06201 Abstracts Collection
Combinatorial and Algorithmic Foundations of Pattern and Association Discovery
— Dagstuhl Seminar —

Rudolf Ahlswede[1], Alberto Apostolico[2] and Vladimir I. Levenshtein[3]

[1] Univ. Bielefeld, DE
ahlswede@mathematik.uni-bielefeld.de
[2] Univ. di Padova, IT
axa@dei.unipd.it
[3] Keldysh Institute - Moscow, RU
leven@keldysh.ru

Abstract. From 15.05.06 to 20.05.06, the Dagstuhl Seminar 06201 "Combinatorial and Algorithmic Foundations of Pattern and Association Discovery" was held in the International Conference and Research Center (IBFI), Schloss Dagstuhl. During the seminar, several participants presented their current research, and ongoing work and open problems were discussed. Abstracts of the presentations given during the seminar as well as abstracts of seminar results and ideas are put together in this paper. The first section describes the seminar topics and goals in general. Links to extended abstracts or full papers are provided, if available.

Keywords. Data compression, pattern matching, pattern discovery, search, sorting, molecular biology, reconstruction, genome rearrangements

06201 Executive Summary – Combinatorial and Algorithmic Foundations of Pattern and Association Discovery

The focus of this seminar has been on the completely new scenario and on the wild paradigm shift that are forced by the recent progresses of ICT (information and communication technology). The new scenario is that data and information accumulate at a pace that makes it no longer fit for direct human inspection. The paradigm shift is that, in contrast to a primeval, persistent tenet of traditional information science and technology, the bottleneck in communication is no longer represented by the channel or medium but rather by the limited perceptual bandwidth of the final user: more and more often, the time and resources that need to be invested in order to gain access to information happens to be disproportionate to fruition time and value, thereby defying the very purpose of access. Consequently, the challenge of maximizing the throughput to the final user has taken up entirely new meanings. The implications brought about

Dagstuhl Seminar Proceedings 06201
Combinatorial and Algorithmic Foundations of Pattern and Association Discovery
http://drops.dagstuhl.de/opus/volltexte/2006/787

by such a dramatic change in perspectives have barely begun to be perceived. A science and engineering of discovery is developing to meet these challenges, which promises to revolutionize many facets of human activity beginning with the basic notions and practices of scientific investigation itself.

The goals of this seminar have been (1) to identify and match recently developed methods to specific tasks and data sets in a core of application areas; next, based on feedback from the specific applied domain, (2) to fine tune and personalize those applications, and finally (3) to identify and tackle novel combinatorial and algorithmic problems, in some cases all the way to the development of novel software tools.

Keywords: Data compression, pattern matching, pattern discovery, search, sorting, molecular biology, reconstruction, genome rearrangements

Joint work of: Ahlswede, Rudolf; Apostolico, Alberto; Levenshtein, Vladimir I.

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2006/792

Suballowable sequences of permutations

Andrei Asinowski (Technion - Haifa, IL)

We define a suballowable sequence of permutations as an infinite periodic sequence of permutations of the set {1, 2, ..., n} such that:
1. Each term in the second half period is the reverse of the corresponding term in the first half period;
2. In the course of each half period, each pair of labels switches just once.

This notion generalizes that of an allowable sequence introduced in 1980 by Goodman and Pollack as a combinatorial encoding of configurations of points in the plane.

Combinatorial properties of allowable sequences were used in solving several problems in Discrete Geometry (configurations of points, arrangements of lines, k-sets, $\leq k$-sets etc.). Some of these properties hold also for suballowable sequences.

We show a characterization of allowable sequences in the class of suballowable sequences, and prove a Helly-type result on labels in sets of permutations which form (sub)allowable sequences. This result may be helpful in determining minimal sets of permutations which are not realizable by configurations of points, and in other related problems.

Keywords: Sequences of permutations, allowable sequences, suballowable sequences, configurations of points

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2006/783

On security of statistical databases

Harout Aydinian (Universität Bielefeld, D)

A Statistical Database (SDB) is a database that returns only statistical information (such as averages, sums etc.), rather than actual records in the database, to user queries for statistical data analysis.

Sometimes, by correlating enough statistics, protected data about an individual can be inferred. When users can infer protected information in the SDB from responses to queries, the SDB is said to be compromised. The SDB security problem is to limit the use of the SDB (introducing certain restrictions) for the prevention of compromise. The goal is to maximize the number of available queries without compromise. One of the natural restrictions for the prevention of database compromise is to allow only SUM queries, that is only certain sums of individual records are available for the users. We discuss security problems for databases where only SUM queries with certain constraints are allowed.

Assume there are n numeric records $\{z_1, \ldots, z_n\}$ stored in a database. The problem is to find the largest number of subset sums of $\{z_1, \ldots, z_n\}$ (may be with some other constraints) that can be disclosed such that none of numbers z_i (or even sums of small subsets) can be determined from these sums. We present tight bounds for this number under constraints on size or dimension of query subsets.

Keywords: Statistical database, database security, sum queries

Joint work of: Ahlswede, Rudolf; Aydinian, Harout

Capacity of Quantum Arbitrarily Varying Channels

Vladimir Blinovsky (Russian Academy of Sciences - Moscow, RUS)

We prove that the average error capacity C_q of a quantum arbitrarily varying channel (QAVC) equals 0 or else the random code capacity $\bar{C}$ (Ahlswede's dichotomy)

Keywords: Arbitrarily varying channel, capacity, quantun channel

Joint work of: Blinovsky, Vladimir; Ahlswede Rudolf

Non–binary error correcting codes with noiseless feedback, localized errors, or both

Christian Deppe (Universität Bielefeld, D)

We investigate non–binary error correcting codes with noiseless feedback, localized errors, or both. It turns out that the Hamming bound is a central concept.

For block codes with feedback we present here a coding scheme based on an idea of erasions, which we call the **rubber method**. It gives an optimal rate for big error correcting fraction τ ($> \frac{1}{q}$) and infinitely many points on the Hamming bound for small τ.

We also consider variable length codes with all lengths bounded from above by n and the end of a word carries the symbol $\square$ and is thus recognizable by the decoder. For both, the $\square$-model with feedback and the $\square$-model with localized errors, the Hamming bound is the exact capacity curve for $\tau < 1/2$. Somewhat surprisingly, whereas with feedback the capacity curve coincides with the Hamming bound also for $1/2 \leq \tau \leq 1$, in this range for localized errors the capacity curve equals 0.

Also we give constructions for the models with both, feedback and localized errors.

Keywords: Error-correcting codes, localized errors, feedback, variable length codes

Joint work of: Ahlswede, Rudolf; Deppe, Christian; Lebedev, Vladimir

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2006/784

See also: submitted to Annals of EAS

Subwords in reverse-complement order

Péter Erdős (Alfréd Rényi Inst. of Mathematics - Budapest, H)

We examine finite words over an alphabet $\Gamma = \{a, \bar{a}; b, \bar{b}\}$ of pairs of letters, where each word $w_1 w_2 ... w_t$ is identical with its *reverse complement* $\bar{w}_t ... \bar{w}_2 \bar{w}_1$ (where $\bar{\bar{a}} = a, \bar{\bar{b}} = b$). We seek the smallest k such that every word of length n, composed from Γ, is uniquely determined by the set of its subwords of length up to k. Our almost sharp result ($k \sim 2n/3$) is an analogue of a classical result for "normal" words.

This classical problem originally was posed by M.P. Schützenberger and I. Simon, and gained a considerable interest for several researchers, foremost by Vladimir Levenshtein.

Our problem has its roots in bioinformatics.

Keywords: Reverse complement order, Reconstruction of words, Microarray experiments

Joint work of: Erdős, Péter; Ligeti, Péter; Sziklai, Péter; Torney, David C.

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2006/785

Some applications of pattern discovery in structural bioinformatics and biochemical networks

David Roger Gilbert (University of Glasgow, GB)

I describe some pattern discovery techniques and their applications for protein structures at the topological level. I give a quick introduction to biological background to bioinformatics.

Then I introduce some pattern discovery techniques and their applications for protein structures at the topological level, describing:
* Definition of data structure
* Pattern matching
* Pattern discovery
* Compression measure
* Structure comparison

Keywords: Bioinformatics, pattern discovery, protein structure, topology

L-Identification for Sources

Christian Heup (Universität Bielefeld, D)

In classical source coding the output of a source is encoded in such a way that it is uniquely decodable so that the decoder knows what the output was. If there exists a probability distribution on the set of all possible outputs we can compute the expected length of the codewords. It is a well-known fact that this expected length is lower-bounded by Shannon's classical entropy:

$$H(p_1, ..., p_n) = -\sum_{i=1}^{n} p_i \cdot \log(p_i).$$

In identification for sources the decoder must be able to decide for any possible output v of the source whether the actual output u coincides with v or not. This done by comparing th codewords c_u and c_v "bit by bit". This means that if the first both letters of c_u and c_v are not we stop the identification process by answering "$u \neq v$". If th first letters are equal we compare the second both letters and so on. The "identification time of u and v" is then the number of comparisons until we know the answer.

Again, if we have a prob. dist. on the set of possible outputs we can compute (for a given code) the expected value of the number of comparisons. It was shown by Ahlswede in "Identification Entropy" that this is lower-bounded by:

$$H_{\mathrm{ID}}(p_1, ..., p_n) = 2(1 - \sum_{i=1}^{n} p_i^2).$$

In this talk we generalize the idea of identification. Instead of a single output of the source we consider vectors $u^L = (u_1, ..., u_L)$ of possible outputs and want to answer the question "$\exists\ i \in \{1, ..., L\} : v = u_i$?". We examine the expected value of the number of questions and establish the so-called "identification entropy of order L" as a lower-bound.

Keywords: Source-coding, Identification, Entropy

Search when the lie depends on the target

Gyula O.H. Katona (Alfréd Rényi Institute of Mathematics, H)

Let X be a finite set of n elements and let $x \in X$ be an unknown element what we want to find by asking questions of type "is $x \in A$?" for certain subsets $A \subset X$. In the traditional model with lie, the answer can be wrong (say once during the whole search) independently of x and A. In the present model the answer can be wrong only for some elements x.

Therefore one question is based on a partition of X of 3 classes: (A, L, B). If $x \in A$ then the answer is YES (1), in the case $x \in B$ the answer NO (0), while $x \in L$ may lead to both answers. Hence the answer 1 tells us that $x \in A \cup L$, the answer 0 contains the information $x \in B \cup L$.

The practical problem is to minimize the number of questions needed to find the unknow element. There are two different models. The first one is the adaptive search, where the choice of the next question (partition with three classes) may depend on the previous answers. The the lenght of the search is the number of questions in the worst case. The mathematical problem is to find the minimum of the lenght of the search if $|L| \geq k$ holds for every question, where k is a given integer. We found the best adaptive search for given n and k.

In the case of the non-adaptive search the set of questions $(A_i, L_i, B_i)(1 \leq i \leq m)$ is given in advance. The mathematical problem is to find the minimum of m under the condition that the answers for these questions determine the unknown element and $|L_i| \geq k(1 \leq i \leq m)$. The solution is a consequence of the following new statement: there is a matching in the graph of the n-dimensional cube which contains $\lfloor 2^{n-1}/n \rfloor$ edges in each direction.

Keywords: Search, Lie

Joint work of: Katona, Gyula O.H.; Tichler, Krisztián

Vertex reconstruction in Cayley graphs

Elena Konstantinova (Sobolev Institute of Mathematics - Novosibirsk, RUS)

We present recent results in the reconstruction of vertices in Cayley graphs $Cay(G, S)$ where the symmetric group S_n and the hyperoctahedral group $\mathbb{Z}_2 \wr S_n$ are considered as a group G.

The generating sets S are specified by the applications in coding theory, molecular biology and computer science. Permutations as well as signed permutations are used to represent sequences of genes in chromosomes, and global rearrangements like reversals (interval inversions) and transpositions correspond to evolutionary changes.

Permutations are also used in the representations of interconnection networks which are modeled by Cayley graphs generating by transpositions (star graphs, bubble sort graphs, transposition graphs) and reversals (unburnt and burnt pancake graphs).

We investigate the structural properties of transposition and reversal Cayley graphs. It is proved that the reversal Cayley graph on B_n generated by reversals with changing sign does not contain C_3, C_5 nor bipartite subgraphs $K_{2,3}$ and contains C_4. The reversal Cayley graph on S_n does not contain C_3 nor bipartite subgraphs $K_{2,4}$ and contains bipartite subgraphs $K_{3,3}$. We show that for any $n \geq 2$ an unknown signed permutation considered as a vertex of the reversal Cayley graph on B_n is uniquely reconstructible from any 3 distinct signed permutations being at the reversal distance at most one from the unknown signed permutation. Under the same conditions for any $n \geq 3$ an unknown permutation considered as a vertex of the reversal Cayley graph on S_n is uniquely reconstructible from 4 distinct permutations. The similar results are presented for the transposition Cayley graphs on S_n and B_n.

Vertex reconstruction in Cayley graphs

Elena Konstantinova (Sobolev Institute of Mathematics - Novosibirsk, RUS)

In this report paper we collect recent results on the vertex reconstruction in Cayley graphs $Cay(G, S)$. The problem is stated as the problem of reconstructing a vertex from the minimum number of its r-neighbors that are vertices at distance at most r from the unknown vertex. The combinatorial properties of Cayley graphs on the symmetric group S_n and the signed permutation group B_n with respect to this problem are presented. The sets of generators of S are specified by applications in coding theory, computer science, molecular biology and physics.

Keywords: Reconstruction problems, Cayley graphs, the symmetric group, the signed permutation group, sorting by reversals, pancake problem

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2006/786

Algorithms for finding small dominating sets in cubic connected graphs

Alexandr V. Kostochka (Univ. of Illinois - Urbana, USA)

In 1996, Reed proved that the domination number $\gamma(G)$ of every n-vertex graph G with minimum degree at least 3 is at most $3n/8$ and conjectured that if G is cubic and connected, then $\gamma(G) \leq \lceil n/3 \rceil$. Kawarabayashi, Plummer and Saito proved that the conjecture is 'close to the truth' for cubic graphs of large girth. The proofs of the above results are algorithmic.

We disprove the conjecture by constructing a sequence $\{G_k\}_{k=1}^{\infty}$ of connected cubic graphs with $\lim\limits_{k\to\infty} \frac{\gamma(G_k)}{|V(G_k)|} \geq \frac{8}{23} = \frac{1}{3} + \frac{1}{69}$.

On the other hand, we gave a polynomial time algorithm that finds a dominating set of size at most $4n/11$ in any n-vertex cubic connected graph G with $n > 8$. Also, an improvement of the result by Kawarabayashi, Plummer and Saito will be discussed.

Keywords: Domination number, Cubic graphs

Joint work of: Kostochka, Alexandr; Stodolsky, Burak

Databases of Combinatorial Objects

Reinhard Laue (Universität Bayreuth, D)

An algorithmic approach to compute canonical forms for combinatorial objects of nested data types is presented. A database of graphs with search capability for isomorphism types accessible via the Internet is shown as an example together with applications to the visualization of t-designs.

Keywords: Canonical Forms, Combinatorial Objects, Graph Database

Joint work of: Kohnert, Axel; Wang, Min; Nedden, Max; Krec, Achim ; Laue, Reinhard

Databases of Combinatorial Objects

Reinhard Laue (Universität Bayreuth, D)

Canonical forms of combinatorial objects are used to retrieve these objects in a database. The data types of the objects may be built employing the constructors of forming sets, sequences, or cyclically permutable sequences arbitrarily deeply nested, during runtime. Thus, the approach is suitable for data from unknown sources like the Internet in an XML-format. The programs have been tested with t-designs, graphs, and chemical structures. Special visualizations of graphs are stored in a database prototype and are accessible in a client server version using any modern browser via the Internet. A request may specify a graph or an SQL where-statement concerning some other properties. The set of properties presently is small but will be extended. A canonical form is computed to identify the isomorphism type of a graph. The resulting graphs are displayed by a Java program. An XML-format is used for loading and saving graphs for further processing. The database includes solids derived from the Platonic solids, chemical graphs, and some lattices. It presently has about 100000 entries. We present an application to the visualization of some t-designs. Among these are the famous Witt-designs, several biplanes, and infinite series of resolvable 3-designs. Attempts are made to retrieve graphs that are similar to a given graph. Since we use graphs to display symmetries, graphs are manipulated to obtain a bigger automorphism group, so that they might be contained in our database. A successful search for a similar more symmetric graph may be used to take over the placement of the vertices or even be used for correcting a symmetry breaking error.

Keywords: Canonical form, Combinatorial Objects, Graphdrawings, Database, Resolvable Designs

Two conjectures on reconstruction of graphs

Vladimir I. Levenshtein (Keldysh Institute - Moscow, RUS)

New problems of the reconstruction of unknown simple connected graphs $G(V,E)$ with a set of V of labelled vertices are considered. It is assumed that for each vertex x of V the set $B_r(x,G)$ of all vertices at distance at most r $(r > 1)$ from x is available. A graph $G(V,E)$ is called r-reconstructible, if any graph $G'(V,E')$ (with the same set V of vertices) is identical to $G(V,E)$ (i.e., $E = E'$) under the condition that $B_r(x,G) = B_r(x,G')$ for all x of V. Analogously it is determined a graph $G(V,E)$ which is r-reconstructible up to isomorphism. Denote by $t(r)$ the minimum girth of graphs $G(V,E)$ without terminal vertices which are r-reconstructible. We conjecture that $t(r) = 2r + 3$ and find an upper bound to $t(r)$ which shows that this conjecture is true for $r = 2, 3, 4$ and 5. We also conjecture that any graph of girth at least 6 is 2-reconstructible up to isomorphism. It is proved that such a graph is 2-reconstructible exactly, if there exists an edge which does not belong to a hexagon. It is also proved that any hexagonal animal (a connected planar graph consisting of hexagons which either do not intersect or have a common edge) is 2-reconstructible up to isomorphism.

Solving Classical String Problems on Compressed Texts

Yury Lifshits (Steklov Inst. - St. Petersburg, RUS)

How to solve string problems, if instead of input string we get only program generating it?

Is it possible to solve problems faster than just "generate text + apply classical algorithm"?

In this paper we consider strings generated by straight-line programs (SLP). These are programs using only assignment operator. We show new algorithms for equivalence, pattern matching, finding periods and covers, computing fingerprint table on SLP-generated strings. From the other hand, computing the Hamming distance is NP-hard.

Main corollary is an $O(n2{*}m)$ algorithm for pattern matching in LZ-compressed texts.

Keywords: Pattern matching, Compressed text

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2006/798

Tiling Periodicity

Yury Lifshits (Steklov Inst. - St. Petersburg, RUS)

According to tiling periodicity the string XXYY is periodic, since it could be covered by two copies of partially defined string X_Y. We examine this notion, study its properties and present an algorithm for finding minimal tiling periods.

Keywords: Periodicity

Notes on the double cut and join operation in genome rearrangement

Julia Mixtacki (Universität Bielefeld, D)

Genome rearrangements have been modeled by a variety of operations such as inversions, translocations, transpositions, block interchanges, fusions and fissions.

The problem of sorting multichromosomal genomes can be stated as: Given two genomes A and B, the goal is to find a shortest sequence of rearrangement operations that transforms A into B. The length of such a shortest sequence is called the distance between A and B. Clearly, the solutions depend on what kind of rearrangement operations are allowed.

Given their prevalence in eukaryotic genomes, the usual choices of operations includes translocations, fusions, fissions and inversions. However, there are some indications that transpositions should also be included in the set of operations, but the lack of theoretical results showing how to include transpositions in the models led to algorithms that simulate transpositions as sequences of inversions.

The double cut and join operation (also called DCJ operation), introduced by Yancopoulos et al. (2005), allows to model all the classical operations while simplifying the algorithms. In this talk, I show that this operation can be extended to the most general type of genomes with a mixed collection of linear and circular chromosomes. I will also describe a graph structure that allows simplifying the theory and distance computation considerably, as neither a capping nor a concatenation of the linear chromosomes is necessary.

Gapped Permutation Patterns with Bounded Length

Laxmi Parida (IBM TJ Watson Research Center, USA)

Permutations on sequences have been successfully applied to model gene clusters on genomes for the purposes of comparing genomes and also discovering orthologs amongst other applications. Here we explore the problem of discovering gapped permutation patterns with a bounded length. It turns out that bounding the length of the permutation pattern leads to an efficient algorithm design. We present an efficient algorithm along the lines of ordered enumeration of patterns.

Keywords: Pattern discovery, data mining, clusters, patterns, motifs, permutation patterns, gapped permutation patterns

On the Monotonicity of the String Correction Factor for Words with Mismatches

Cinzia Pizzi (University of Helsinki, FIN)

The string correction factor is the term by which the probability of a word w needs to be multiplied in order to account for character changes or "errors" occurring in at most k arbitrary positions in that word. The behavior of this factor, as a function of k and of the word length, has implications on the number of candidates that need to be considered and weighted when looking for subwords of a sequence that present unusually recurrent replicas within some bounded number of mismatches. Specifically, it is seen that over intervals of mono- or bitonicity for the correction factor, only some of the candidates need be considered.

This mitigates the computation and leads to tables of over-represented words that are more compact to represent and inspect. In recent work, expectation and score monotonicity has been established for a number of cases of interest, under *i.i.d.* probabilistic assumptions. The present paper reviews the cases of bi-tonic behavior for the correction factor, concentrating on the instance in which the question is still open.

Keywords: Pattern discovery, Motif, Over-represented word, Monotone score, Correction Factor

Joint work of: Apostolico, Alberto; Pizzi, Cinzia

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2006/789

Universal Codes as a Basis for Time Series Testing

Boris Ryabko (Russian Academy of Sc. - Novosibirsk, RUS)

We suggest a new approach to hypothesis testing for ergodic and stationary processes. In contrast to standard methods, the suggested approach gives a possibility to make tests, based on any lossless data compression method even if the distribution law of the codeword lengths is not known. We apply this approach to the following four problems: goodness-of-fit testing (or identity testing), testing for independence, testing of serial independence and homogeneity testing and suggest nonparametric statistical tests for these problems. It is important to note that practically used so-called archivers can be used for suggested testing.

Joint work of: Ryabko, Boris; Astola, Jaakko

Sequence prediction for non-stationary processes

Daniil Ryabko (IDSIA - Lugano, CH)

We address the problem of sequence prediction for nonstationary stochastic processes. In particular, given two measures on the set of one-way infinite sequences over a finite alphabet, consider the question whether one of the measures predicts the other. We find some conditions on local absolute continuity under which prediction is possible.

Keywords: Sequence prediction, probability forecasting, local absolute continuity

Joint work of: Ryabko, Daniil; Hutter, Marcus

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2006/790

Local Minimax Learning of Approximately Polynomial Functions

Konstantin Rybnikov (Univ. of Massachusetts - Lowell, USA)

Suppose we have a number of noisy measurements of an unknown real-valued function f near point of interest $\mathbf{x}_0 \in \mathbb{R}^d$. Suppose also that nothing can be assumed about the noise distribution, except for zero mean and bounded covariance matrix. We want to estimate f at $\mathbf{x} = \mathbf{x}_0$ using a general linear parametric family $f(\mathbf{x}; \mathbf{a}) = a_0 h_0(\mathbf{x}) + ... + a_q h_q(\mathbf{x})$, where $\mathbf{a} \in \mathbb{R}^q$ and h_i's are bounded functions on a neighborhood B of $\mathbf{x}_0$ which contains all points of measurement. Typically, B is a Euclidean ball or cube in $\mathbb{R}^d$ (more generally, a ball in an l_p-norm). In the case when the h_i's are polynomial functions in $x_1, \ldots, x_d$ the model is called locally-polynomial. In particular, if the h_i's form a basis of the linear space of polynomials of degree at most two, the model is called locally-quadratic (if the degree is at most three, the model is locally-cubic, etc.). Often, there is information, which is called context, about the function f (restricted to B) available, such as that it takes values in a known interval, or that it satisfies a Lipschitz condition. The theory of local minimax estimation with context for locally-polynomial models and approximately locally polynomial models has been recently initiated by Jones. In the case of local linearity and a bound on the change of f on B, where B is a ball, the solution for squared error loss is in the form of ridge regression, where the ridge parameter is identified; hence, minimax justification for ridge regression is given together with explicit best error bounds. The analysis of polynomial models of degree above 1 leads to interesting and difficult questions in real algebraic geometry and non-linear optimization.

We show that in the case when f is a probability function, the optimal (in the minimax sense) estimator is effectively computable (with any given precision), thanks to Tarski's elimination principle.

Joint work of: Rybnikov, Konstantin; Jones, Lee

Counting Suffix Arrays and Strings

Jens Stoye (Universität Bielefeld, D)

Suffix arrays are used in various application and research areas like data compression or computational biology. In this work, our goal is to characterize the combinatorial properties of suffix arrays and their enumeration. For fixed alphabet size and string length we count the number of strings sharing the same suffix array and the number of such suffix arrays. Our methods have applications to succinct suffix arrays and build the foundation for the efficient generation of appropriate test data sets for suffix array based algorithms. We also show that summing up the strings for all suffix arrays builds a particular instance for some summation identities of Eulerian numbers.

Some Results for Identification for Sources and its Extension to Liar Models

Zlatko Varbanov (Veliko Tarnovo University, BG)

Let $(\mathcal{U}, P)$ be a source, where $\mathcal{U} = \{1, 2, \ldots, N\}, P = \{P_1, P_2, \ldots, P_N\}$, and let $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$ be a binary prefix code (PC) for this source with $||c_u||$ as length of c_u. Introduce the random variable U with $\text{Prob}(U = u) = p_u$ for $u = 1, 2, \ldots, N$ and the random variable C with $C = c_u = (c_1, c_2, \ldots, c_{u||c_u||})$ if $U = u$. We use the PC for noiseless identification, that is user u wants to know whether the source output equals u, that is, whether C equals c_u or not.

The user iteratively checks whether C coincides with c_u in the first, second, etc. letter and stops when the first different letter occurs or when $C = c_u$. What is the expected number $L_{\mathcal{C}}(P, u)$ of checkings?

In order to calculate this quantity we introduce for the binary tree $T_{\mathcal{C}}$, whose leaves are the codewords $c_1, c_2, \dots, c_N$, the sets of leaves $\mathcal{C}_{ik}(1 \leq i \leq N; 1 \leq k)$, where $\mathcal{C}_{ik} = \{c \in \mathcal{C} : c$ coincides with c_i exactly until the k'th letter of $c_i\}$.

If C takes a value in $\mathcal{C}_{uk}, 0 \leq k \leq ||c_u|| - 1$, the answers are k times "Yes" and 1 time "No". For $C = c_u$ the

$$L_{\mathcal{C}}(P, u) = \sum_{k=0}^{||c_u||-1} P(C \in \mathcal{C}_{uk})(k+1) + ||c_u||P_u.$$

For code $\mathcal{C}$, $L_{\mathcal{C}}(P) = \max L_{\mathcal{C}}(P, u)$, $1 \geq u \geq N$, is the expected number of checkings in the worst case and $L(P) = \min L_{\mathcal{C}}(P)$ is this number for the best code $\mathcal{C}$.

Let $P = P^N = \{\frac{1}{N}, \dots, \frac{1}{N}\}$. We construct a prefix code $\mathcal{C}$ in the following way. In each node (starting at the root) we split the number of remaining codewords in proportion as close as possible to $(\frac{1}{2}, \frac{1}{2})$.

It is known that

$$\lim_{N \to \infty} L_{\mathcal{C}}(P^N) = 2$$

(Ahlswede, Balkenhol, Kleinewachter, 2003)

We know that $L(P) \leq 3$ for all P (Ahlswede, Balkenhol, Kleinewachter, 2003). Also, the problem to estimate an universal constant $A = \sup L(P)$ for general $P = (P_1, \dots, P_N)$ was stated (Ahlswede, 2004). We compute this constant for uniform distribution and this code $\mathcal{C}$.

$$\sup_N L_{\mathcal{C}}(P^N) = 2 + \frac{log_2(N-1) - 2}{N}$$

Also, we consider the average number of checkings, if code $\mathcal{C}$ is used: $L_{\mathcal{C}}(P, P) = \sum P_u L_{\mathcal{C}}(P, u)$, for $u \in \mathcal{U}$. We calculate the exact values of $L_{\mathcal{C}}(P^N)$ and $L_{\mathcal{C}}(P^N, P^N)$ for some N.

Other problem is the extension of identification for sources to liar models. We obtain a upper bound for the expected number of checkings $L_{\mathcal{C}}(P^N; e)$, where e is the maximum number of lies.

$$L_{\mathcal{C}}(P^N; e) \leq M_{\mathcal{C}}(P^N; e) = (e+1)L_{\mathcal{C}}(P^N) + e; \quad \lim_{N \to \infty} M_{\mathcal{C}}(P^N; e) = 3e + 2$$

Keywords: Identification for sources, lies prefix code

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2006/782

Buffer Management with limited knowledge of future arrivals

Christian Wischmann (Universität Bielefeld, D)

In Buffer Management for Network Switches we deal with a switch having m input ports and n output ports. At each time step t, a unlimited number of data packets may arrive at each input port and at most 1 can be transmitted from each output port. Every port is equipped with a buffer that can store a fixed number of packets. If a buffer is full while new packets arrive, these packets are lost. There is a multitude of different models of this problem. The packets may have varying value or size, for example. The quality of a buffer management algorithm is measured by its competitive ratio $c \leq T_{OPT}(\sigma)/T_{ALG}(\sigma)$, for all finite packet arrival sequences σ, where T denotes the total number of packets transmitted by the regarded online algorithm and an optimal offline algorithm.

In this talk we consider the possibility of the online algorithm to have a limited knowledge of the future packet arrivals: it is able to see the packet arrivals of the next φ time steps. Thus it approaches the abilities of the offline algorithm, which has $\varphi = \infty$. Our model is quite simple, our packets have unit size and value and all input buffers have the same size B. Since we set $n = 1$ and have no output buffer, we need not bother about scheduling between input and output ports. An algorithm only has to decide which non-empty port to serve at each time step.

Having φ knowledge of the future we show, that for $\varphi = B \cdot (m-1)$ all work-conserving online algorithms are optimal. For general φ we show there exist online algorithms satisfying the upper bound $c \leq 2 - \frac{2\varphi+1}{2Bm}$.

Keywords: Buffer Management, Creating Order, Competitive Analysis