04161 Abstracts Collection
Detecting Local Patterns
— Dagstuhl Seminar —

Jean-François Boulicaut[1], Katharina Morik[2] and Arno Siebes[3]

[1] INSA Lyon, FR
Jean-Francois.Boulicaut@insa-lyon.fr
[2] Univ. Dortmund, DE
morik@ls8.informatik.uni-dortmund.de
[3] Utrecht University, NL
siebes@cs.uu.nl

Abstract. From 12.04.04 to 16.04.04, the Dagstuhl Seminar 04161 "Detecting Local Patterns" 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.

Detecting Cheating

Niall Adams (Imperial College London, GB)

Students sometimes cheat. In particular, they sometimes copy coursework assignments from each other. Such copying is often detected by the markers, since the copied script and the original will be unusually similar. However, one cannot rely on such subjective assessment - perhaps there are many scripts or perhaps the student has sought to disguise the copying by changing words or other aspects of the answers. We describe an attempt to develop a pattern discovery method for detecting cheating, based on measures of the similarities between scripts, where similarity is defined in syntactic rather than semantic terms. This problem differs from many other pattern discovery problems because the peaks will typically be very low: normally only one or two cheating students will copy from any given other student.

Dagstuhl Seminar Proceedings 04161
Detecting Local Patterns
http://drops.dagstuhl.de/opus/volltexte/2006/467

Comparison of sequential pattern discovery algorithms in finding patterns in text

Helena Ahonen-Myka (University of Helsinki, FIN)

Several methods for discovering sequential patterns have been in the past. Each method usually has some special application or type of data in mind, although the methods are certainly general and applicable in several domains. In this paper we compare some of these methods on the basis of how they could be applied to mining sequential patterns in text. We discuss first the underlying assumptions concerning features of the data and patterns, for instance, representation of the data, size of the alphabet (set of items, set of words), length of the input sequences, number of the input sequences, and sparseness of the frequent patterns in the data. We also try to characterize the text domain using these features and hence formulate a set of requirements for this domain. Also some possible constraints are discussed, e.g. maximum gap between the items of a sequence in the instances and maximum length of frequent sequences. We then compare the methods according to these requirements found. Finally we try to suggest some targets for development. The following methods, at least, are covered in this study: Ahonen-Myka. Discovery of frequent word sequences in text. The ESF Exploratory Workshop on Pattern Detection and Discovery in Data Mining, Imperial College, London, 16-19 September, 2002. Agrawal and Srikant. Mining sequential patterns. International Conference on Data Engineering, 1995. Mannila and Toivonen. Discovering generalized episodes using minimal occurrences. KDD 1996. Tsoukatos and Gunopulos. Efficient mining of spatiotemporal patterns. SSTD 2001. Zaki. SPADE: an efficient algorithm for mining frequent sequences, Machine Learning, 2000.

Recent results in constrained frequent pattern mining

Francesco Bonchi (CNR - Pisa, I)

Frequent itemsets play an essential role in many data mining tasks that try to find interesting patterns from databases, such as association rules, correlations, sequences, episodes, classifiers, clusters and many more. Although the collection of all frequent itemsets is typically very large, the subset that is really interesting for the user usually contains only a small number of itemsets. Therefore, the paradigm of constraint-based mining was introduced. Constraints provide focus on the interesting knowledge, thus reducing the number of patterns extracted to those of potential interest. Additionally, they can be pushed deep inside the mining algorithm in order to achieve better performance. For this reason the problem of how to push different types of constraints into the frequent itemsets computation has been studied a lot. In this talk we will describe our recent algorithmic results, namely ExAnte and ExAMiner, which were introduced in 2003 in order to exploit monotone constraints in frequent pattern computation.

As a brand new result, we will show how tough constraints can be pushed in such Apirori-like computation.

Joint work of: Bonchi, Francesco; Giannotti, Fosca

Towards Inductive Databases for gene Expression Data Analysis

Jean-François Boulicaut (INSA - Lyon, F)

We are designing new data mining techniques on gene expression data, more precisely techniques that provide a priori interesting bi-sets, i.e., sets of biological situations (or objects) and associated sets of genes (or attributes). The so-called (formal) concepts are important special cases of a priori interesting bi-sets in derived boolean expression matrices, e.g., matrices that encode over-expression of genes. Indeed, using various sources of information (e.g., gene ontology), we can post-process extracted bisets and provide putative transcription modules, i.e., one of the typical knowledge molecular biologists are looking for. In this talk, we will not only specify prototypical gene expression data analysis tasks within the inductive database framework but also illustrate how the recent work on itemset or sequential pattern constraint-based mining provides more or less generic algorithms for an efficient evaluation of biologically relevant inductive queries. We will consider real-life examples from ongoing research projects with molecular biologists. The first one concerns the analysis of global human SAGE data (human gene expression in around 100 different biological situations) and the second one concerns human gene expression data (DNA chip) before and after insulino administration. This research is a joint work with Jérémy Besson, Ruggero Pensa and several biologists: Sylvain Blachon, Olivier Gandrillon, and Sophie Rome.

Keywords: Local pattern discovery, inductive databases, bioinformatics,gene expression data analysis, frequent patterns, closed sets, sequences, constraint-based mining

Local and Global Evaluation in Rule Learning

Johannes Fürnkranz (Darmstadt, D)

Separate-and-conquer or covering rule learning algorithms learn rule sets one rule at a time. Traditionally, the learned rules are evaluated locally, i.e., solely with respect to the number of positive and negative examples that are covered by the rule. Global context is provided by removing the examples that are covered by previous rules. In the first part of the talk, we will briefly review recent results on the relation between local and global evaluation in covering algorithms. In the second part of the talk, we will argue that conventional rule learning heuristics have at least two fundamental shortcomings, which should be addressed by a rule learning heuristic: 1) performance measures on the training data are optimistically biased 2) incomplete rules should not be assessed by their own performance, but by their potential of being refined into a high-quality rule We will then report on first results for addressing the first of these two questions by meta-learning a predictor for the true accuracy of a rule, and will briefly discuss our ideas for investigating the second problem.

Keywords: Inductive rule learning, evaluation metrics

A Frequent Pattern Query Language with Optimizations

Fosca Giannotti (CNR - Pisa, I)

In this paper we study data mining query language and optimizations in the context of a Logic-based Knowledge Discovery Support Environment. i.e., a flexible knowledge discovery system with capabilities to obtain, maintain, represent, and utilize both induced and deduced knowledge. In particular, we focus on frequent pattern queries, since this kind of query is at the basis of many mining tasks, and it seems appropriate to be encapsulated in a knowledge discovery system as a primitive operation. We introduce an inductive language for frequent pattern queries, which is simple enough to be highly optimized and expressive enough to cover the most of interesting queries. Then we define an optimized constraint-pushing operational semantics for our inductive language. This semantics is based on a frequent pattern mining operator, which is able to exploit as much as possible the given set of constraints, and which can adapt its behavior to the characteristics of the given input set of data.

Miming Frequent Queries

Bart Goethals (University of Antwerp, B)

During the last decade, a lot of algorithms have been developed for mining frequent several types of patterns in several types of databases. Nevertheless, a system that can efficiently mine patterns in an arbitrary relational database is still nonexistent. Inspired by the work done in ILP, we further explore the opportunity of mining conjunctive queries. In this paper, we describe different approaches to efficiently mine queries belonging to small and simple subclasses of conjunctive queries, and show that still a lot of interesting patterns can be expressed as such.

Local Patterns in Time Series

Frank Höppner (FH Wolfenbüttel, D)

The starting point for this paper is the definition of pattern detection given by David Hand: pattern detection is the unsupervised detection of local regions with anomalously high data density, which represent real underlying phenomena. We discuss some aspects of his definition and the close relationship between clustering and pattern detection, before we investigate how to utilize clustering algorithms for pattern detection. An extension of an existing clustering algorithm is proposed to identify local patterns that are flagged as being significant according to some statistical measure on the data density.

Querying an Inductive Database for Frequent Patterns

Rosa Meo (University of Torino, I)

The problem of mining association rules from very large databases and, more generally, that of extracting frequent sets satisfying user defined constraints has been widely investigated in the last decade.

Constraint-based mining languages are the main key factor of inductive databases proposed by Mannila and Imielinski, in order to leverage decision support systems. Inductive databases will become really effective only when efficient optimizers for the mining languages will be available, i.e., if it will be possible to execute a query exploiting the available information in the database, such as the constraints in the schema, the indices or the results of previously executed queries. In this talk, we present a well-known query language proposed in 1996, MINE RULE, to extract association rules from relational databases. We show the flexibility and the expressive power of this mining language by applying it to some practical problems, such as WEB log analysis and online trading. Queries extracting frequent itemsets are today called iceberg queries and are generally very expensive to compute. As a consequence, in order to speed up execution time it makes sense to try to factorize the effort already done by the DBMS with previous queries.

In this talk we present the problem of query rewriting for frequent itemset mining, that is the determination of a relational expression on a set of queries whose result is equivalent to the result of a given query for every database on the same schema. In the past, query rewriting has been widely used in relational databases, in data warehouses and in statistical database systems. In this talk we present conditions under which query rewriting in constraint-based mining languages is possible. We show that the proposed approach is feasible and advantageous, by means of some experiments with a prototype optimizer. At the moment, the implemented optimizer recognizes equivalent queries, and exploits such equivalences to avoid heavy computations.

Furthermore, it can be easily extended to recognize query containment, and saving heavy computations in this case as well.

Keywords: Inductive databases, constraint based languages, query equivalence, query containment, dominance, functional dependency

Visualizing Very Large Graphs using Clustering Neighborhoods

Dunja Mladenic (Jozef Stefan Institute - Ljubljana, SLO)

We presents a method for visualization of collection of large graphs, such as a collection of Web pages in a 2D space.

The main contribution here is in representation change to enable better handling of the data. The idea of the method consists from three major steps: (1) First, we transform a graph into a sparse matrix, where for each vertex in the graph there is one sparse vector. Sparse vectors have non-zero components for the vertices linked closely to the vertex represented by the vector. (2) Next, we perform hierarchical clustering (e.g. hierarchical K-Means) on the set of sparse vectors resulting in the hierarchy of clusters. (3) In the last step, we map hierarchy of clusters into a 2D space in the way that more similar clusters appear closely on the picture.

The effect of the whole procedure is that we assign unique X and Y coordinates to each vertex, in a way that vertices or groups of vertices on several levels of hierarchy that are stronger connected in a graph are place closer in the picture. The method is particular useful for power distributed graphs. We show applications of the method on several examples including visualization of a large web site graph, company collaboration graph and cross-sell recommendation graph.

Joint work of: Mladenic, Dunja; Grobelnik, Marko

Features for Learning Local Pattern in Time-Stamped Data

Katharina Morik (Universität Dortmund, D)

Time-stamped data occur frequently in real-world databases. The goal of analysing time-stamped data is very often to find a small group of objects (customers, machine parts,...) which is important for the business at hand. In contrast, the majority of objects obey well-known rules and is not of interest for the analysis. In terms of a classification task, the small group means that there are very few positive examples and within them, there is some sort of a structure such that the small group differs significantly from the majority. We may consider such a learning task learning a local pattern.

Depending on the goal of the data analysis, different aspects of time are relevant, e.g., the particular date, the duration of a certain state, or the number of different states. From the given data, we may generate features that allow us to express the aspect of interest. Here, we investigate the aspect of state change and its representation for learning local patterns in time-stamped data. Besides a simple Boolean representation indicating a change, we use frequency features from information retrieval. We transfer Joachim's theory for text classification to our task and inestigate its fit to local pattern learning. The approach has been implemented within the MiningMart system and was successfully applied to real-world insurance data.

Joint work of: Morik, Katharina; Köpke, Hanna

Modelling the noise to find patterns

Arno Siebes (Utrecht University, NL)

David Hand characterized patterns as a component between a model and the random component. In the analysis of micro-array data, much of the "signal" seems to be masked by noise. In other words, to detect the patterns, we first have to "subtract" the noise from the data. To do make this work we make a modell of the noise and look for patterns on top of this model.

Finding Correspondence Patterns Between Partitions

Daniel Sánchez (Universidad de Granada, E)

In some occasions, data comes in the form of (or can be interpreted as) a set of objects on which several partitions are defined. The different partitions could be obtained as a result of different automatic clustering processes, or they can be provided by one or more experts on the universe of objects, even by using different criteria. In this framework, a correspondence pattern is a kind of matching between partitions on the basis of the matching between groups in each partition.

This kind of pattern provides useful information for several tasks, such as data fusion problems, measurement of matching degrees between classifications, and estimation of resemblance between expert criteria. In this work we introduce several kinds of correspondences and relations between them. We also show how they can be interpreted as association rules on different sets of transactions, each set obtained from the collection of partitions by different translation processes.

Joint work of: Sanchez, D.; Delgado, M.; Serrano, J.M.; Vila, M.A.

Subgroup Discovery: A summary of recent work on scalability and significance

Stefan Wrobel (Fraunhofer Inst. - St. Augustin, D)

The task of subgroup discovery consists of finding, within a given population space, subgroups with unusual statistical properties. Subgroup discovery is thus a particular kind of local pattern discovery, and has been of interest in the KDD community ever since the pioneering work of Kloesgen and others.

In this talk, I will summarize a line of research on subgroup discovery that has primarily focused on finding subgroup discovery algorithms that are scalable to larger data sets. Firstly, I will describe the algorithmic optimizations made in Midos, a first-order ("relational") subgroup discovery system. I will then report on joint work published with Tobias Scheffer over the past years on achieving scalability through sampling.

Interestingly, besides leading to a very fast algorithm, this also results in an approach that can provide guarantees against spurious or random discoveries despite examining large hypothesis spaces.

Local Patterns: What Do They Really Mean?

Liu Xiaohui (Brunel University, GB)

Local patterns are deviations from a background model, and are detected on the basis of data configurations relative to the model. There have been a large body of work in both statistical and AI communities on how to detect such patterns, and to a certain extent, on testing whether the patterns are simply due to distortion in the data or random variation because of the stochastic data generating process. What appears lacking, however, is the effort to understand what these patterns really mean in a practical setting once they are identified. In this talk, we explore the issues concerned with relating meanings to local patterns and explore how much we could automate this important task. Real-world applications will be used to help formulate the concepts, particularly those from biomedical domains where local patterns are common and they are often mixed with noise.