# Doctoral dissertations

# Using and Extending Itemsets in Data Mining: Query Approximation, Dense Itemsets, and Tiles

**Jouni K. Seppänen**

*Dissertation for the degree of Doctor of Science in Technology on 31 May 2006.*

**External examiners:**
Gautam Das (University of Texas at Arlington)
Bart Goethals (University of Antwerp)
**Opponent:**
Dimitrios Gunopulos (University of California at Riverside)

**Abstract:**
Frequent itemsets are one of the best known concepts in data mining, and there is active research in itemset mining algorithms. An itemset is frequent in a database if its items co-occur in sufficiently many records. This thesis addresses two questions related to frequent itemsets. The first question is raised by a method for approximating logical queries by an inclusion-exclusion sum truncated to the terms corresponding to the frequent itemsets: how good are the approximations thereby obtained? The answer is twofold: in theory, the worst-case bound for the algorithm is very large, and a construction is given that shows the bound to be tight; but in practice, the approximations tend to be much closer to the correct answer than in the worst case. While some other algorithms based on frequent itemsets yield even better approximations, they are not as widely applicable.

The second question concerns extending the definition of frequent itemsets to relax the requirement of perfect co-occurrence: highly correlated items may form an interesting set, even if they never co-occur in a single record. The problem is to formalize this idea in a way that still admits efficient mining algorithms. Two different approaches are used. First, dense itemsets are defined in a manner similar to the usual frequent itemsets and can be found using a modification of the original itemset mining algorithm. Second, tiles are defined in a different way so as to form a model for the whole data, unlike frequent and dense itemsets. A heuristic algorithm based on spectral properties of the data is given and some of its properties are explored.

# Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition

**Mathias Creutz**

*Dissertation for the degree of Doctor of Science in Technology on 15 June 2006.*

**External examiners:**
Richard Wicentowski (Swarthmore College, Pennsylvania)
Jukka Heikkonen (Helsinki University of Technology)
**Opponents:**
James H. Martin (University of Colorado)
Wray Buntine (Helsinki Institute of Information Technology)

**Abstract:**
In order to develop computer applications that successfully process natural language data (text and speech), one needs good models of the vocabulary and grammar of as many languages as possible. According to standard linguistic theory, words consist of morphemes, which are the smallest individually meaningful elements in a language. Since an immense number of word forms can be constructed by combining a limited set of morphemes, the capability of understanding and producing new word forms depends on knowing which morphemes are involved (e.g., "water, water+s, water+y, water+less, water+less+ness, sea+water").

Morpheme boundaries are not normally marked in text unless they coincide with word boundaries. The main objective of this thesis is to devise a method that discovers the likely locations of the morpheme boundaries in words of any language. The method proposed, called Morfessor, learns a simple model of concatenative morphology (word forming) in an unsupervised manner from plain text. Morfessor is formulated as a Bayesian, probabilistic model. That is, it does not rely on predefined grammatical rules of the language, but makes use of statistical properties of the input text.

Morfessor situates itself between two types of existing unsupervised methods: morphology learning vs. word segmentation algorithms. In contrast to existing morphology learning algorithms, Morfessor can handle words consisting of a varying and possibly high number of morphemes. This is a requirement for coping with highly-inflecting and compounding languages, such as Finnish. In contrast to existing word segmentation methods, Morfessor learns a simple grammar that takes into account sequential dependencies, which improves the quality of the proposed segmentations.

Morfessor is evaluated in two complementary ways in this work: directly by comparing to linguistic reference morpheme segmentations of Finnish and English words and indirectly as a component of a large (or virtually unlimited) vocabulary Finnish speech recognition system. In both cases, Morfessor is shown to outperform state-of-the-art solutions.

The linguistic reference segmentations were produced as part of the current work, based on existing linguistic resources. This has resulted in a morphological gold standard, called Hutmegs, containing analyses of a large number of Finnish and English word forms.

## Blind Source Separation for Interference Cancellation in CDMA Systems

**Karthikesh Raju**

*Dissertation for the degree of Doctor of Science in Technology on 11 August 2006.*

**External examiners:**
Lars Rasmussen (University of South Australia)
Asoke Nandi (University of Liverpool)
**Opponent:**
Jürgen Lindner (Universität Ulm)

**Abstract:**

Communication is the science of "reliable" transfer of information between two parties, in the sense that the information reaches the intended party with as few errors as possible. Modern wireless systems have many interfering sources that hinder reliable communication. The performance of receivers severely deteriorates in the presence of unknown or unaccounted interference. The goal of a receiver is then to combat these sources of interference in a robust manner while trying to optimize the trade-off between gain and computational complexity.

Conventional methods mitigate these sources of interference by taking into account all available information and at times seeking additional information e.g., channel characteristics, direction of arrival, etc. This usually costs bandwidth. This thesis examines the issue of developing mitigating algorithms that utilize as little as possible or no prior information about the nature of the interference. These methods are either semi-blind, in the former case, or blind in the latter case.

Blind source separation (BSS) involves solving a source separation problem with very little prior information. A popular framework for solving the BSS problem is independent component analysis (ICA). This thesis combines techniques of ICA with conventional signal detection to cancel out unaccounted sources of interference. Combining an ICA element to standard techniques enables a robust and computationally efficient structure. This thesis proposes switching techniques based on BSS/ICA effectively to combat interference. Additionally, a structure based on a generalized framework termed as denoising source separation (DSS) is presented. In cases where more information is known about the nature of interference, it is natural to incorporate this knowledge in the separation process, so finally this thesis looks at the issue of using some prior knowledge in these techniques. In the simple case, the advantage of using priors should at least lead to faster algorithms.

# Approaches for Content-Based Retrieval of Surface Defect Images

**Jussi Pakkanen**

*Dissertation for the degree of Doctor of Science in Technology on 20 October 2006.*

**External examiners:**
Matti Niskanen (University of Oulu)
Andreas Rauber (Vienna University of Technology)
**Opponent:**
Pasi Koikkalainen (University of Jyväskylä)

**Abstract:**

There are two properties which all industrial manufacturing processes try to optimize: speed and quality. Speed can also be called throughput and tells how much products can be created in a specified time. The higher speeds you have the better. Quality means the perceived goodness of the finished product. Broken or defective products simply don't sell, so they must be eliminated.

These are contradicting goals. The larger the manufacturing volumes, the less time there is to inspect a single product, or the more inspectors are required. A good example is paper manufacturing. A single paper machine can produce a sheet of paper several meters wide and several hundred kilometers long in just a few hours. It is impossible to inspect these kinds of volumes by hand.

In this thesis the indexing and retrieval of defect images taken by an automated inspection machine is examined. Some of the images taken contain serious defects such as holes, while others are less grave. The goal is to try to develop automated methods to find the serious fault images from large databases using only the information in the images. This means that there are no annotations. This is called content-based image retrieval, or CBIR.

This problem is examined in two different ways. First the PicSOM CBIR tool's suitability for this task is evaluated. PicSOM is a platform for content-based image retrieval developed at the Laboratory of Computer and Information Science, Helsinki University of Technology. PicSOM has earlier been succesfully applied to various different CBIR tasks.

The other part involves developing new algorithms for efficient indexing of large, high-dimensional databases. The Evolving Tree (ETree), a novel hierarchical, tree-shaped, self-organizing neural network is presented and analyzed. It is noticeably faster than classical methods, while still obtaining good results.

The suitability and performance of both CBIR and ETree on this problem is evaluated using several different experiments. The results show that both approaches are applicable for this real world quality inspection problem with good results.

# Advanced source separation methods with applications to spatio-temporal datasets

**Alexander Ilin**

*Dissertation for the degree of Doctor of Science in Technology on 3 November 2006.*

**External examiners:**
Mark A. Girolami, (University of Glasgow)
Aki Vehtari (Helsinki University of Technology)
**Opponent:**
Luis Borges de Almeida, (Technical University of Lisbon)

## Abstract:

Latent variable models are useful tools for statistical data analysis in many applications. Examples of popular models include factor analysis, state-space models and independent component analysis. These types of models can be used for solving the source separation problem in which the latent variables should have a meaningful interpretation and represent the actual sources generating data. Source separation methods is the main focus of this work.

Bayesian statistical theory provides a principled way to learn latent variable models and therefore to solve the source separation problem. The first part of this work studies variational Bayesian methods and their application to different latent variable models. The properties of variational Bayesian methods are investigated both theoretically and experimentally using linear source separation models. A new nonlinear factor analysis model which restricts the generative mapping to the practically important case of post-nonlinear mixtures is presented. The variational Bayesian approach to learning nonlinear state-space models is studied as well. This method is applied to the practical problem of detecting changes in the dynamics of complex nonlinear processes.

The main drawback of Bayesian methods is their high computational burden. This complicates their use for exploratory data analysis in which observed data regularities often suggest what kind of models could be tried. Therefore, the second part of this work proposes several faster source separation algorithms implemented in a common algorithmic framework. The proposed approaches separate the sources by analyzing their spectral contents, decoupling their dynamic models or by optimizing their prominent variance structures. These algorithms are applied to spatio-temporal datasets containing global climate measurements from a long period of time.

## Bayesian Inference in Nonlinear and Relational Latent Variable Models

**Tapani Raiko**

*Dissertation for the degree of Doctor of Science in Technology on 1 December 2006.*

**External examiners:**
Jouko Lampinen (Helsinki University of Technology)
Petri Myllymäki (University of Helsinki)
**Opponent:**
Ole Winther (Danmarks Tekniske Universitet)

**Abstract:**

Statistical data analysis is becoming more and more important when growing amounts of data are collected in various fields of life. Automated learning algorithms provide a way to discover relevant concepts and representations that can be further used in analysis and decision making.

Graphical models are an important subclass of statistical machine learning that have clear semantics and a sound theoretical foundation. A graphical model is a graph whose nodes represent random variables and edges define the dependency structure between them. Bayesian inference solves the probability distribution over unknown variables given the data. Graphical models are modular, that is, complex systems can be built by combining simple parts. Applying graphical models within the limits used in the 1980s is straightforward, but relaxing the strict assumptions is a challenging and an active field of research.

This thesis introduces, studies, and improves extensions of graphical models that can be roughly divided into two categories. The first category involves nonlinear models inspired by neural networks. Variational Bayesian learning is used to counter overfitting and computational complexity. A framework where efficient update rules are derived automatically for a model structure given by the user, is introduced. Compared to similar existing systems, it provides new functionality such as nonlinearities and variance modelling. Variational Bayesian methods are applied to reconstructing corrupted data and to controlling a dynamic system. A new algorithm is developed for efficient and reliable inference in nonlinear state-space models.

The second category involves relational models. This means that observations may have distinctive internal structure and they may be linked to each other. A novel method called logical hidden Markov model is introduced for analysing sequences of logical atoms, and applied to classifying protein secondary structures. Algorithms for inference, parameter estimation, and structural learning are given. Also, the first graphical model for analysing nonlinear dependencies in relational data, is introduced in the thesis.

# Compaction of C-Band Synthetic Aperture Radar Based Sea Ice Information for Navigation in the Baltic Sea

**Juha Karvonen**

*Dissertation for the degree of Doctor of Science in Technology on 8 December 2006.*

**External examiners:**
Markku Hauta-Kasari (University of Joensuu)
Matti Leppäranta (University of Helsinki)
**Opponent:**
Torbjörn Eltoft (University of Tromsö)

**Abstract:**

In this work operational sea ice synthetic aperture radar (SAR) data products were improved and developed. The main idea is to deliver the essential SAR-based sea ice information to end-users (typically on ships) in a compact and user-friendly format. The operational systems at Finnish Institute of Marine Research (FIMR) are based on the Canadian SAR-satellite Radarsat-1.

The operational sea ice classification, developed by the author with colleagues, has been further developed. An incidence angle correction algorithm to normalize the backscattering over the SAR incidence angle range for Baltic Sea ice has been developed. The algorithm is based on SAR backscattering statistics over the Baltic Sea.

A SAR segmentation algorithm based on pulse-coupled neural networks has been developed and tested. The parameters have been tuned suitable for the operational data in use at FIMR. The sea ice classification is based on this segmentation and the classification is segment-wise rather than pixel-wise.

To improve distinguishing between sea ice and open water an open water detection algorithm based on segmentation and local autocorrelation has been developed. Also ice type classification based on higher-order statistics and independent component analysis has been studied.

A compression algorithm for compressing sea ice SAR data for visual use has been developed. This algorithm is based on the wavelet decomposition, zero-tree structure and arithmetic coding. Also some properties of the human visual system were utilized.

SAR-based ice thickness estimation has been developed and evaluated. This method uses the ice thickness history derived from digitized ice charts, made daily at the Finnish Ice Service, as its input, and updates this chart based on the novel SAR data. The result is an ice thickness chart representing the ice situation at the SAR acquisition time in higher resolution than in the manually made ice thickness charts. For the evaluation a helicopter-borne ice thickness measuring instrument, based on electromagnetic induction and laser altimeter, was used.

# Adaptive combinations of classifiers with application to on-line handwritten character recognition

**Matti Aksela**

*Dissertation for the degree of Doctor of Science in Technology on 29 March, 2007.*

**External examiners:**
David Windridge (University of Surrey)
Jarmo Hurri (University of Helsinki)
**Opponent:**
Robert P.W. Duin (Delft University of Technology)

**Abstract:**

Classifier combining is an effective way of improving classification performance. User adaptation is clearly another valid approach for improving performance in a user-dependent system, and even though adaptation is usually performed on the classifier level, also adaptive committees can be very effective. Adaptive committees have the distinct ability of performing adaptation without detailed knowledge of the classifiers. Adaptation can therefore be used even with classification systems that intrinsically are not suited for adaptation, whether that be due to lack of access to the workings of the classifier or simply a classification scheme not suitable for continuous learning.

This thesis proposes methods for adaptive combination of classifiers in the setting of on-line handwritten character recognition. The focal part of the work introduces adaptive classifier combination schemes, of which the two most prominent ones are the Dynamically Expanding Context (DEC) committee and the Class-Confidence Critic Combining (CCCC) committee. Both have been shown to be capable of successful adaptation to the user in the task of on-line handwritten character recognition. Particularly the highly modular CCCC framework has shown impressive performance also in a doubly-adaptive setting of combining adaptive classifiers by using an adaptive committee.

In support of this main topic of the thesis, some discussion on a methodology for deducing correct character labeling from user actions is presented. Proper labeling is paramount for effective adaptation, and deducing the labels from the user's actions is necessary to perform adaptation transparently to the user. In that way, the user does not need to give explicit feedback on the correctness of the recognition results.

Also, an overview is presented of adaptive classification methods for single-classifier adaptation in handwritten character recognition developed at the Laboratory of Computer and Information Science of the Helsinki University of Technology, CIS-HCR. Classifiers based on the CIS-HCR system have been used in the adaptive committee experiments as both member classifiers and to provide a reference level.

Finally, two distinct approaches for improving the performance of committee classifiers further are discussed. Firstly, methods for committee rejection are presented and evaluated. Secondly, measures of classifier diversity for classifier selection, based on the concept of diversity of errors, are presented and evaluated.

The topic of this thesis hence covers three important aspects of pattern recognition: on-line adaptation, combining classifiers, and a practical evaluation setting of handwritten character recognition. A novel approach combining these three core ideas has been developed and is presented in the introductory text and the included publications.

To reiterate, the main contributions of this thesis are: 1) introduction of novel adaptive committee classification methods, 2) introduction of novel methods for measuring classifier diversity, 3) presentation of some methods for implementing committee rejection, 4) discussion and introduction of a method for effective label deduction from on-line user actions, and as a side-product, 5) an overview of the CIS-HCR adaptive on-line handwritten character recognition system.

## Dimensionality Reduction for Visual Exploration of Similarity Structures

**Jarkko Venna**

*Dissertation for the degree of Doctor of Science in Technology on 8 June, 2007.*

**External examiners:**
Pasi Fränti (University of Joensuu)
Oleg Okun (University of Oulu)
**Opponent:**
Michel Verleysen (Université catholique de Louvain)

**Abstract:**
Visualizations of similarity relationships between data points are commonly used in exploratory data analysis to gain insight on new data sets. Answers are searched for questions like: Does the data consist of separate groups of points? What is the relationship of the previously known interesting data points to other data points? Which points are similar to the points known to be of interest? Visualizations can be used both to amplify the cognition of the analyst and to help in communicating interesting similarity structures found in the data to other people.

One of the main problems faced in information visualization is that while the data is typically very high-dimensional, the display is limited to only two or at most three dimensions. Thus, for visualization, the dimensionality of the data has to be reduced. In general, it is not possible to preserve all pairwise relationships between data points in the dimensionality reduction process. This has lead to the development of a large number of dimensionality reduction methods that focus on preserving different aspects of the data. Most of these methods were not developed to be visualization methods, which makes it hard to assess their suitability for the task of visualizing similarity structures. This problem is made more severe by the lack of suitable quality measures in the information visualization field.

In this thesis a new visualization task, visual neighbor retrieval, is introduced. It formulates information visualization as an information retrieval task. To assess the performance of dimensionality reduction methods in this task two pairs of new quality measures are introduced and the performance of several dimensionality reduction methods are analyzed. Based on the insight gained on the existing methods, three new dimensionality reduction methods (NeRV, fNeRV and LocalMDS) aimed for the visual neighbor retrieval task, are introduced. All three new methods outperform other methods in numerical experiments; they vary in their speed and accuracy.

A new color coding scheme, similarity-based color coding, is introduced in this thesis for visualization of similarity structures, and the applicability of the new methods in the task of creating graph layouts is studied. Finally, new approaches to visually studying the results and convergence of Markov Chain Monte Carlo methods are introduced.

## Language Models for Automatic Speech Recognition: Construction and Complexity Control

**Vesa Siivola**

*Dissertation for the degree of Doctor of Science in Technology on 3 September, 2007.*

**External examiners:**
Krister Lindén (University of Helsinki)
Imre Kiss (Nokia Research Center)
**Opponent:**
Dietrich Klakow (Universität des Saarlandes)

**Abstract:**
The language model is one of the key components of a large vocabulary continuous speech recognition system. Huge text corpora can be used for training the language models. In this thesis, methods for extracting the essential information from the training data and expressing the information as a compact model are studied.

The thesis is divided in three main parts. In the first part, the issue of choosing the best base modeling unit for the prevalent language modeling method, n-gram language modeling, is examined. The experiments are focused on morpheme-like subword units, although syllables are also tried. Rule-based grammatical methods and unsupervised statistical methods for finding morphemes are compared with the baseline word model. The Finnish cross-entropy and speech recognition experiments show that significantly more efficient models can be created using automatically induced morpheme-like subword units as the basis of the language model.

In the second part, methods for choosing the n-grams that have explicit probability estimates in the n-gram model are studied. Two new methods specialized on selecting the n-grams for Kneser-Ney smoothed n-gram models are presented, one for pruning and one for growing the model. The methods are compared with entropy-based pruning and Kneser pruning. Experiments on Finnish and English text corpora show that the proposed pruning method gives considerable improvements over the previous pruning algorithms for Kneser-Ney smoothed models and also is better than entropy pruned Good-Turing smoothed model. Using the growing algorithm for creating a starting point for the pruning algorithm further improves the results. The improvements in Finnish speech recognition over the other Kneser-Ney smoothed models were significant as well.

To extract more information from the training corpus, words should not be treated as independent tokens. The syntactic and semantic similarities of the words should be taken into account in the language model. The last part of this thesis explores, how these similarities can be modeled by mapping the words into continuous space representations. A language model formulated in the state-space modeling framework is presented. Theoretically, the state-space language model has several desirable properties. The state dimension should determine, how much the model is forced to generalize. The need to learn long-term dependencies should be automatically balanced with the need to remember the short-term dependencies in detail. The experiments show that training a model that fulfills all the theoretical promises is hard: the training algorithm has high computational complexity and it mainly finds local minima. These problems still need further research.

## Advances in variable selection and visualization methods for analysis of multivariate data

**Timo Similä**

*Dissertation for the degree of Doctor of Science in Technology on 19 October 2007.*

**External examiners:**
Risto Ritala (Tampere University of Technology)
Patrik Hoyer (University of Helsinki)
**Opponent:**
Volker Tresp (Siemens Corporate Technology)

**Abstract:**
This thesis concerns the analysis of multivariate data. The amount of data that is obtained from various sources and stored in digital media is growing at an exponential rate. The data sets tend to be too large in terms of the number of variables and the number of observations to be analyzed by hand. In order to facilitate the task, the data set must be summarized somehow. This work introduces machine learning methods that are capable of finding interesting patterns automatically from the data. The findings can be further used in decision making and prediction. The results of this thesis can be divided into three groups.

The first group of results is related to the problem of selecting a subset of input variables in order to build an accurate predictive model for several response variables simultaneously. Variable selection is a difficult combinatorial problem in essence, but the relaxations examined in this work transform it into a more tractable optimization problem of continuous-valued parameters. The main contribution here is extending several methods that are originally designed for a single response variable to be applicable with multiple response variables as well. Examples of such methods include the well known lasso estimate and the least angle regression algorithm.

The second group of results concerns unsupervised variable selection, where all variables are treated equally without making any difference between responses and inputs. The task is to detect the variables that contain, in some sense, as much information as possible. A related problem that is also examined is combining the two major categories of dimensionality reduction: variable selection and subspace projection. Simple modifications of the multiresponse regression techniques developed in this thesis offer a fresh approach to these unsupervised learning tasks. This is another contribution of the thesis.

The third group of results concerns extensions and applications of the self-organizing map (SOM). The SOM is a prominent tool in the initial exploratory phase of multivariate analysis. It provides a clustering and a visual low-dimensional representation of a set of high-dimensional observations. Firstly, an extension of the SOM algorithm is proposed in this thesis, which is applicable to strongly curvilinear but intrinsically low-dimensional data structures. Secondly, an application of the SOM is proposed to interpret nonlinear quantile regression models. Thirdly, a SOM-based method is introduced for analyzing the dependency of one multivariate data set on another.

## Methods for exploring genomic data sets: application to human endogenous retroviruses

**Merja Oja**

*Dissertation for the degree of Doctor of Science in Technology on 14 December 2007.*

**External examiners:**
Juho Rousu (University of Helsinki)
Tero Aittokallio (University of Turku)
**Opponent:**
Hiroshi Mamitsuka (Kyoto University)

**Abstract:**

In this thesis exploratory data analysis methods have been developed for analyzing genomic data, in particular human endogenous retrovirus (HERV) sequences and gene expression data. HERVs are remains of ancient retrovirus infections and now reside within the human genome. Little is known about their functions. However, HERVs have been implicated in some diseases. This thesis provides methods for analyzing the properties and expression patterns of HERVs.

Nowadays the genomic data sets are so large that sophisticated data analysis methods are needed in order to uncover interesting structures in the data. The purpose of exploratory methods is to help in generating hypotheses about the properties of the data. For example, by grouping together genes behaving similarly, and hence presumably having similar function, a new function can be suggested for previously uncharacterized genes. The hypotheses generated by exploratory data analysis can be verified later in more detailed studies. In contrast, a detailed analysis of all the genes of an organism would be too time consuming and expensive.

In this thesis self-organizing map (SOM) based exploratory data analysis approaches for visualization and grouping of gene expression profiles and HERV sequences are presented. The SOM-based analysis is complemented with estimates on reliability of the SOM visualization display. New measures are developed for estimating the relative reliability of different parts of the visualization. Furthermore, methods for assessing the reliability of groups of samples manually extracted from a visualization display are introduced.

Finally, a new computational method is developed for a specific problem in HERV biology. Activities of individual HERV sequences are estimated from a database of expressed sequence tags using a hidden Markov mixture model. The model is used to analyze the activity patterns of HERVs.