm, Sweden
jakobn@kth.se

20th April 2007

## ABSTRACT

The width of a resolution proof is the maximal number of literals in any clause of the proof. The space of a proof is the maximal number of clauses kept in memory simultaneously if the proof is only allowed to infer new clauses from clauses currently in memory.

Both of these measures have previously been studied and related to the resolution refutation size of unsatisfiable CNF formulas. Also, the minimum refutation space of a formula has been proven to be at least as large as the minimum refutation width, but it has been open whether space can be separated from width or the two measures coincide asymptotically.

We prove that there is a family of $k$-CNF formulas for which the refutation width in resolution is constant but the refutation space is non-constant, thus solving a problem mentioned in several previous papers.

Our result has appeared as [2]. A full version of this paper is available as [1].

## References

[1] J. Nordström. Narrow proofs may be spacious: Separating space and width in resolution. Technical Report TR05-066, Revision 02, Electronic Colloquium on Computational Complexity (ECCC), Nov. 2005. Available at `http://www.eccc.uni-trier.de/eccc/`.

[2] Jakob Nordström. Narrow proofs may be spacious: Separating space and width in resolution (Extended abstract). In *Proceedings 38th Annual ACM Symposium on Theory of Computing (STOC '06)*, pages 507–516, May 2006.

# Modular Answer Set Programming

Emilia Oikarinen

Department of Computer Science and Engineering
Helsinki University of Technology, Finland
emilia.oikarinen@tkk.fi
www.tcs.tkk.fi/~eoikarin

## ABSTRACT

Answer set programming is a declarative rule-based constraint programming paradigm. In answer set programming (ASP) the problem at hand is solved declaratively by writing down a logic program the answer sets of which correspond to the solutions of the problem, and then computing the answer sets of the program using a special purpose search engine. Currently ASP programs are typically viewed as integral entities. This becomes problematic as programs become more complex, and the sizes of program instances grow. In ASP there is a lack of mechanisms, available in other modern programming languages, that ease program development by allowing re-use of code or breaking programs into smaller pieces, modules. Even though modularity has been studied extensively in conventional logic programming, there are only few approaches how to incorporate modularity into ASP.

We accommodate Gaifman and Shapiro's program modules [1] in conventional logic programming to the context of ASP resulting in a simple and intuitive notion for *ASP program modules under the stable model semantics*. A module interacts through an input/output interface, and full compatibility of the module system and the stable model semantics is achieved by restricting positive recursion between modules. One of the main results is a module theorem showing that the ASP module system is compositional with respect to the stable model semantics, i.e., module-level stability implies program-level stability, and vice versa, as long as the answer sets of the submodules are compatible. We also introduce a notion *modular equivalence*, that is a proper congruence relation for composition of modules.

The results appear in [3, 4, 5] and are joint work with Tomi Janhunen. The module system is later extended to the case of disjunctive program in [2].

## References

[1] H. Gaifman and E. Y. Shapiro. Fully abstract compositional semantics for logic programs. In *Proc. ACM-POPL'89*, pp. 134–142. ACM Press.

[2] T. Janhunen, E, Oikarinen, H. Tompits, and S. Woltran. Modularity aspects of disjunctive stable models. In *Proc. LPNMR'07*. To appear.

[3] E. Oikarinen. Modular answer set programming. Research Report A106, Helsinki University of Technology, Laboratory for Theoretical Computer Science, Espoo, Finland, December 2006.

[4] E. Oikarinen. Modularity in SMODELS programs. In *Proc. LPNMR'07*. To appear.

[5] E. Oikarinen and T. Janhunen. Modular equivalence for normal logic programs. In *Proc. ECAI'06*, pp. 412–416. IOS Press.

# EFFICIENT DATA STREAM FILTERING

Panu Silvasti

Department of Computer Science, Helsinki University of Technology, Finland
panu.silvasti@tkk.fi
www.cs.hut.fi/~psilvast

## ABSTRACT

One important and popular research subject in data stream management has been publish-subscribe systems based on XML document filtering, e.g. [1, 3, 4]. In XML publish-subscribe system users specify their interest in profiles that are expressed in XPath [2] language. The system processes a stream of XML documents and delivers notification or content to interested subscribers. The number of interested users and stored profiles can be very large, thousands or millions.

Recently publish-subscribe systems have appeared for every day use, such as *Google alerts* and stock information delivery by *Yahoo.com*. In [5] XML document filtering techniques have been applied in routing real-time air traffic control data.

I am studying and developing efficient algorithms for XML document filtering. The primary problem to be addressed is defined as the *filtering problem*: Given a set of XPath expressions (XPEs), identify those expressions that match a given XML document.

The XPath profiles are stored into the system, which should have nearly constant throughput for filtering, regardless of the number of profiles. Also the memory usage of the algorithm should be feasible, even when the number of profiles is large.

## References

[1] Mehmet Altinel and Michael J. Franklin. Efficient filtering of XML documents for selective dissemination of information. In *VLDB*, pages 53–64, 2000.

[2] Anders Berglund, Scott Boag, Don Chamberlin, Mary F. Fernandez, Michael Kay, Jonathan Robie, , and Jrme Simon. XML path language (xpath) 2.0 w3c working draft 16. technical report wd-xpath20-20020816. Technical report, World Wide Web Consortium, 2002.

[3] Yanlei Diao, Mehmet Altinel, Michael J. Franklin, Hao Zhang, and Peter Fischer. Path sharing and predicate evaluation for high-performance XML filtering. *ACM Trans. Database Syst.*, 28(4):467–516, 2003.

[4] Joonho Kwon, Praveen Rao, Bongki Moon, and Sukho Lee. FiST: scalable XML document filtering by sequencing twig patterns. In *VLDB '05: Proceedings of the 31st international conference on Very large data bases*, pages 217–228. VLDB Endowment, 2005.

[5] Alex C. Snoeren, Kenneth Conley, and David K. Gifford. Mesh-based content routing using XML. In *SOSP '01: Proceedings of the eighteenth ACM symposium on Operating systems principles*, pages 160–173, New York, NY, USA, 2001. ACM Press.

# Drawing Large Graphs Using Progressive Multidimensional Scaling

Christian Pich

Department of Computer & Information Science, University of Konstanz, Germany
christian.pich@uni-konstanz.de
http://www.inf.uni-konstanz.de/~pich/

## ABSTRACT

We present a novel sampling-based approximation technique for classical multidimensional scaling that yields an extremely fast layout algorithm suitable even for very large graphs [1]. It produces layouts that compare favorably with other methods for drawing large graphs, and it is among the fastest methods available. In addition, our approach allows for progressive computation, i.e. a rough approximation of the layout can be produced even faster, and then be refined until satisfaction.

## References

[1] U. Brandes and C. Pich. Eigensolver Methods for Progressive Multidimensional Scaling of Large Data. *Proc. 14th Intl. Symp. Graph Drawing (GD '06).* LNCS 4372, pp. 42–53. Springer-Verlag, 2007.

# Trajectory Pattern Mining

Fabio Pinelli

Department of Engeneering of Information, University of Pisa, Italy

KDD lab ISTI - CNR- Pisa, Italy

fabio.pinelli@isti.cnr.it
www-kdd.isti.cnr.it/˜pinelli

## ABSTRACT

The increasing pervasiveness of location-acquisition technologies (GPS, GSM networks, etc.) is leading to the collection of large spatio-temporal datasets and to the opportunity of discovering usable knowledge about movement behaviour, which fosters novel applications and services. In this paper, we move towards this direction and develop an extension of the sequential pattern mining paradigm [3] with temporal annotations [1] that analyzes the trajectories of moving objects. We introduce trajectory patterns (*T-Pattern*) as concise descriptions of frequent behaviours, in terms of both space (i.e., the regions of space visited during movements: *ROI*) and time (i.e., the duration of movements: annotations). In this setting, we provide a general formal statement of the novel mining problem and then study several different instantiations of different complexity.

We defined three different kinds of *T-Patterns*. The first one (*static*) takes in input predefined *ROI*'s, then, after a translation of the sequences of points in sequence of *ROI*'s, we run the MiSTA algorithm [1]. The second is still a static version of the algorithm and it computes the first set of Regions of Interest which are the most popular regions over a grid applied to the space. As previously, we translate the sequences of points as sequence of *ROI*'s. Finally we run MiSTA algorithm. The last algorithm is a dynamic version, it is based on MiSTA. We compute the set of *ROI*'s taking in account only the trajectories belong in a projected databases, being based on PrefixSpan algorithm [2], and any time we apply the translation procedure. For each algorithm we extract the most frequent *T-patterns* as sequences of *ROI*'s and typical transition time.

The various approaches are then empirically evaluated over real data and synthetic benchmarks, comparing their strengths and weaknesses. Full version of this paper is submitted to *KDD07*. This work is jointly done with F. Giannotti, M. Nanni and D. Pedreschi.

## References

[1] F.Giannotti, N.Nanni, D.Pedreschi. Mining annotated Sequences *Proc. SIAM Conference on Data Mining,*, pages 346 - 357. SIAM, 2006.

[2] J. Pei et al. Prefixspan: Mining sequential patterns by prefix-projected growth. *In ICDE*, pages 215225, 2001.

[3] R. Agrawal and R. Srikant. Mining sequential patterns. *In Proceedings of ICDE, 1995.*

# Mining statistical information of fault-tolerant itemsets in transactional database

Ardian Kristanto Poernomo

Nanyang Technological University, Singapore
ardi0002@ntu.edu.sg

## ABSTRACT

This work addresses the problem of mining frequent fault-tolerant itemsets (FT-itemsets) in transactional database [5, 2]. We formulate the general definition of FT-itemset in terms of basic constraints, and show how existing and new variants of this problem can be derived from combinations of those constraints. In the general FT-itemset mining problem, an itemset may be supported by a large number of maximal support transaction-sets, i.e. exponential to the number of transactions. Therefore, complete search technique [4] is feasible only in small databases. Previous approaches to solve this problem either relax the constraint [2], tighten the constraint [1], or perform approximation [5]. As the transactions are not important in itemset mining, we propose to mine only the statistical information – size, maximum value, and mean, of the support transaction-sets.

We design a new framework for mining statistical information of FT-itemsets, *$ILP_{FT}$ framework*, based on backtracking algorithm, Integer Linear Programming (ILP), and aggregation statistics, and demonstrate its application for all problem variants. Experimental studies show that our proposed technique increases the efficiency of the mining process by more than an order of magnitude, compared to complete search technique.

## References

[1] J. Besson, R. G. Pensa, C. Robardet, and J-F. Boulicaut. Constraint-based mining of fault-tolerant patterns from boolean data. In *Proc. of KDID*, pages 55–71, 2005.

[2] J. Pei, A. K. H. Tung, and J. Han. Fault-tolerant frequent pattern mining: Problems and challenges. In *Proc. of DMKD*, 2001.

[3] A. K. Poernomo, and V. Gopalkrishnan. Mining statistical information of fault-tolerant itemset in transactional database. Technical Report, Centre for Advanced Information Systems, Nanyang Technological University, Singapore, November 2006, *www.cais.ntu.edu.sg/~vivek/pubs/TR-1106.pdf*

[4] K. Sim, J. Li, V. Gopalkrishnan, and G. Liu. Mining maximal quasi-bicliques to co-cluster stocks and financial ratios for value investment. In *Proc. of ICDM*, 2006.

[5] C. Yang, U. M. Fayyad, and P. S. Bradley. Efficient discovery of error-tolerant frequent itemsets in high dimensions. In *Proc. of SIGKDD*, pages 194–203, 2001.

# Integrating Pattern Mining in Relational Databases

Adriana Prado

Department of Computer Science, University of Antwerp, Belgium
adriana.prado@ua.ac.be
www.adrem.ua.ac.be/~aprado

## ABSTRACT

Almost a decade ago, Imielinski and Mannila introduced the concept of an *Inductive Database*, in which a *Knowledge and Data Discovery Management System* (KDDMS) manages KDD applications just as DBMSs successfully manage business applications. Basically, besides allowing the user to query the data, the KDDMS should also give users the ability to query patterns or models extracted from these data. In this context, several researchers proposed extensions to the popular SQL, as a natural way to express such mining queries.

In our work, we aim at extending the DBMS itself, not the query language. That is, we propose an approach in which the user can query the collection of all possible patterns as if these are stored in relational tables. The main challenge is how this storage can be implemented effectively. After all, the amount of all possible patterns can be extremely large and impractical to store. To solve this problem, we propose to keep these pattern tables virtual. That is, as far as the user is concerned, all possible patterns are stored, but on the physical layer, no such complete tables exist. Instead, whenever the user queries such a pattern table, or *virtual mining view*, an efficient data mining algorithm is triggered by the DBMS, which materializes at least those tuples needed to answer the query. Afterwards, the query can be executed as if the patterns were there all the time. Notice that the user can now query mining results by using a standard relational query language, such as SQL. Of course, this assumes the user poses constraints in his or her query, asking for only a subset of all possible patterns, which should then be detected and exploited by the data mining algorithm. As a first step towards this goal, we propose a constraint extraction algorithm starting from a collection of simple constraints. To extract the constraints from a given SQL-query, the algorithm works on an equivalent expression tree of relational algebra. Starting from the leaves, going bottom-up until the root, it determines for every node $n$ in the expression tree which tuples should be in the views in order to answer to the *subquery* associated with that node, that is, the query represented by the subtree rooted at $n$.

This is a joint work with Toon Calders, from Eindhoven University of Technology, and Bart Goethals, from the University of Antwerp. The full version of this work appears in [1], where we show how this approach can be implemented for the popular association rule and frequent set mining problems.

## References

[1] T. Calders, B. Goethals and A. Prado. Integrating Pattern Mining in Relational Databases. *PKDD Conference*, LNCS, Springer, 454–461, 2006.

# Residual Variance for Dependency Measure

Nima Reyhani Amaury Lendasse

Lab of Computer and Information Science,
Department of Computer Science and Engineering
Helsinki University of Technology, Finland
nreyhani@cis.hut.fi

## ABSTRACT[1]

In this work, the residual variance estimation [1] is applied as an empirical dependency measure. The main idea is to use the notion of predictability as a basis for dependency definition. Considering any nonlinear regression function between two random variables, the power of regression residuals or residual variance defines the dependency between random variables. The residuals variance can be directly computed by estimators without finding the best curve fit. The conditions on which two random variables are independent according to the estimated residuals variance are discussed are also discussed.

In comparision, the residual varaince can not acheive the accuracy lower bound of Mutual Information, however, practically, the residual variance shows more accurate estimation when the data samples is very small compare to the dimensionality of the problem.

The dependency measure finds wide areas of applications in signal processing and machine learning. In this work, residual variance estimation is applied for Independent Component Analysis [2] and input selection.

## References

[1] A. Munk, N. Bissantz and Freitag, G. On difference-based variance estimation in nonparamteric regression when the covariate is high dimensional. *Journal of the Royal Statistical Society B(Methodological) 67* (2005), 901–919.

[2] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis.* John Wiley and Sons, 2001.

[1] This work continues within a Biomedical Signal Processing framework under supervision of Erkki Oja and Ricardo Vigario.

# New Solutions to Basic Searching Problems

Milan Ružić

Computational Logic and Algorithms Group, IT University of Copenhagen
milan@itu.dk
www.itu.dk/~milan

## ABSTRACT

Searching problems are among fundamental issues in computer science, being of great practical and theoretical interest. The primary focus of our research are the dictionary problem, and related to it – the prefect hashing problem. A dictionary stores a subset of some universe and it can answer membership queries on the set. If a search is successful, then some information associated with the searched key may be returned. Some dictionaries also support more general predecessor queries.

The dictionary problem has been well studied, and many solutions have been given, going beyond classical comparison-based structures. The solutions offer different characteristics regarding space usage, time bounds, model of computation, and universe in question. A challenge is to simultaneously achieve good time and space requirements, without constraints on the universe size. The understanding of the complexity of *deterministic* solutions to the dictionary problem is far from complete. An important segment of our research is devising new and more efficient algorithms for selecting "good" hash functions deterministically. Although randomized dictionaries reached a stage of high development there still are some open questions.

In our research we found: a new type of hash functions useful in a couple of settings [3], a new insight into a well-known family of functions [4], a use of expander graphs to get efficient dictionaries in the parallel disk model [1], an analysis of the classical linear probing scheme with realistic hash functions [2].

## References

[1] Mette Berger, Esben Rune Hansen, Rasmus Pagh, Mihai Patrascu, Milan Ružić, and Peter Tiedemann. Deterministic load balancing and dictionaries in the parallel disk model. In *Proceedings of the 18th Annual ACM Symposium on Parallel Algorithms and Architectures*, pages 299–307. ACM, 2006.

[2] Anna Pagh, Rasmus Pagh, and Milan Ružić. Linear probing with constant independence. In *Proceedings of the 39th ACM Symposium on Theory of Computing*, 2007.

[3] Milan Ružić. Making deterministic signatures quickly. In *Proceedings of the 18th ACM-SIAM Symposium on Discrete Algorithms*, pages 900–909. ACM and SIAM, 2007.

[4] Milan Ružić. Uniform deterministic dictionaries. Submitted manuscript. Very preliminary version appeared as: Uniform Algorithms for Deterministic Construction of Efficient Dictionaries. In *12th Ann. European Symposium on Algorithms (ESA '04).*

# Bulk Updates in Search Trees

Riku Saikkonen

Department of Computer Science and Engineering
Helsinki University of Technology, Finland
rjs@cs.hut.fi

## ABSTRACT

I present some results from my research on bulk update operations in AVL and B-trees. Bulk updates (bulk insertion and bulk deletion) insert or delete a set of keys in a single operation, using more efficient rebalancing than with the individual insertion and deletion operations. One motivation for bulk updates is that in some applications of databases, updates to indices commonly occur in groups, and bulk updates give much more effective methods of performing the updates than if the updates were processed one at a time.

One subtopic in my research is that of efficient algorithms for bulk updates in AVL trees. This research is based on earlier work presented in [2]. When certain conditions hold, the number of rotations needed for a bulk update is worst-case logarithmic in the size of the bulk. We have performed a detailed complexity analysis that also considers constant factors and the amortized complexity of sequences of bulk operations. I also present some results from an extensive set of experiments done on the algorithms.

Another subtopic is how to perform bulk deletion efficiently in a database setting, specifically in a B-tree primary index of a large database table. Here, methods for efficient concurrency control and recovery are as important as the actual bulk deletion algorithm. Our paper [1] gives an efficient locking protocol that supports bulk deletion using logical partitioning of the key space. Our recoverable bulk delete algorithm works in two phases: a scan phase which acquires all necessary locks and marks records for deletion, and a rebalance phase where the records are actually deleted. One factor that makes our algorithm efficient is that the rebalance phase does not need to visit subtrees in which every record is to be deleted.

## References

[1] Timo Lilja, Riku Saikkonen, Seppo Sippu and Eljas Soisalon-Soininen. Online bulk deletion. In *IEEE 23rd International Conference on Data Engineering* (*ICDE 2007*). IEEE Computer Society, 2007.

[2] Eljas Soisalon-Soininen and Peter Widmayer. Amortized complexity of bulk updates in AVL-trees. In *8th Scandinavian Workshop on Algorithm Theory* (*SWAT 2002*), Lecture Notes in Computer Science vol. 2368. Springer–Verlag Berlin Heidelberg, 2002.

# Algorithms with $q$-Grams for Two String Matching Problems

Leena Salmela

Department of Computer Science and Engineering,
Helsinki University of Technology, Finland
lsalmela@cs.hut.fi
http://www.cs.hut.fi/u/lsalmela/

## ABSTRACT

**Multiple pattern matching** We present three algorithms for exact string matching of multiple patterns with very large pattern sets. Such methods are needed in anti-virus scanning, intrusion detection, content scanning and filtering and in specific data mining problems. Currently a tool for biologists is under development based on our algorithms. Our algorithms are filtering methods, which apply $q$-grams and bit parallelism.

The HG algorithm builds a table which is used to check if a given $q$-gram appears in any of the patterns in a given position. These are used in the right-to-left scanning of a text window to calculate the shift. The SOG algorithm transforms the set of patterns into a single generalized pattern that contains classes of $q$-grams. This generalized pattern is then used by the shift-or algorithm to find the candidate matches. The BG algorithm uses the BNDM algorithm to search with the generalized pattern. HG and BG are sublinear on average while SOG is linear on average.

We ran extensive experiments with the algorithms and compared them with various versions of earlier algorithms, e.g. different trie implementations of the Aho-Corasick algorithm. All of our algorithms showed to be substantially faster than earlier solutions for sets of 1,000-10,000 patterns and the good performance of two of them continues to 100,000 patterns. The gain is due to the improved filtering efficiency caused by $q$-grams. Further details can be found in [1] which is joint work with Jorma Tarhio and Jari Kytöjoki.

**Parameterized pattern matching** Two strings parameterize match if there is a bijection defined on the alphabet that transforms the first string character by character into the second string. We present algorithms that solve this problem in sublinear time on average for moderately repetitive patterns. Our algorithms are based on the Boyer-Moore-Horspool algorithm for exact matching. In our algorithms the shift is calculated based on the last $q$ characters of the previously examined text window. After the shift these characters will be aligned so that they parameterize match with the pattern. We present algorithms for both one and two dimensional parameterized matching. Further details can be found in [2] which is joint work with Jorma Tarhio.

## References

[1] L. Salmela, J. Tarhio, J. Kytöjoki. Multi-pattern string matching with $q$-grams. *ACM Journal of Experimental Algorithmics* 11:1–19, 2006.

[2] L. Salmela, J. Tarhio. Sublinear algorithms for parameterized matching. In: *Proc. CPM'06, LNCS 4009*, 254–264, 2006.

# Distributed Optimisation Algorithms for Multihop Wireless Networks

André Schumacher

Laboratory for Theoretical Computer Science,
Helsinki University of Technology, Finland
Andre.Schumacher@tkk.fi
www.tcs.hut.fi/~schumach/

## ABSTRACT

Due to their wireless and decentralised nature, ad hoc networks are attractive in a variety of applications. However, these properties also pose significant challenges to their developers. In cases where conventional wired networks usually rely on some kind of centralised entity, ad hoc network nodes have to cooperate in a distributed and self-organising manner. Additional side constraints, such as energy consumption, have to be taken into account as well. The radio transmission channel is limited in bandwidth and shared among nearby nodes. Two properties of algorithms are particularly desirable in an ad hoc context. First, an algorithm should be mathematically justified. Second, an algorithm should be distributed and non-hierarchical. Each node should follow a simple set of rules to cooperate in computing the optimum. Neither the size nor the number of messages should grow rapidly with the size of the network. We address the problem of load balancing in wireless multi-hop networks by applying a network flow optimisation algorithm. A mathematical model of the underlying problem is formulated as a linear program and a distributed approximation algorithm is implemented in the form of a multipath routing protocol. The algorithm is based on shortest-path computations that are integrated into a source-routing protocol. Our simulations show a gain of 14% to 69% in the throughput, depending on the setup, compared to a standard routing protocol for high network load. The full version of this paper appears in [1, 2]. This research has been joint work with Harri Haanpää, Satu Elisa Schaeffer, and Pekka Orponen. The simulations were performed in collaboration with Shreyas Prasad.

## References

[1] Shreyas Prasad, André Schumacher, Harri Haanpää, and Pekka Orponen. Balanced multipath source routing. In *Proceedings of the 21st International Conference on Information Networking (ICOIN'07, Estoril, Portugal, January 2007)*, to appear.

[2] André Schumacher, Harri Haanpää, Satu Elisa Schaeffer, and Pekka Orponen. Load balancing by distributed optimisation in ad hoc networks. In J. Cao, I. Stojmenovic, X. Jia, and S. K. Das, editors, *Mobile Ad-hoc and Sensor Networks*, volume 4325/2006 of *Lecture Notes in Computer Science*, pages 873–884, Berlin / Heidelberg, 2006. Springer-Verlag.

# EFFICIENT DATA STREAM FILTERING

Panu Silvasti

Department of Computer Science, Helsinki University of Technology, Finland
panu.silvasti@tkk.fi
www.cs.hut.fi/~psilvast

## ABSTRACT

One important and popular research subject in data stream management has been publish-subscribe systems based on XML document filtering, e.g. [1, 3, 4]. In XML publish-subscribe system users specify their interest in profiles that are expressed in XPath [2] language. The system processes a stream of XML documents and delivers notification or content to interested subscribers. The number of interested users and stored profiles can be very large, thousands or millions.

Recently publish-subscribe systems have appeared for every day use, such as *Google alerts* and stock information delivery by *Yahoo.com*. In [5] XML document filtering techniques have been applied in routing real-time air traffic control data.

I am studying and developing efficient algorithms for XML document filtering. The primary problem to be addressed is defined as the *filtering problem*: Given a set of XPath expressions (XPEs), identify those expressions that match a given XML document.

The XPath profiles are stored into the system, which should have nearly constant throughput for filtering, regardless of the number of profiles. Also the memory usage of the algorithm should be feasible, even when the number of profiles is large.

## References

[1] Mehmet Altinel and Michael J. Franklin. Efficient filtering of XML documents for selective dissemination of information. In *VLDB*, pages 53–64, 2000.

[2] Anders Berglund, Scott Boag, Don Chamberlin, Mary F. Fernandez, Michael Kay, Jonathan Robie, and Jrme Simon. XML path language (xpath) 2.0 w3c working draft 16. Technical report wd-xpath20-20020816. Technical report, World Wide Web Consortium, 2002.

[3] Yanlei Diao, Mehmet Altinel, Michael J. Franklin, Hao Zhang, and Peter Fischer. Path sharing and predicate evaluation for high-performance XML filtering. *ACM Trans. Database Syst.*, 28(4):467–516, 2003.

[4] Joonho Kwon, Praveen Rao, Bongki Moon, and Sukho Lee. FiST: scalable XML document filtering by sequencing twig patterns. In *VLDB '05: Proceedings of the 31st international conference on Very large data bases*, pages 217–228. VLDB Endowment, 2005.

[5] Alex C. Snoeren, Kenneth Conley, and David K. Gifford. Mesh-based content routing using XML. In *SOSP '01: Proceedings of the eighteenth ACM symposium on Operating systems principles*, pages 160–173, New York, NY, USA, 2001. ACM Press.

# Network-Oblivious Algorithms

Francesco Silvestri

Department of Information Engineering, University of Padova, Italy
francesco.silvestri@dei.unipd.it
www.dei.unipd.it/~silvest1

## ABSTRACT

Communication is a major factor determining the performance of algorithms on current parallel computing systems. Reducing the communication requirements of algorithms is then of paramount importance, if they have to run efficiently on physical machines. Recognition of this fact has motivated a large body of results in algorithm design and analysis, but these results do not yet provide a coherent and unified theory of the communication requirements of computations. One major obstacle toward such a theory lies in the fact that communication is defined only with respect to a specific mapping of a computation onto a specific machine structure. Furthermore, the impact of communication on performance depends on the latency and bandwidth properties of the machine. In this scenario, algorithm design, optimization, and analysis can become highly machine dependent, which is undesirable from the economical perspective of developing efficient and portable software.

It is natural to wonder whether algorithms can be designed that, while independent of any machines, are nevertheless efficient for a wide set of machines. In other words, we are interested in exploring the world of efficient *network-oblivious* algorithms, in the same spirit as the exploration of *cache-oblivious* algorithms [2]. We develop a framework where the concept of network-obliviousness and of algorithmic efficiency are precisely defined. In this framework, a network-oblivious algorithm is designed in a model of computation where the only parameter is the problem's input size. Then, the algorithm is evaluated on a model with two parameters, capturing parallelism and granularity of communication. We show that, for a wide class of network-oblivious algorithms, optimality in the latter model implies optimality in a block-variant of the Decomposable BSP model, which effectively describes a wide class of parallel platforms. We illustrate our framework by providing optimal network-oblivious algorithms for a few key problems, and also establish some negative results.

This abstract is based on results appeared in [1].

## References

[1] G. Bilardi, A. Pietracaprina, G. Pucci, and F. Silvestri. Network-Oblivious Algorithms. In *Proc. of 21st International Parallel and Distributed Processing Symposium*, 2007.

[2] M. Frigo, C. Leiserson, H. Prokop, and S. Ramachandran. Cache-Oblivious Algorithms. In *Proc. of 40th Symp. on Foundations of Computer Science*, 1999.

# Counting minimum weighted dominating sets

Alexey A. Stepanov

Department of Informatics, University of Bergen, Bergen, Norway
ljosha@ljosha.org
ljosha.org

## ABSTRACT

The dominating set problem is one of the classical NP-complete graph optimization problem which fits into the broader class of domination and covering problems. Hundreds of papers have been written on them (see e.g. the survey [1] by Haynes et al.). The fastest known algorithm computes a minimum dominating set of a graph in time $\mathcal{O}(1.5137^n)$ [2]. The algorithm from [2] cannot be used to compute a dominating set of minimum weight (in a weighted graph), however, as it was observed in [3], the problem can be solved in time $\mathcal{O}(1.5780^n)$ by similar techniques. In the same paper it was shown that all minimal dominating sets can be listed in time $\mathcal{O}(1.7697^n)$ (later improved to $\mathcal{O}(1.7170^n)$), which implies that minimum dominating sets can be counted in this time.

In this paper we give an algorithm that counts minimum weight dominating sets in a weighted graph on $n$ vertices in time $\mathcal{O}(1.5535^n)$. The basic idea is as follows: First we turn the instance of the dominating set problem to the instance of a set cover problem and perform branching on large sets and sets of size three containing elements of high degree. When branching is complete, we turn the instance of set cover into an instance of red/blue domination on bipartite graphs and use dynamic programming to count all solutions.

The novel and the most difficult part of the paper is the analysis of the algorithm. To analyze the running time we need to investigate the behavior of the pathwidth of a graph as a function of the measure of the corresponding set cover instance. The difficulty here is to find the measure of the problem that "balances" branching and dynamic programming parts of the algorithm. To choose the right measure we express the bounds on pathwidth as a linear program.

Joint work with Fedor V. Fomin.

## References

[1] Teresa W. Haynes and Stephen T. Hedetniemi. Domination in graphs. *Monographs and Textbooks in Pure and Applied Mathematics*, 209, 1998.

[2] Fedor V. Fomin, Fabrizio Grandoni, and Dieter Kratsch Measure and Conquer: Domination – A Case Study. *Proceedings of the 32nd International Colloquium on Automata, Languages and Programming (ICALP 2005)*, 191–203, 2005.

[3] Fedor V. Fomin, Fabrizio Grandoni, Artem V. Pyatkin, and Alexey A. Stepanov Bounding the Number of Minimal Dominating Sets: a Measure and Conquer Approach. *Proceedings of the 16th Annual International Symposium on Algorithms and Computation (ISAAC 2005)*, 573–582, 2005.

# A distributed approximation scheme for sleep scheduling in sensor networks*

Jukka Suomela

Helsinki Institute for Information Technology HIIT,
Department of Computer Science, University of Helsinki, Finland
jukka.suomela@cs.helsinki.fi
http://www.cs.helsinki.fi/jukka.suomela/

## ABSTRACT

We investigate the theoretical feasibility of near-optimal, distributed sleep scheduling in energy-constrained sensor networks with pairwise sensor redundancy. In this setting, an optimal sleep schedule is equivalent to an optimal fractional domatic partition of the associated redundancy graph.

We present a set of realistic assumptions on the structure of the communication and redundancy relations; for the family of networks meeting these assumptions, we develop an efficient distributed approximation scheme for sleep scheduling. For any $\epsilon > 0$, we demonstrate that it is possible to schedule the sensing activities of the nodes in a local and distributed manner so that the ratio of the optimum lifetime to the achieved lifetime of the network is at most $1 + \epsilon$. The computational effort (time, memory and communication) required at each node depends on $\epsilon$ and the parameters of the network family, but given so-called anchor nodes (a set of nodes meeting certain density constraints) and locally unique node identifiers, the effort is *independent* of the actual network at hand; in particular, the required effort at each node remains constant as the size of the network is scaled up.

The full version of this paper appears in [1]. This is joint work with Patrik Floréen, Petteri Kaski and Topi Musto.

## References

[1] P. Floréen, P. Kaski, and J. Suomela. A distributed approximation scheme for sleep scheduling in sensor networks. In *Proc. 4th Annual IEEE Communications Society Conference on Sensor, Mesh, and Ad Hoc Communications and Networks (SECON, San Diego, CA, USA, June 2007)*, 2007. To appear.

*This research was supported in part by the Academy of Finland, Grants 117499 and 116547, by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, and by Helsinki Graduate School in Computer Science and Engineering (Hecse).

# A Simple Model to Detect User Intent from Query Logs

Libertad Tansini and Bjorn Boden

Department of Computer Science,
Chalmers University of Technology, Gothenburg, Sweden
libertad@cs.chalmers.se

## ABSTRACT

With the explosion of available information from Internet and other sources, search engines must leverage an increased knowledge of user behavior to satisfy the users. This is where User Intent comes in, see [3] and [5], more recently [1] and [4]. Who is better suited than the users themselves to tell us their intent for a query, and how do we obtain this information without asking explicitly? Search engine query logs are the best implicit source from where to extract common behavioral patterns, since they register choices and preferences.

It seems reasonable to differentiate queries based on the user that posed them. We propose a model where the classification is done on user-query pairs and pages. We use an agglomerative clustering algorithm [2] to group user-query pairs and pages. The clusters represent user behaviors or Interest Groups. We obtain on one hand groups of related queries and documents, based on User Intent. On the other hand it is possible to detect groups of similar users, those belonging to the same Interest Groups.

We aim at finding users desires beyond the classification into informational, transactional and navigational [3], but to also include topics, and even a combination of all the previous.

The tests are done on a real query log, the TodoCl log, where we investigate the composition of the groups. The results indicate that the groups are indeed coherent and additionally we confirm that users may be placed in several groups. As expected, we also found some large degenerate groups. They consist of several strongly connected subgroups. To solve this problem we can use a threshold $\theta$ constraining the similarity of objects when clustering. Alternatively, we could use a new similarity measure suggested as future work.

We propose improvements to this model and a search aid-system inspired on Yahoo!s MindSet (http://mindset.research.yahoo.com) to improve the quality of search results.

## References

[1] Baeza-Yates, R., Calderon-Benavides, L. & Gonzalez-Caro, C. The Intention Behind Web Queries. In *Proceedings of SPIRE2006*, 98–109, 2006.

[2] Beeferman, D. & Berger A. Agglomerative Clustering of a Search Engine Query Log. In *Proceedings of ACM SIGKDD International Conference*, 407–415, 2000.

[3] Broder, A. A Taxonomy of Web Search. *SIGIR Forum*, 36(2):3–10, 2002.

[4] Jansen, B. J., Booth, D. L. & Spink A. Determining the User Intent of Web Search Engine Queries. In *Proceedings of WWW2007*, 2007.

[5] Rose, D. E. & Levinson D. Understanding User Goals in Web Search. In *Proceedings of WWW2004*, 13–19, 2004.

# Generic Representations of Sequences in an Algorithm Library

Jarkko Toivonen

Department of Computer Science, University of Helsinki, Finland
jarkko.toivonen@cs.helsinki.fi

## ABSTRACT

A *sequence* is a popular way to specify a subset. Many algoritms use sequences to pass subset information. Passing sequences is a light operation since the actual elements aren't stored in the sequence but in a *container*. All the elements in the sequence can be traversed by starting from the beginning and following a successor relation until all the elements have been enumerated. For example, in an array a sequence can be specified by the first and last index of the sequence. The successor of an element is found by increasing the index.

In C++ a sequence is specified by two iterators, first points to the first element and second points to the last element. This method is used throughout in the standard library of C++, and is also a huge success in other libraries as well. Still, there a problems with this method. For one, the single concept of sequence is specified by two objects. This can be demonstrated with an example of two functions `f` and `g`, both taking a sequence as a parameter and also returning one. Suppose that we want connect these two functions as `sequence2 = g(f(sequence1))`. But since the return value is just one object instead of two, the functions composition isn't straigthforward.

Also, the idea of representing a sequence with two iterators isn't natural for many sequences. In C a string of characters is specified by a pointer to the first element. We know we are at the end of the sequence when we encounter the first element which is null byte. When passing a string as a parameter to a function, it would be costly to compute the end iterator in advance.

Third example comes in the setting of complex data structures, like graphs. It would be hard (or inefficient) for the user to traverse the data structure by always instructing which edge to follow. Instead we can let the data structure do the traversing. Then at each element a callback function is called. This allows the user to perform an operation for each element in the sequence. This kind of callback interface is widely used in data structures as the only access method. Note that it is not possible to implement iterators on top of callback interface.

We propose a concept in C++ called *range* to represent sequences as a single object. These ranges are further divided into categories. Our previous examples belong each to a different range category. Different ranges can look totally dislike, and so we spend a large part in studying the features of different ranges and the range categories.

Our work allows more generic algorithms and also provides better tools for combining the algorithms.

The work is joint with Juha Kärkkäinen and is part of the GLAS algorithm library.

# Applying data mining techniques for measuring software quality

Nikos Tsirakis

Department of Computer Engineering and Informatics, University Of Patras, Greece
tsirakis@ceid.upatras.gr
http://students.ceid.upatras.gr/~tsirakis

## ABSTRACT

Software is playing a crucial role in modern societies. The demand for software quality is increasing and is setting it as a differentiator which can determine the success or failure of a software product. Moreover delivering high quality products is becoming not just a competitive advantage but a necessary factor for companies to be successful [1]. There are many quality measures but a thorough evaluation of quality can arise from the use of an ISO standard [2]. On the other hand data mining and its ability to deal with large volumes of data and to uncover hidden patterns has been proposed as a means to support some quality parameters such us the evaluation and assessment of the maintainability of industrial scale software systems [3]. Data mining is employed to support semi-automated software maintenance [4] and comprehension and provide practical insights into systems specifics, assuming limited prior familiarity. Since software engineering repositories consist of text documents (e.g. mailing lists, bug reports, execution logs), the mining of textual artifacts is requisite for many important activities in software engineering: tracing of requirements, retrieval of components from a repository, identification and prediction of software failures, etc. Finally by applying mining techniques we can extract useful information and predict individual actions about users and calculate aggregate measures regarding the software quality.

## References

[1] Tian J. Quality-Evaluation Models and Measurements. *IEEE Software*, pp: 84-91, 2004.

[2] ISO/IEC 9126-1, Software Engineering Product Quality International Standard. Geneva 2001.

[3] Kanellopoulos Y., Dimopoulos T., Tjortjis C. and Makris C. Mining Source Code Elements for Comprehending Object-Oriented Systems and Evaluating Their Maintainability. *ACM SIGKDD Explorations v8.1, Special Issue on Successful Real-World Data Mining Applications*, June 2006.

[4] P. Antonellis D. Antoniou Y. Kanellopoulos, C. Makris E. Theodoridis C. Tjortjis N.Tsirakis, A Data Mining Methodology for Evaluating Maintainability according to ISO/IEC-9126 Software Engineering-Product Quality Standard. *11th European Conference on Software Maintenance and Reengineering (CSMR)*, March 2007.

# Locality and Graph Kernels for Biomedical Data Analysis

Evgeni Tsivtsivadze

Turku Centre for Computer Science, University of Turku, Finland
evgeni@cs.utu.fi

## ABSTRACT

We present locality and graph kernels and describe their applications to biomedical data analysis. Our locality kernels take advantage of local correlation within sequential data [1, 2]. They are designed to measure similarity within a small window constructed around matching features. Moreover, locality kernels use a range of feature similarity evaluations within a window, namely *position insensitive matching*: only features that match are taken into account irrespective of their position, *position sensitive matching*: features that match but have different positions are penalized, *strict matching*: only features that match and have the same positions are taken into account. This makes locality kernels in particular applicable to the task of remote homology detection in proteins, where the similarity of two or more protein sequences may also imply structural and functional similarity. We use these kernels together with regularized least-squares (RLS) for recognition of previously unseen families from SCOP database. Our experiments show that RLS with kernels incorporating positional information performs better than other baseline methods.

Recently, several kernel functions that operate on data consisting of graphs have been presented. Here, we concentrate on designing graph representations and adapting the kernels for these graphs [3]. In particular, we propose graph representations for dependency parses and analyze the applicability of several variations of the graph kernels for the problem of parse ranking in the domain of biomedical texts. The parses used in the study are generated with the link grammar parser from annotated sentences of BioInfer [4] corpus. The results indicate that designing the graph representation is as important as designing the kernel function that is used as the similarity measure of the graphs.

## References

[1] E. Tsivtsivadze, T. Pahikkala, J. Boberg, T. Salakoski. Locality-convolution kernel and its application to dependency parse ranking. IEA/AIE'06, Volume 4031 of Lecture Notes in Computer Science, Springer, 2006

[2] E. Tsivtsivadze, J. Boberg, T. Salakoski. Locality kernels for protein classification. *Submitted.*

[3] T. Pahikkala, E. Tsivtsivadze, J. Boberg, T. Salakoski. Graph kernels versus graph representations: a case study in parse ranking. *Mining and Learning with Graphs*, Proceedings of the ECML'06 workshop, 2006

[4] S. Pyysalo, F. Ginter, J. Heimonen, J. Björne, J. Boberg, J. Järvinen, T. Salakoski. BioInfer: A corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 2007

# Finding Representative Sets of Bucket Orders from Partial Rankings

Antti Ukkonen
(with Heikki Mannila)

Department of Computer Science and Engineering,
Helsinki University of Technology, Finland
antti.ukkonen@tkk.fi
http://www.cis.hut.fi/aukkonen

## ABSTRACT

Rankings and partial rankings (rankings that do not cover all the items) arise naturally in, e.g., web search, web log mining, collaborative filtering, and different scientific applications. Given a collection of partial rankings, we consider the problem of finding a collection of bucket orders (total orders with ties) that can be used to describe the input collection of rankings well.

We formulate the appropriate optimization problem and give simple and efficient algorithms for the task of finding a good clustering representation for a set of rankings. The basic algorithm is a straightforward k-means type of approach: start from a random set of components, and assign each input ranking to the component it is closest to. Then update the components based on the assignment and iterate until convergence.

This algorithm is, however, sensitive to the choice of initial points. We describe two different methods for finding an initial grouping of the rankings. The first method uses a graph model that turns out to be identical to a simple case of the planted partition model [2]; we solve the problem by using a spectral method. The second approach is based on mapping the input rankings to Euclidean space. Once a clustering if the rankings is obtained, we use the algorithm presented in [1] to construct a bucket order separately for each cluster.

We give a comprehensive set of experimental results of the performance of the algorithms on synthetic and real datasets from two different application domains. First we demonstrate that in case of clickstream data the method finds interesting groups of users based on the sequences in which they visited different parts of a website. Second, we analyze two voting datasets, where each vote is a partial ranking of the candidates. Our results indicate that it is possible to group voters to meaningful clusters based on their preferences. Moreover, in each case the resulting bucket orders can be used to interpret the differences between the clusters.

## References

[1] A. Gionis. H. Mannila. A. Ukkonen and K. Puolamäki Algorithms for Finding Bucket Orders from Data *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[2] A. Condon and R. M. Karp. Algorithms for graph partitioning on the planted partition model. *Random Structures and Algorithms*, 18(2):116–140, 2001.

# Space-efficient Algorithms for Document Retrieval

Niko Välimäki

Department of Computer Science, University of Helsinki, Finland
niko.valimaki@cs.helsinki.fi
www.cs.helsinki.fi/~nvalimak

## ABSTRACT

We study the *Document Listing* problem, where a collection $D$ of documents $d_1, \ldots, d_k$ of total length $\sum_i d_i = n$ is to be preprocessed, so that one can later efficiently list all the ndoc documents containing a given query pattern $P$ of length $m$ as a substring. Muthukrishnan [1] gave an optimal solution to the problem; with $O(n)$ time preprocessing, one can answer the queries in $O(m + \text{ndoc})$ time. We improve the space-requirement of the Muthukrishnan's solution from $O(n \log n)$ bits to $|CSA| + 2n + n \log k(1 + o(1))$ bits, where $|CSA| \leq n \log |\Sigma|(1 + o(1))$ is the size of any suitable *compressed suffix array* ($CSA$), and $\Sigma$ is the underlying alphabet of the documents. The time requirement depends on the $CSA$ used, but we can obtain e.g. the optimal $O(m + \text{ndoc})$ time when $|\Sigma|, k = O(\text{polylog}(n))$. For general $|\Sigma|, k$ the time requirement becomes $O(m \log |\Sigma| + \text{ndoc} \log k)$. Sadakane [2] has developed a similar space-efficient variant of the Muthukrishnan's solution; we obtain a better time requirement in most cases, but a slightly worse space requirement.

Full version of this paper appears in [3]. Joint work with Veli Mäkinen.

## References

[1] S. Muthukrishnan. Efficient algorithms for document retrieval problems. In *Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms (SODA'02)*, pages 657–666, 2002.

[2] K. Sadakane. Space-efficient data structures for flexible text retrieval systems. *Journal of Discrete Algorithms*, 5(1):12–22, 2007.

[3] N. Välimäki and V. Mäkinen. Space-efficient Algorithms for Document Retrieval. Accepted to *18th Annual Symposium on Combinatorial Pattern Matching (CPM 2007)*, Canada, July 9-11, 2007.

# Investigating Learning Behavior with Experiment Databases

Joaquin Vanschoren

Department of Computer Science, Katholieke Universiteit Leuven, Belgium
joaquin.vanschoren@cs.kuleuven.be
www.cs.kuleuven.be/~joaquin

## ABSTRACT

Gaining insights into the behavior of learning algorithms generally involves studying their performance on many different datasets, under various parameter settings. A useful method to learn effectively from previous learning episodes are experiment databases: databases designed to store detailed descriptions of a large number of learning experiments, including the used algorithms, parameter settings, datasets (and preprocessing techniques), evaluation environment, and a large number of measured performance criteria. The experiments themselves are selected to cover a wide range of conditions.

After being populated, these databases allow us to investigate a wide range of questions on algorithm behavior by just querying the database and interpreting the returned results, or by using data mining methods to automatically discover patterns in learning algorithm performance. For instance, one could compare or rank algorithms by querying for their (average) performance on a specific (kind of) dataset, investigate the effect of certain parameters or dataset properties, or automatically find conditions in which an algorithm generally performs well (or badly).

These databases also improve the interpretability of obtained results. Given that the stored experiments cover a wide range of conditions, the conditions under which any observed trend is valid are listed explicitly in the query (and conditions may need to be added to find a clear trend). Also, the experiments can easily be repeated as every aspect of the experimental setup is stored. Furthermore, we can combine theoretical aspects of machine learning research with empirical assessment by also storing detailed properties of the featured datasets (e.g. dataset size, skewness, entropy,. . . ) and algorithms (e.g. model type, time complexity, proneness to bias/variance, sensitivity to noise,...), which can be added at any time and queried on.

By putting these databases online, they serve as a repository of experimental results that can be (re)used (and added to) by various researchers to easily test hypotheses, obtain new insights, or to validate (and possibly refine) previous research. We believe such databases may become a valuable resource for machine learning researchers and practitioners alike.

This research is joint work with Anneleen Van Assche, Celine Vens and Hendrik Blockeel. A full version of this paper appears in [1].

## References

[1] J. Vanschoren, A. Van Assche, C. Vens and H. Blockeel. Meta-learning from Experiment Databases: An Illustration. *Proceedings of the Annual Machine Learning Conference of Belgium and The Netherlands (Benelearn 2007)*, 2007.

# A simple storage scheme for strings achieving entropy bounds

Rossano Venturini

Department of Computer Science, University of Pisa, Italy
rventurini@di.unipi.it
www.di.unipi.it/~rossano

## ABSTRACT

The key problem addressed in this paper consists of representing a string $S[1, n]$ drawn from an alphabet $\Sigma$ within compressed space, and still be able to extract any $\ell$-long substring of $S$ in optimal $\mathcal{O}(1 + \ell/\log_{|\Sigma|} n)$ time. The compressed space usually means space close to the $k$-th order empirical entropy of $S$, which is a lower bound to the space achieved by any $k$-th order compressor.

The first result of this kind is due to Sadakane and Grossi [3]. They proposed a compressed storage scheme for a string $S$ which is *provably better* than storing $S$ as a plain array of symbols. The Sadakane-Grossi's storage scheme is able to achieve the optimal time bound using a sophisticated combination of various techniques: Ziv-Lempel's string encoding and succinct dictionaries.

González and Navarro [2] proposed a simpler storage scheme achieving the same query time and an improved space bound:

$$nH_k(S) + \mathcal{O}\left(\frac{n}{\log_{|\Sigma|} n}\,(k \log |\Sigma| + \log\log n)\right) \tag{1}$$

This storage scheme exploits a statistical encoder (namely, Arithmetic) on most of $S$'s substrings.

We propose a very simple storage scheme that: (1) drops the use of any compressor (either statistical or Lz-like), and deploys only binary encodings and tables; (2) matches the space bound of Eqn. (1) simultaneously over all $k = o(\log_{|\Sigma|} n)$. We also exploit this storage scheme to achieve a corollary result. In fact, it can be used upon the Burrows-Wheeler Transformed string $\texttt{bwt}(S)$ in order to achieve an interesting compressed-space bound which depends on the $k$-th order entropy of both the strings $S$ and $\texttt{bwt}(S)$.
Full version of this paper appears in [1] and SODA 2007. Joint work with P. Ferragina.

## References

[1] P. Ferragina and R. Venturini A simple storage scheme for strings achieving entropy bounds. *Theoretical Computer Science*, 372(1):115–121, 2007.

[2] R. González and G. Navarro. Statistical encoding of succinct data structures. In *Procs CPM*, LNCS 4009, 295–306, 2006.

[3] K. Sadakane and R. Grossi. Squeezing succinct data structures into entropy bounds. In *Procs ACM-SIAM SODA*, 1230–1239, 2006.

# Experiences from mixture model clustering in the analysis of complex didiseases

Jaana Wessman

Department of Computer Science, University of Helsinki, Finland
jaana.wessman@cs.helsinki.fi
www.cs.helsinki.fi/~jantikai

## ABSTRACT

In medical genetics, a *complex disease* is one that is not explainable by a single genetic variant, but is caused by several factors, both genetic and environmental, contributing to the phenotype. The analysis of the etiology (causes) of such diseases is difficult due to the combinatorial number of possible causes under consideration.

As a further complication, many common complex diseases are not complex only in their etiology but also in their clinical outcome. A clinical diagnosis while being useful for selection of treatment might not be optimal for etiological or genetic analysis since they poorly reflect the underlying pathology. To create potential new diagnostic classes, we have performed clustering studies on large phenotype datasets of several complex diseases, exploring this as a method to identify in a non-supervised way subgroups of individuals that might share a genetic etiological basis of the disease. We can then use these groups in standard association analysis, to find out if different set of associated gene variants will be identified with these cluster diagnoses potentially representing so-called ”endophenotypes”.

Many methods exist for clustering. Mixture model clustering has several advantages, most importantly 1) the model assumptions are explicit and the model can be interpreted as a description of the underlying population, not only of the individuals analyzed, and 2) missing data can be easily handled by iterative imputation in the context of the model.

I my poster for SADA07, I present some practical experiences from the use of normal and/or naive Bayesian mixture models on phenotype datasets on several complex diseases (schizophrenia, migrane, the metabolic syndrome). I also present a novel way to define a mixture model over the phenotypes of pairs of siblings, or in the ideal case identical twins, to directly look for *heritable* clusters.

The work has been conducted as a joint work between the groups of Heikki Mannila (Helsinki Insititute of Information Technology) and Leena Peltonen (National Public Health Institute). The shchizophrenia part of this work has been submitted to Molecular Psychiatry.

# Clustering of Sparse Probablity Distributions for N-gram Language Models

Sami Virpioja

Adaptive Informatics Research Centre, Helsinki University of Technology, Finland
sami.virpioja@tkk.fi
www.cis.hut.fi/svirpioj

## ABSTRACT

Statistical language models try to determine the probability distribution over text strings, usually observed as sequences of words, $(w_1 \dots w_k) = w_1^k$. The main issue is the high sparsity of the language data: No matter how large a text corpus is collected, is covers only a tiny fraction of the meaningful sentences. The most widely used models are the $N$-gram models. In practice, they are just huge collections of probability distributions $P(w_i|w_{i-n+1}^{i-1})$ with $n = 1, \dots, N$. Like the language data, also the distributions are very sparse, i.e. most of the $w_i$:s for an observed history $w_{i-n+1}^{i-1}$ are zeros.

A way to reduce the dimensionality, especially with morphologically rich languages, is to use sub-word units instead of words. Models based on statistical morphs found by the Morfessor algorithm [1] have been successfully applied to large vocabulary continuous speech recognition [2]. The reduced lexicon size makes it computationally feasible to cluster the $n$-gram histories $w_{i-n+1}^{i-1}$ according to the distributions $P(w_i|w_{i-n+1}^{i-1})$. In previous work, this kind of $N$-gram models lead to improvements when used in speech recognition [3]. However, the applied incremental clustering algorithm was very slow for building large models.

Now we have studied faster clustering techniques for such sparse distributions. For a large $N$, the number of the $n$-gram histories to cluster may be millions, and even in models based on statistical morphs, the dimensionality is at least several thousands. In order to cope with this, we exploit the sparsity of the distributions by dividing the data points according to the most dominant dimensions. For suitable data, the obtained subsets are small enough to significantly reduce the complexity, and any standard clustering method can then be used for them. When we applied this technique to the $N$-gram data, we obtained improvements in the training performance at the expense of somewhat less optimal clustering.

## References

[1] M. Creutz and K. Lagus. Unsupervised Models for Morpheme Segmentation and Morphology Learning *ACM Transactions on Speech and Language Processing*, 4(1), 2007.

[2] T. Hirsimäki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, and J. Pylkkönen. Unlimited Vocabulary Speech Recognition with Morph Language Models Applied to Finnish. *Computer Speech and Language*, 20(4):515–541, 2006.

[3] S. Virpioja and M. Kurimo. Compact N-gram Models by Incremental Growing and Clustering of Histories. In *Proceedings of the International Conference on Spoken Language Processing (Interspeech 2006 — ICSLP)*, pp. 1037–1040, 2006.

# Reconstruction of Putative Gene Orders in Multifurcating Phylogenetic Trees Using Conserved Intervals

Roland Wittler

Graduiertenkolleg Bioinformatik, Universität Bielefeld, Germany
roland@cebitec.uni-bielefeld.de
www.cebitec.uni-bielefeld.de/˜roland

## ABSTRACT

The orders of genes in genomes provide extensive information. In comparative genomics, differences or similarities of gene orders are determined to predict functional relations of genes or phylogenetic relations of genomes. For this purpose, various combinatorial models can be used to identify groups of genes that are co-located in a set of genomes.

Bergeron *et al.* introduced an algorithm to label internal nodes of a binary phylogenetic tree with sets of putative gene orders [1, 2], following the Fitch-Hartigan algorithm for the most parsimonious tree annotation [4, 5]. Thereby *conserved intervals* not only build the base for the calculations but also represent the complex sets of gene orders in a very compact way [3].

In this poster, the concept of conserved intervals is introduced and some set-theoretic operations are defined. Then the labeling algorithm, including our new expansion to multifurcating trees, is described. Finally results for real data (the mitochondrial RNA of 30 species of *bilateria*) are shown, whose calculation was first enabled by our implementation.

## References

[1] A. Bergeron, M. Blanchette, A. Chateau, and C. Chauve. Operations on sets of conserved intervals. Technical report, Université du Québec à Montreal, 2004.

[2] A. Bergeron, M. Blanchette, A. Chateau, and C. Chauve. Reconstructing ancestral gene orders using conserved intervals. In $4^{th}$ *Workshop on Algorithms in Bioinformatics (WABI'04)*, Lecture Notes in Bioinformatics. Springer, 2004.

[3] A. Bergeron and J. Stoye. On the similarity of sets of permutations and its applications to genome comparison. In $9^{th}$ *Annual International Conference on Computing and Combinatorics (COCOON 2003)*, volume 2697 of *Lecture Notes in Bioinformatics*, pages 68–79. Springer, 2003.

[4] W. Fitch. Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology*, 10:406–416, 1971.

[5] J. A. Hartigan. Minimum mutation fits to a given tree. *Biometrics*, 29:53–65, 1973.

# Multiple Target Tracking in Anonymized Trajectory Databases

Nikolay Vyahhi

Department of Computer Science, St.Petersburg State University, Russia
vyahhi@gmail.com

## ABSTRACT

Multiple target tracking (MTT) is a well-studied technique in the field of radar technology which associates anonymized measurements with the appropriate object trajectories [Reid79] [Konstantinova04]. This technique, however, suffers from a combinatorial explosion, since each new measurement may potentially be associated with any of the existing tracks; consequently, it is not scalable to a large number of objects.

Given a history of object movements, where the corresponding object ids have been removed. A goal is to track all objects from its starting locations, to make the most probable tracks from given subsequent timestamps.

New polynomial time algorithm for MTT problem will be presented in the poster with theoretical basis and experimental results. Also open questions and challenging problems will be formulated.

Collaboration work with Panos Kalnis (National University of Singapore), Spiros Bakiras (City University of New York), Gabriel Ghinita (National University of Singapore).

## References

[Reid79] Donald Reid. An Algorithm for Tracking Multiple Targets. *IEEE Transactions of Automatic Control*, AC-27, Dec 1979.

[Konstantinova04] Pavlina Konstantinova, Milen Nikolov, Tzvetan Semerdjiev. A Study of Clustering Applied to Multiple Target Tracking Algorithm. *International Conference on Computer Systems and Technologies - CompSysTech*, IIIA.22-1–IIIA.22-6, 2004.

# Approximated Geodesic Updates with Principal Natural Gradients

Zhirong Yang

Laboratory of Computer and Information Science
Helsinki University of Technology
P.O. Box 5400, FI-02015 TKK, Espoo, Finland
zhirong.yang@hut.fi
http://www.cis.hut.fi/~rozyang

## ABSTRACT

We propose a novel optimization algorithm which overcomes two drawbacks of Amaris natural gradient updates for information geometry. First, prewhitening the tangent vectors locally converts a Riemannian manifold to an Euclidean space so that the additive parameter update sequence approximates geodesics. Second, we prove that dimensionality reduction of natural gradients is necessary for learning multidimensional linear transformations. Removal of minor components also leads to noise reduction and better computational efficiency. The proposed method demonstrates faster and more robust convergence in the simulations on recovering a Gaussian mixture of artificial data and on discriminative learning of ionosphere data. Full version of this paper will appear in [1].

## References

[1] Zhirong Yang and Jorma Laaksonen. Approximated Geodesic Updates with Principal Natural Gradients. In *Proceedings of The 2007 International Joint Conference on Neural Networks (IJCNN 2007)*, Orlando, USA, August 2007.

# Visual Exploration and Analysis of Financial Data

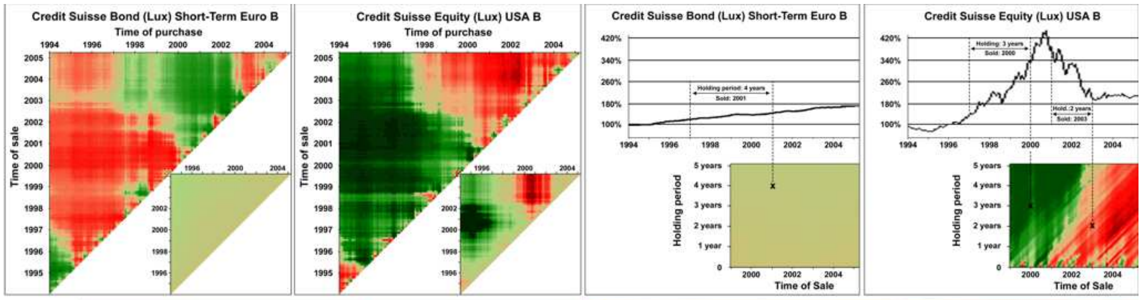
Hartmut Ziegler

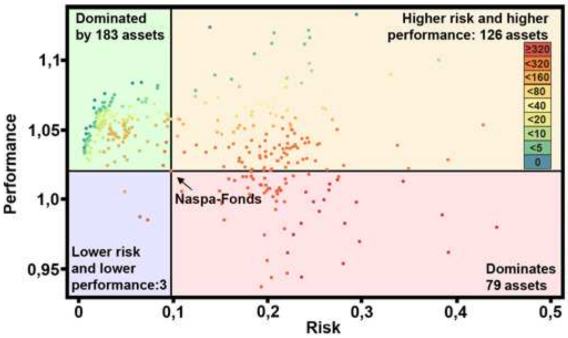
Department of Computer Science, University of Konstanz, Germany
ziegler@inf.uni-konstanz.de
http://infovis.uni-konstanz.de/~ziegler

## ABSTRACT

Visual analysis of financial data is a challenging task due to the large amounts of data that is created in this particular domain every day. Companies like Reuters generate 50.000 data updates per second for the stock market. However, in financial analysis, the most important and most common visualization technique for time series data that is used by investors and analysts for decades is still the traditional line chart.

We have developed an advanced pixel-based approach that extends the ability of traditional line chart techniques, and visualizes the growth of all possible time intervals in only one image [1]. Our technique allows to analyze the growth of a fund (intra-asset analysis) as well as the performance of a fund for each time interval matched against the whole market (inter-asset analysis). Thus, each asset generates a unique finger print revealing the characteristics of an assets behavior. Further visualization techniques specialize on the evaluation of long-term investments, and allow to use Weight Matrices in order to focus on an investors region of interest. We also implemented techniques for performance/risk analysis as well as Dominance Plots and Pareto efficiency curves to analyze sets of assets.





[1] H. Ziegler, T. Nietzschmann, D. A. Keim, "Relevance Driven Visualization of Financial Performance Measures", in Proc. Eurographics/IEEE-VGTC Symposium on Visualization (EUROVIS'07), Sweden, 2007

# Poster abstracts of Hecse students

# Symbolic Model Checking for UML State Machines

Jori Dubrovin

Laboratory for Theoretical Computer Science,
Helsinki University of Technology, Finland
Jori.Dubrovin@tkk.fi

## ABSTRACT

In *formal verification*, the aim is to conclude that a software system fulfills its specification by using mathematically rigorous reasoning. The specification can be, for example, a temporal logic formula that all executions must fulfill, or a desired execution that the system must be able to produce. A modern area of verification is *symbolic model checking* [1], where the idea is to construct a Boolean representation of a finite-state model of the system and to automatically analyze the set of all possible executions of the model by manipulating Boolean formulas.

My current research topic is to develop symbolic model checking methods for UML state machine models. The work is closely related to the industrial project ”Symbolic Methods for UML Behavioural Diagrams”, where the goal is to verify properties of distributed, asynchronous message passing systems such as embedded controllers or communication protocols, and to find bugs in such systems. In particular, I study the applicability of *bounded model checking* (BMC) [2] on concurrent software systems. Bounded model checking is a symbolic technique where a Boolean satisfiability (SAT) solver is used to check whether any execution of the system of up to $k$ steps violates a given property.

In the project, we have designed Boolean encodings for the control logic of UML state machines, for asynchronous events, and for Java-like data manipulation. The encodings are implemented in a translation from a subset of UML to the input language of the symbolic model checker NuSMV. Efficiency of BMC is improved by extending *alternative execution semantics* [3] to be used with UML. The semantics allow firing several independent transitions in the same step, exploiting the concurrency in the system. The next task is to extend the translation to use state-of-the-art satisfiability modulo theories (SMT) solvers, capable of handling also non-Boolean formulas.

## References

[1] E. Clarke, O. Grumberg, D. Peled. *Model Checking*, 1999.

[2] A. Biere, A. Cimatti, E. Clarke, Y Zhu. Symbolic model checking without BDDs. In Tools and Algorithms for Construction and Analysis of Systems, *LNCS*, 1579:193–207, 1999.

[3] J. Rintanen, K. Heljanko, I. Niemelä. Planning as satisfiability: parallel plans and algorithms for plan search. *Artificial Intelligence*, 170(12-13):1031–1080, 2006.

# Approximate Bayesian Inference in Astrophysical Problems

Markus Harva

Adaptive Informatics Research Centre, Helsinki University of Technology, Finland
markus.harva@tkk.fi
http://www.cis.hut.fi/~mha/

## ABSTRACT

I have been working on approximate Bayesian inference. One application area has been astrophysical data analysis in collaboration with University of Birmingham (School of Computer Science; School of Physics and Astronomy). Two finished projects are presented in the poster.

**Analysis of galaxy spectra**. The goal is to separate a set of underlying star population spectra from a set of observed galaxy spectra. A natural approach to this problem is factor analysis (FA), with positivity constraints both on the mixing process as well as on the factors. We developed a method [2] which does not suffer from the certain technical difficulties that alternative approaches have. The method was able to find a physically meaningful decomposition of the spectra, especially, it was able to separate young and old star populations [1].

**Gravitational lens time delay estimation**. Gravitational lensing occurs when the light coming from a distant quasar is bent by the gravitational potential of an intermediate galaxy. This results in multiple images of the quasar being observed. Since the paths the light travels differ in length, there is a delay between the fluctuations in the observed intensities of the different images. From the delays one can compute important cosmological quantities (e.g. the Hubble constant). The estimation of the delays is made problematic by the uneven sampling rate and the low SNR, as shown by Cuevas-Tello et al. [4]. We developed an approach [3] to the problem based on Bayesian modelling.

## References

[1] L. Nolan, M. Harva, A. Kabán, and S. Raychaudhury. A data-driven Bayesian approach for finding young stellar populations in early-type galaxies from their UV-optical spectra. *Monthly Notices of the Royal Astronomical Society*, 366(1):321–338, 2006.

[2] M. Harva and A. Kabán. Variational learning for rectified factor analysis. *Signal Processing*, 87(3):509–527, 2007.

[3] M. Harva and S. Raychaudhury. Bayesian estimation of time delays between unevenly sampled signals. In *Proc. Int. Workshop on Machine Learning for Signal Processing (MLSP'06)*, pages 111–116. Maynooth, Ireland, 2006.

[4] J. C. Cuevas-Tello, P. Tino, and S. Raychaudhury. How accurate are the time delay estimates in gravitational lensing? *Astronomy and Astrophysics*, 454(3):695–706, 2006.

# On Compressing N-gram Language Models

Teemu Hirsimäki

Department of Computer Science and Engineering,
Helsinki University of Technology, Finland
teemu.hirsimaki@tkk.fi
http://www.cis.hut.fi/thirsima/

## ABSTRACT

A statistical language model is the largest part in a large-vocabulary continuous speech recognition system. While the memory resources are often not the main concern in research systems, consumer systems have to take the memory issues into account. Representing the language models efficiently affects the recognition accuracy directly on systems with limited memory resources.

In the literature, several methods have been proposed to control the size of n-gram language models. Entropy pruning [1] can be used to reduce n-gram statistics considerably before the recognition accuracy starts to degrade. Whittaker and Raj, on the other hand, have proposed methods for storing the language model compactly during recognition: The n-gram probabilities and backoff weights can be quantized and the n-gram tree structure can be compressed while maintaining reasonable access times [2, 3].

The current work [4] presents an extension to the compressed data structure of Whittaker and Raj. The idea is based on how the distribution of *leaf n-grams* (an n-gram that is not a prefix for any higher-order n-gram) is changed in pruning. While unpruned models contain leaf n-grams only on the highest order, pruned models also contain a lot of leaf n-grams on lower orders, since pruning algorithms typically prune more high-order n-grams. For leaf n-grams, some of the model parameters can be omitted, which results in memory savings, but requires some modifications to the original data structure. When comparing the proposed and original data structures on Finnish 6-gram models and English 4-gram models, up to 30 % memory savings can be obtained depending on the level of quantization and pruning used in the initial n-gram model.

## References

[1] Andreas Stolcke. Entropy-based pruning of backoff language models. *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 270–274.

[2] E.W.D. Whittaker and B. Raj. Comparison of width-wise and length-wise language model compression. *Proc. EUROSPEECH*, 2001, pp. 733–736.

[3] B. Raj and E.W.D. Whittaker. Lossless compression of language model structure and word identifiers. *Proc. ICASSP*, 2003, pp. 388–391.

[4] T. Hirsimäki. On Compressing N-gram Language Models. *Proc. ICASSP*, 2007, pp. IV-949-952.

# Distributed Constraint-Based Search in Grids

Antti E.J. Hyvärinen

Department of Computer Science and Engineering, TKK, Finland
antti.hyvarinen@tkk.fi
http://www.tcs.hut.fi/~aehyvari/

## ABSTRACT

Constraint-based search, and specifically propositional satisfiability (SAT) checking, has successfully been applied in industrial circuit verification, test pattern generation, planning, bounded model checking, cryptanalysis, and other areas which require potentially significant computing power. Several studies performed on distributed and parallel solving of SAT problems (see, for example, [1, 2, 3]) show that load balancing issues are among the main challenges in such applications. The growing interest from the industry on SAT based applications has guaranteed an active research field and currently emerging low-latency parallel computing environments together with the developing grid infrastructure promise to give more freedom to the design of parallel algorithms. The current results include the following:

(*i*) The foundations of a novel distribution method called scattering have been laid, and the effect of the heuristic in scattering has been experimentally evaluated [1].

(*ii*) Scattering has been implemented (the SATU distributed SAT solver, `http://www.tcs.hut.fi/Software/satu/`) and tested on a large grid, NorduGrid.

In the work, SAT problems are solved using distributed DPLL type algorithms. The work is a balanced combination between theoretical and practical work. The results from (*i*) emphasize the dramatical effect of the heuristic to the performance of the system. Significant progress has been made in heuristics for subproblem construction of scattering: known difficult problems have been efficiently solved using the method. In (*ii*), a considerable amount of effort has been invested into adjusting the implementation to a grid. A general purpose load-balancer developed as part of the work has enabled us to conveniently scale to one of the biggest users of the Finnish M-Grid resources available through NorduGrid. The software is currently in use in a hospital in Geneva, Switzerland.

## References

[1] A.E.J. Hyvärinen, T. Junttila, and I. Niemelä. A distribution method for solving SAT in grids. In *SAT, volume 4121 of LNCS*, pages 430-435. Springer, 2006.

[2] H. Zhang, M.P. Bonacina, and J. Shiang. PSATO: A distributed propositional prover and its application to quasigroup problems. *J. Symb. Comput.*, 21(4):543–560, 1996.

[3] M. Böhm and E. Speckenmeyer. A fast parallel SAT-solver – Efficient Workload Balancing. *Ann. Math. Artif. Intell.*, 17(3-4):381–400, 1996

# Algorithm Visualizations as Interactive Learning Material

Ville Karavirta

Department of Computer Science, Helsinki University of Technology, Finland
vkaravir@cs.hut.fi
http://www.cs.hut.fi/u/vkaravir/

## ABSTRACT

Software Visualization (SV) and its sub-fields, Algorithm visualization (AV) and animation (AA), have been used in education for a few decades. The goal of AA is to help humans to understand difficult pieces of program code by providing a view on a higher level of abstraction.

Algorithm animation has been found to be educationally effective, provided that it is interactive enough. Several AA systems have been developed for educational puposes. However, the main reason for not taking full advantage of AA in teaching is often lack of time of the educators. Based on this, we propose data exchange between AA systems as a suitable solution to lower the effort needed to produce AVs.

We have defined a taxonomy of algorithm animation languages based on a thorough survey of existing AV systems and the languages they use. In the survey, our goal was to find the requirements for a common language to be used in data exchange between the systems. By using the taxonomy as a guideline for a new language, we have introduced an XML language for describing algorithm animations, Xaal, or eXtensible Algorithm Animation Language. In addition, we have implemented a set of tools to support data exchanging. A more thorough discussion about this research can be found in [1, 3].

Furthermore, we have analyzed students' usage of interactive AV exercises [2]. In the study, we divided the students into five groups based on their usage of resubmissions using a clustering technique. The groups used significantly different resubmission strategies. In the future, the aim is to find out what kind of learning material students need, how different student groups use the materials, as well as how it benefits learning.

## References

[1] Ville Karavirta. Integrating algorithm animation systems. In *Proceedings of the Fourth Program Visualization Workshop*, Accepted for publication.

[2] Ville Karavirta, Ari Korhonen, and Lauri Malmi. On the use of resubmissions in automatic assessment systems. *Computer Science Education*, 16(3):229 – 240, September 2006.

[3] Ville Karavirta, Ari Korhonen, and Lauri Malmi. Taxonomy of algorithm animation languages. In *SoftVis '06: Proceedings of the 2006 ACM symposium on Software visualization*, pages 77–85, New York, NY, USA, September 2006. ACM Press.

# Local Dependent Components

Arto Klami

Helsinki Institute for Information Technology and
Adaptive Informatics Research Centre,
Helsinki University of Technology, Finland
arto.klami@tkk.fi
http://www.cis.hut.fi/aklami/

## ABSTRACT

Canonical correlation analysis (CCA) is a classical method for studying dependencies between two data sets of paired samples $(x, y)$. CCA maximizes the mutual information between representations extracted from $x$ and $y$ (for normally distributed data), which makes it a predecessor of the more recent methods that find dependencies between sets of variables, such as various co- and discriminative clustering methods and information bottleneck.

A main problem with these models is that they overfit easily to small data sets. In practical data analysis tasks, such as data fusion in bioinformatics, the number of dimensions may even exceed the number of data points, and the existing regularization methods are insufficient. The recent finding [1] that CCA can be interpreted as a generative model was very promising since it opened the road to Bayesian treatments, and hence to rigorous ways of including both prior knowledge and complexity control.

The generative model of CCA assumes normally distributed data and linear components, which are very restrictive assumptions in practical data analysis tasks. It would make sense to make these assumptions locally, however, and search for local dependencies between data sets. This requires very good complexity control methods since the effective number of data points per CCA will decrease. In addition, locality should be defined in the sense of the covariance matrix the CCA introduces, not in the original metric of the data space.

We introduce a fully Bayesian mixture of CCAs [2] which helps in avoiding overfitting, and makes the analysis local. Furthermore, it extends the traditional CCA by finding a decomposition of local variation into common (shared) and data set-specific components, instead of finding only features shared by the data sets. The model is formulated as a Dirichlet process mixture, and a split-merge procedure based on Gibbs sampling is used for approximative inference. The quantities used in data analysis, such as degree of local dependency, loadings of the components relevant for the dependencies, etc are computed from the posterior.

## References

[1] F. R. Bach and M. I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical Report 688, Dept. of Statistics, Univ. of California, Berkeley, 2005.

[2] Arto Klami and Samuel Kaski. Local Dependent Components In *Proceedings ICML 2007*, accepted for publication.

# Emergence of conjunctive visual features by quadratic independent component analysis

Jussi T. Lindgren

Department of Computer Science, University of Helsinki, Finland
jtlindgr@cs.helsinki.fi

## ABSTRACT

In previous studies, quadratic modelling of natural images has resulted in cell models that react strongly to edges and bars. Here we apply quadratic Independent Component Analysis to natural image patches, and show that up to a small approximation error, the estimated components are computing conjunctions of two linear features. These conjunctive features appear to represent not only edges and bars, but also inherently two- dimensional stimuli, such as corners. In addition, we show that for many of the components, the underlying linear features have essentially V1 simple cell receptive field characteristics. Our results indicate that the development of the V2 cells preferring angles and corners may be based on the principle of unsupervised sparse coding of natural images.

This is joint work with Aapo Hyvärinen. The presented poster and its corresponding paper appeared originally elsewhere [1].

## References

[1] J. T. Lindgren and A. Hyvärinen. Emergence of conjunctive visual features by quadratic independent component analysis *Advances in Neural Information Processing Systems (NIPS)*, in press, 2007.

# Errors in concurrent programming assignments

Jan Lönnberg

Department of Computer Science and Engineering
Helsinki University of Technology, Finland
jlonnber@cs.hut.fi
http://www.cs.hut.fi/~jlonnber/

## ABSTRACT

Students' solutions to programming assignments provide material that can be used to improve several interlinked processes. The student-submitted assignment solutions can be used to evaluate and improve the students' learning, the teaching and the assignments by identifying weaknesses in the students' knowledge, skill and understanding of the assignments. Information on defects in students' programs can also be used as a starting point for the development of testing and debugging methodologies and tools. Concurrency further complicates the programming process by introducing nondeterminism and its effect on debugging has seen little research.

For the reasons outlined above, I am examining student submissions from the three programming assignments on the concurrent programming course at HUT. The first data was collected during the Autumn 2005 course[1] (see [1]), and more detailed data has been collected during the Autumn 2006 course[2]. In 2005, students were required to submit both the actual program source code and a brief report outlining how their solution works with an emphasis on concurrency-related behaviour. Defects were found in the programs using a combination of testing and manual analysis and students' explanations of how their code works were used to deduce the underlying mistakes. From the 2006 version of the course, more detailed data was collected by requiring students submitting corrections to failed assignments to describe how they designed the flawed part of their program, how they tried to determine whether it was correct and how they fixed it. I have also used semi-structured interviews with selected students to find weaknesses in their knowledge of concurrency and in their software design, testing and debugging metholodologies that would explain the defects observed in their code.

The end results of this work will be information on weaknesses in students' concurrent programming knowledge and skills and the effects of these on the concurrent programs they produce that will be used to improve the teaching of concurrent programming as well as design testing and debugging tools and methods that address these problems.

## References

[1] Jan Lönnberg. Student errors in concurrent programming assignments. In Anders Berglund and Mattias Wiggberg, editors, *Proceedings of the 6th Baltic Sea Conference on Computing Education Research, Koli Calling 2006*, pages 145–146, Uppsala, Sweden, 2007. Uppsala University.

[1] http://www.cs.hut.fi/Studies/T-106.420/main.html

[2] http://www.cs.hut.fi/Studies/T-106.5600/english.shtml

# Pragmatic aspects in computer-supported negotiations of virtual enterprise contracts

Janne Metso

Department of Computer Science, University of Helsinki, Finland
janne.metso@cs.helsinki.fi
http://www.cs.helsinki.fi/janne.metso/

## ABSTRACT

In todays modern world, business organizations must concentrate on their core business ideas in order to remain competitive and at the same time outsource other functionality. The specialized enterprises need to network with other enterprises for the provision of competitive services. The result is a networked business organization, a virtual enterprise. The virtual enterprises are based on contracts that define the purpose of the virtual enterprise, the participants in it, and the rules of the virtual enterprise. The contract is expressed in an electronic form, as an eContract. Connecting organizations through the Internet requires new infrastructure support which consists of partner discovery and management of virtual enterprise life-cycle. A breeding environment is used to gather partners together and help through the early phases of the life-cycle, such as contract negotiation.

Pragmatic aspects in the virtual enterprise negotiations consist of issues like modeling the goals of participating enterprises, different social aspects in the negotiation process itself, and the execution of the negotiation process. The enterprises control their actions in a business domain using policies. The policies dictate which negotiations are taken part of and which issues are important in them. Social aspects include modeling and taking into account the business environment that an enterprise has. The enterprise might want to favor certain strategic partners that it already has. In the negotiation process itself it is important to solve dependencies between different issues or variables. Another important issue in the negotiation process is using a contract language to model the contract.

As a result of this work, we expect to create tools for enterprises to use during negotiations, and automating negotiation in routine cases, when entering business networks. The tools are used to model business policies and to help factor in the environment. A rule based system will provide the policy information for the automated negotiation system which will help in calculation of expected utility and risk in a given negotiation situation. Providing such assistance will help a decision maker in an enterprise to determine if joining a virtual enterprise is worth wile or not and what negotiation issues are important to consider.

The formal basis for the negotiation system will be in distributed constraint satisfaction problems (DCSP). Virtual enterprise negotiations are an extension to the basic DCSP model with multiple constraints for each agent (participant in a VE) and multiple variables. In the case of VE negotiations each agent has its own constraints for the variables and not all variables are shared by all parties. Social aspects have an effect on the behaviour of a negotiation strategy of an enterprise. When a business is participating in multiple simultaneous virtual enterprises, it cannot be only self-interested, but it needs to take into account the goals and needs of its valued partners.

# A Feature Selection Methodology for Steganalysis

Yoan Miche, Benoit Roue, Amaury Lendasse, and Patrick Bas

Laboratory of Computer and Information Science,
Helsinki University of Technology, Finland
ymiche@cis.hut.fi

## ABSTRACT

This poster presents a methodology to select features before training a classifier based on Support Vector Machines (SVM). In this study 23 features presented in [1] are analysed. A feature ranking is performed using a fast classifier called K-Nearest-Neighbours combined with a forward selection. The result of the feature selection is afterward tested on SVM to select the optimal number of features. This method was tested with the Outguess steganographic software and 14 features were selected while keeping the same classification performances. Results confirm that the selected features are efficient for a wide variety of embedding rates.

Full version of this paper appears in [2].

## References

[1] Fridrich J. Feature-Based Steganalysis for JPEG Images and its Implications for Future Design of Steganographic Schemes *6th Information Hiding Workshop*, LNCS vol. 3200:67–81, 2004.

[2] Miche Y., Roue B., Lendasse A., and Bas P. A Feature Selection Methodology for Steganalysis *Multimedia Content Representation, Classification and Security, International Workshop, MRCS 2006*, LNCS vol. 4105:49–56, 2006

# Generalization to Unseen Cases

Teemu Roos

Department of Computer Science, University of Helsinki, Finland
teemu.roos@cs.helsinki.fi
www.cs.helsinki.fi/teemu.roos

## ABSTRACT

In the No Free Lunch theorems of machine learning theory, the generalization performance of a hypothesis is measured strictly in terms of its behavior on *unseen cases*, i.e. cases that are different from those in the training set from which the hypothesis was inferred. This performance measure, usually called *off-training set error*, is different from the standard generalization error, which is used in almost all other theoretical machine learning research. Unlike the standard generalization error, off-training set error may differ significantly from the error a hypothesis makes on the training data, with high probability, even with large sample sizes. Still, we derive a data-dependent bound on the difference between off-training-set and standard generalization error. Our result is based on a new bound on the so called *missing mass*, which for small samples is stronger than existing bounds based on Good-Turing estimators. As we demonstrate on a number of benchmark data-sets, our bound gives nontrivial generalization guarantees in many practical cases. In light of these results, we show that certain claims made in the No Free Lunch literature are overly pessimistic.

This work has been published in [1]. Joint work with Peter Grünwald, Petri Myllymäki, and Henry Tirri.

## References

[1] T. Roos, P. Grünwald, P. Myllymäki, and H. Tirri. Generalization to Unseen Case. *Advances in Neural Information Processing Systems*, 18:1129–1136, MIT Press, Cambridge, MA, 2006.

# Trust based on evidence

Sini Ruohomaa

Department of Computer Science, University of Helsinki, Finland
sini.ruohomaa@cs.helsinki.fi
http://www.cs.helsinki.fi/sini.ruohomaa/

## ABSTRACT

In open distributed systems, the set of potential collaboration partners is large and highly dynamic. An adaptive soft security approach is needed to allow new collaborations to form, on one hand, while on the other hand quickly adapting self-defense mechanisms according to evidence of the new partners' good or bad behaviour.

Trust management protects assets coming under risk when collaborating. Evidence-based trust management has three phases: identifying the target of trust, making a context-specific trust decision, and observing the target's behaviour. Their hard security equivalents are authentication, authorization and anomaly or breach detection; the main difference is that the decision-making and observation phases are connected by a feedback loop.

The TuBE trust management system is built into collaboration middleware. The guarded service application does not need to be aware of the trust mechanism around it, as the guard is implemented around it as a wrapper [2]. Trust decisions are made by combining four factors: risk and business importance, reputation, and temporary contextual adjustments to the three. Reputation is built from continuously updated evidence: a reputation system stores and distributes local and globally shared experiences, and provides an aggregated summary of the information for use in trust decisions.

Trust is used both when setting up a collaborative virtual organization, and through its runtime. Partner selection is based on trust, with a strong focus on reputation. During the collaboration, action-specific trust decisions are made when resources must be committed.

The TuBE trust management system extends the Pilarcos middleware, which is built on the Web Services technology. Interesting research questions include the expression and propagation of trust and reputation, application-level observation, and analysis methods used for upkeeping reputation [1].

## References

[1] Sini Ruohomaa and Lea Kutvonen. Trust management survey. In *Proceedings of the iTrust 3rd International Conference on Trust Management*, 23–26, May, 2005, Rocquencourt, France, pages 77–92. Springer-Verlag, LNCS 3477/2005, May 2005.

[2] Sini Ruohomaa, Lea Viljanen, and Lea Kutvonen. Guarding enterprise collaborations with trust decisions — the TuBE approach. In *Proceedings of the First International Workshop on Interoperability Solutions to Trust, Security, Policies and QoS for Enhanced Enterprise Systems (IS-TSPQ 2006)*, March 2006.

# Discriminative MCMC

Jarkko Salojärvi

Laboratory of Computer and Information Science
Helsinki University of Technology
P.O. Box 5400, FI-02015 TKK, FINLAND
jarkko.salojarvi@cis.hut.fi

## ABSTRACT

In generative modeling, it is well-known that usual Bayesian inference is not optimal for generalizing to new data if the model family is incorrect. Arguably the best solution is to improve the model family by incorporating more prior knowledge. However, this is not always possible or feasible, and simplified models are being generally used, often with good results. There are good reasons for still applying Bayesian-style techniques but the general problem of how to best do inference with incorrect model families is still open.

In discriminative modeling, here meaning inference on the distribution $p(y|x)$, the question of using discriminative vs. generative models has attracted a lot of interest. The question has been whether to model $p(y|x)$ directly or to build a generative model for the joint distribution $p(y, x)$ and compute the conditional distribution from that. It is easy to show that point estimates computed by mazimizing the joint likelihood and the conditional likelihood differ. Maximum conditional likelihood works better asymptotically, and it can be optimized with expectation-maximization procedures. Here we extend the work from point estimates to distributions. Standard posterior distribution is optimal if the model family is correct, but is there an extension that would be analogous to standard Bayesian inference while working better for incorrect model families?

Earlier work [1] has proposed the so-called supervised posterior. The posterior has however only been justified heuristically. We give an axiomatic proof, introduce Markov Chain Monte Carlo methods for computing with the posterior, and demonstrate that it works as expected. The inference reduces to standard Bayesian inference if there are no covariates.

For the purpose of regression, the discriminative posterior makes it possible to use more general model structures: in essence any generative model. The gained advantage, compared to using the standard posterior, is that the predictions are more accurate when the model family is incorrect. Compared to Bayesian regression the predictions should be better if the introduced generative model for $x$ is informative.

## References

[1] P. Grünwald, P. Kontkanen, P. Myllymäki, T. Roos, H. Tirri, and H. Wettig. Supervised posterior distributions. presentation at the Seventh Valencia International Meeting on Bayesian Statistics, Tenerife, Spain, 2002. http://homepages.cwi.nl/~pdg/presentationpage.html.

# A Simple Approach for Finding the Globally Optimal Bayesian Network Structure

Tomi Silander

Helsinki Institute for Information Technology,
Department of Computer Science,
University of Helsinki, Finland
tomi.silander@cs.helsinki.fi
http://cosco.hiit.fi/

## ABSTRACT

We study the problem of learning the best Bayesian network structure with respect to a decomposable score such as BDe, BIC or AIC. This problem is known to be NP-hard, which means that solving it becomes quickly infeasible as the number of variables increases. Nevertheless, in this paper we show that it is possible to learn the best Bayesian network structure with over 30 variables, which covers many practically interesting cases. Our algorithm is less complicated and more efficient than the techniques presented earlier. It can be easily parallelized, and offers a possibility for efficient exploration of the best networks consistent with different variable orderings. In the experimental part of the paper we compare the performance of the algorithm to the previous state-of-the-art algorithm. Free source-code and an online-demo can be found at http://b-course.cs.helsinki.fi/bene.

Full version of this paper appears in [1].

## References

[1] T. Silander and P. Myllymäki. A simple approach for finding the globally optimal Bayesian network structure. In *Proceedings of the 22nd Conference on Uncertainty in Artificial I ntelligence (UAI-2006)*, pages 445–452, 2006.

# SOM+EOF for Finding Missing Values

Antti Sorjamaa

Department of Computer Science and Engineering
Helsinki University of Technology, Espoo, Finland
Antti.Sorjamaa@hut.fi
http://www.cis.hut.fi/projects/tsp

## ABSTRACT

The presence of missing values in the underlying time series is a recurrent problem when dealing with databases. This abstract summarizes our method, which combines two projection methods, Self-Organizing Maps and Empirical Orthogonal Functions.

Self-Organizing Maps (SOM) [1] aim to ideally group homogeneous individuals, highlighting a neighborhood structure between classes in a chosen lattice. The SOM algorithm is based on unsupervised learning principle where the training is entirely stochastic, data-driven. The SOM algorithm allows projection of high-dimensional data to a low-dimensional grid. Through this projection and focusing on its property of topology preservation, SOM allows nonlinear interpolation for missing values.

Empirical Orthogonal Functions (EOF) [2] are deterministic, enabling linear projection to a high-dimensional space. They have also been used to develop models for finding missing data [3]. Moreover, EOF models allow continuous interpolation of missing values, but are sensitive to the initialization.

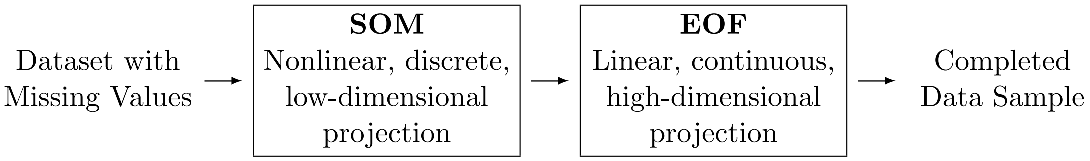


Figure 1: Global methodology, the SOM+EOF, summarized.

The work has been published in the ESANN 2007 conference [4] and it is a joint work with Paul Merlin, Bertrand Maillet and Amaury Lendasse.

## References

[1] T. Kohonen. Self-Organizing Maps. Springer-Verlag, Berlin, 1995.

[2] R. Preisendorfer. Principal Component Analysis in Meteorology and Oceanography. Elsevier, 1988

[3] J. Boyd, E. Kennelly and P. Pistek. Estimation of EOF expansion coefficients from incomplete data. *Deep Sea Research*, 41:1479-1488, 1994.

[4] ESANN 2007 Conference website: http://www.dice.ucl.ac.be/esann/.

# Modeling how varying data quality affects the ability to detect trends in environmental time series

Mika Sulkava

Laboratory of Computer and Information Science
Helsinki University of Technology, Finland
Mika.Sulkava@tkk.fi
www.cis.hut.fi/~sulkava

## ABSTRACT

Detection of changes in ecosystem characteristics is a principal tool for identifying and understanding the effects of anthropogenic activities on the condition and functioning of ecosystems. It is widely known that temporal trends can be blurred by the imprecision of the data. When detected, trends support the assessment of the future development of ecosystems under the present and predicted environmental scenarios. Further, trend detection can potentially provide tools for policy makers when shaping the environmental policy. Research program managers are aware of the difficulties surrounding representative sampling and therefore enforce strict sampling protocols. Standardized sampling can be so effective that the initially much smaller uncertainty in the instrumental analysis becomes substantial. Despite instrumental improvements and the introduction of quality control and quality assurance systems, instrumental analyses are still prone to many uncertainties. The aim of this study is to quantify the time required to detect linear trends in the chemical composition of ecosystem components, given a certain quality of the applied analytical method.

In this study, we present a novel technique and theoretical computations based on weighted linear regression models for the detection of trends in single and combined indices. The theory is clarified with examples from the International Co-operative Programme on Assessment and Monitoring of Air Pollution on Forests (ICP Forests). The results show that, when sampling protocols largely reduce the variability of representative sampling, poor quality of the instrumental analysis blurs the data such that environmental monitoring or long-term ecological research programs can lose the ability to detect trends by causing up to three decades long delay in detecting changes. We can thus conclude that high quality of the instrumental analysis is a prerequisite for a sensitive monitoring program.

Full version of this paper has been accepted for publication [1]. The work has been done together with Sebastiaan Luyssaert and Ivan A. Janssens from University of Antwerp, Pasi Rautio from the Finnish Forest Research Institute, and Jaakko Hollmén from Helsinki University of Technology.

## References

[1] M. Sulkava, S. Luyssaert, P. Rautio, I.A. Janssens, and J. Hollmén. Modeling the effects of varying data quality on trend detection in environmental monitoring. *Ecological Informatics*, accepted for publication, 2007.

# Input Selection Algortihm for Radial Basis Function Networks

Jarkko Tikka

Laboratory of Computer and Information Science
Helsinki University of Technology, Finland
tikka@mail.cis.hut.fi
http://www.cis.hut.fi/tikka/

## ABSTRACT

Radial basis function (RBF) networks have been widely utilized in regression problems. The advantages of RBF networks are that the training of networks is relatively fast and they are capable on universal approximation with non-restrictive assumptions [1]. Fastness of the training is a consequence of simple structure of the RBF networks. They have only one hidden layer, in which each node corresponds to a basis function and a mapping from the hidden layer to the output layer is linear. The activation of hidden nodes are evaluated by the distance between an input vector and a center of the basis function. The disadvantage is their black-box characteristics. Basically, the network includes all the input variables and, in addition, importances of the inputs are not clear at all. However, interpretation or understanding of the underlying process can be increased by selecting the input variables. In addition to interpretability, the rejection of non-informative inputs can improve the generalization capability of the network [2].

Several strategies exist to perform input selection, for instance the filter approach, the wrapper methodology, and the embedded methods [2]. In this work, an input selection algorithm for RBF networks is proposed. The algorithm is based on the weighted Euclidean distance, thus each input dimension has its own weight. The sum of the weights are constrained such that some of the weights tend to be zero and the corresponding inputs are rejected from the final model. The problem is defined as a constrained optimization problem, which takes into account the non-linear dependency between the inputs and the output. The proposed algorithm solves a log-barrier reformulation of the original optimization problem. The estimation of output and the selection of inputs are carried out simultaneously, therefore the algorithm belongs in the the class of embedded input selection methods.

In the experiments, the algorithm was applied to the simulated and real world data sets. The results were convincing in both cases. Full version of this paper appears in [3].

## References

[1] J. Park, I.W. Sandberg. Approximation and Radial-Basis-Function Networks. *Neural Computation*, 5(2):305–316, 1993.

[2] I. Guyon, A. Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Reasearh*, 3(Mar):1157–1182, 2003.

[3] J. Tikka. Input Selection for Radial Basis Function Networks by Constraint Optimization. Submitted for publication to the *17th International Conference on Artificial Neural Networks (ICANN 2007)*.

# Sparse Distributed Representations for Words with Thresholded Independent Component Analysis

Jaakko J. Väyrynen

Department of Computer Science And Engineering,
Helsinki University of Technology, Finland
jaakko.j.vayrynen@tkk.fi
www.cis.hut.fi/~jjvayryn

## ABSTRACT

Independent component analysis (ICA) [1, 2] can be shown to be an extension of latent semantic analysis (LSA) [3], which is based on singular value decomposition (SVD). LSA has been shown to produce distributed representations that perform extremely well in various textual tasks [4]. The representations, however, are latent and individual components do not directly correspond to any specific feature. ICA computes an additional rotation of the space so that the components are interesting. We show that the ICA components for words can be further processed by thresholding to produce a sparse representation. The applicability of the thresholded ICA representation is compared to SVD in a multiple choice vocabulary task with three data sets.

Full version of this paper appears in [5]. This is joint work with Lasse Lindqvist and Timo Honkela.

## References

[1] P. Comon. Independent Component Analysis, a new concept? *Signal Processing*, 36(3):287–314, Apr. 1994.

[2] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* John Wiley & Sons, 2001.

[3] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *J. American Socyecty of Information Science*, 41(6):391–407, 1990.

[4] T. K. Landauer and S. T. Dumais. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240, 1997.

[5] J. J. Väyrynen, L. Lindqvist, and T. Honkela. Sparse distributed representations for words with thresholded independent component analysis. In *Proc. International Joint Conference on Neural Networks (IJCNN 2007).* To appear.

# A New Deconvolution Tool for Image Denoising

Hannes Wettig
Complex Systems Computation Group (CoSCo)
Helsinki Institute for Information Technology (HIIT)
P.O.Box 68 (Department of Computer Science)
FIN-00014 University of Helsinki, Finland
Hannes.Wettig@hiit.fi

## ABSTRACT

We consider the problem of denoising a natural digital image that has been contaminated by white noise, i.e. that has an additive zero-mean i.i.d. Gaussian component to be removed. Typically this is done by first transforming the image using some wavelet basis, which leaves the noise distribution untouched while concentrating the energy of the original image. The simplest way of denoising in the wavelet domain is thresholding, i.e. nulling small entries that can be assumed to contain mainly noise. A mixture model, consisting of one component for the noise and one for the signal, basically amounts to the same thing, as both models can be assumed to peak at zero. A theoretically better way is to deconvolute the observed signal. The denoised signal is then the expectation of the information part, which is optimal for the mean square error. But deconvolution is computationally demanding, for which reason both model components are usually assumed Gaussian. For the noise part this is good by design of the problem. But wavelet transformed natural images behave differently. We propose a new deconvolutionary tool that models the signal with a Laplace (double-exponential) distribution. This is more suitable, although far from perfect. However, we can compute all expectations and derivatives we need, which enables us to switch from a mixture to a convolution model. Our convolution model outperforms the simple Gauss/Gauss-convolution. Up to now our method is not competitive with state-of-the-art algorithms, but we demonstrate that this tool is highly flexible and extendable. Work in progress, paper forthcoming.

# Visualization on Real-time Traffic Simulation Data Stream

Mu Zhou

Department of Computer Science, Helsinki University of Technology, Finland
mu.zhou@hut.fi
www.cs.hut.fi/Research/simutraf

## ABSTRACT

Visualization in Transportation Engineering is ubiquitous, including applications for traffic planning, controlling, monitoring, and analyzing. Visual representation can help to gain insight into complex, abstract traffic data, and thus enhance cognition of the essentials of the traffic process.

Recent developments in transportation engineering research enable the real-time simulation systems to provide functionalities for traffic monitoring and to some extent for traffic prediction. However, integrating a traffic simulator into an Intelligent Transportation System (ITS) application is a difficult task, because it is not straightforward to take the data stream from the simulator directly into ITS applications for traffic analysis. In this research we present a new visualization framework [1] to manipulate the intensive and very dynamic data stream from traffic simulators. With this visualization framework, we can build applications that provide better tools to examine the complicate traffic process by more user-defined details. Moreover, the whole simulation process can be rebuilt according to users needs. These applications can be accessed through Internet by traffic professionals for different purposes. We can also use this approach to build a case library of various traffic events, which can be used for the future traffic prediction. Here the huge volume of simulation data, which are normally dumped, can be fully utilized. Integrating the data with the online simulation system itself, we can even create some more reality for the traffic simulation systems, which combine traffic information with real scenes.

So far, we have built several interesting applications, including a real time 3D traffic map, 3D traffic map animation and video-game-like traffic statistics animation, etc. A group of such application prototypes have been implemented quickly, which proves that our framework is efficient.

## References

[1] Zhou M, Korhonen A, Malmi L, Kosonen I, Luttinen T. Application Framework for Integration of GIS-T with Real-Time Traffic Simulation System. *Journal of the Transportation Research Board*, 2006. Nro 1972, pp. 78-84.