The Role of Soft Computing in Intelligent Data Analysis

(Invited Paper)

Rudolf Kruse, Christian Borgelt, Detlef D. Nauck, Nees Jan van Eck and Matthias Steinbrecher

***Abstract*—The analysis of large amounts of data becomes an ever more important issue in science and business. Users, typcially domain experts, are faced with a huge challenge and require support. In this paper we look how data analysis can be intelligently supported and how soft computing methods can help. As examples, we look at some visualization concepts, approximate matching in frequent pattern mining as a technical example, and some industrial applications concerned with complex analytical scenarios and aspects of automating analytics.**

I. Introduction

Technological advances in computing, data storage, networks and sensors have dramatically increased our ability to access, store and process huge amounts of data. In scientific research and ever increasing in business applications we see ourselves confronted with the need to extract relevant information from huge amounts of data and heterogeneous data sources, like sensors, databases, text archives, images, audio and video streams etc.

Semi-automated analytics supported by interactive visual methods play an important role in facilitating real-time decision making by human users. In this paper we analyze the role soft computing in general and fuzzy systems in particular play in this area.

Computer-supported visual analytics pose several challenges to software design and human-computer interaction. In fact, this is a new research area that already has resulted in the creation of several new research groups at prime institutions worldwide. A major prerequisite for visual analytics is information integration and fusion. Complex scenarios have to integrate several heterogeneous information sources. Dedicated analysis tools have to condense and summarize the information to reach an abstraction level adequate for the decision support. Extracted information has to be presented in an intuitive way allowing direct manipulation and switching between different levels of abstraction.

Moreover, in any reasonably complex analytical scenario context can change at any time. Changes can occur on the level of data sources, user intentions, the overall objective of the analysis or the representation and visualization of results including output devices. This means the whole analytical process has to be adaptive to cope with such possibly unforeseen changes. Given this level of complexity and the fact that human users are typically not experts in analytics a certain level of automation is required.

In the following we look at some challenges in analytics and how soft computing can contribute. We start at looking at selected aspects of visualisation. Then we show how fuzzy approaches can enhance an established analytical methodology that plays an ever increasing role in applications — frequent pattern mining. Finally, we look at a selection of industrial applications and list some trends and opportunities for fuzzy methods in analytical applications.

II. Visualization

In this section we look at two visualization techniques. Concept maps provide an intuitive overview on the prevalence of and associations between concepts extracted from, for example, document collections. The second technique is used to support the visual identification of conspicious data subsets given a Bayesian Network.

A. Concept Maps

A concept map represents the associations between concepts by arranging them in a graphical representation such that closeness reflects strengths of associations. This approach is useful to visualize the content of document collections, for example. Important business applications are the analysis of customer communications (user fora, e-mails, call center dialogs etc) or news items, for example.

As an illustrative example we analyze the way in which the Fuzzy Systems (FS) field is divided into several subfields. The visualizations provide insight into the characteristics of each subfield and their relations. By comparing two visualizations, one based on data from 2006 and one based on present data from 2007, we examine how the FS field has evolved in the last year. The data we use consist of the abstracts of the papers presented or accepted at the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) in 2006 and 2007. Using a fully automatic procedure, concept maps are constructed from the data. These maps visualize the associations between the main concepts in the FS field. Our analysis of the structure and the evolution of the FS field is largely based on the constructed concept maps.

We used the following procedure to construct concept maps of the FS field. First, for each of the abstracts of the papers presented at the FUZZ-IEEE 2006 and the FUZZ-IEEE 2007,

Rudolf Kruse and Matthias Steinbrecher are with the Faculty of Computer Science, Otto-von-Guericke University, Universitätsplatz 2, 39106 Magdeburg, Germany (email: (kruse|msteinbr)@iws.cs.uni-magdeburg.de).

Christian Borgelt is with the European Center for Soft Computing, Edificio Científico-Tecnológico, c/ Gonzalo Gutiérrez Quirós s/n, 33600 Mieres, Asturias, Spain (email: christian.borgelt@softcomputing.es).

Detlef D. Nauck is with the Intelligent Systems Research Centre, Research & Venturing, BT Group, Orion pp 1/12, BT Adastral Park, Ipswich, IP5 3RE UK (email: detlef.nauck@bt.com).

Nees Jan van Eck is with the Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands (email: nvaneck@few.eur.nl).

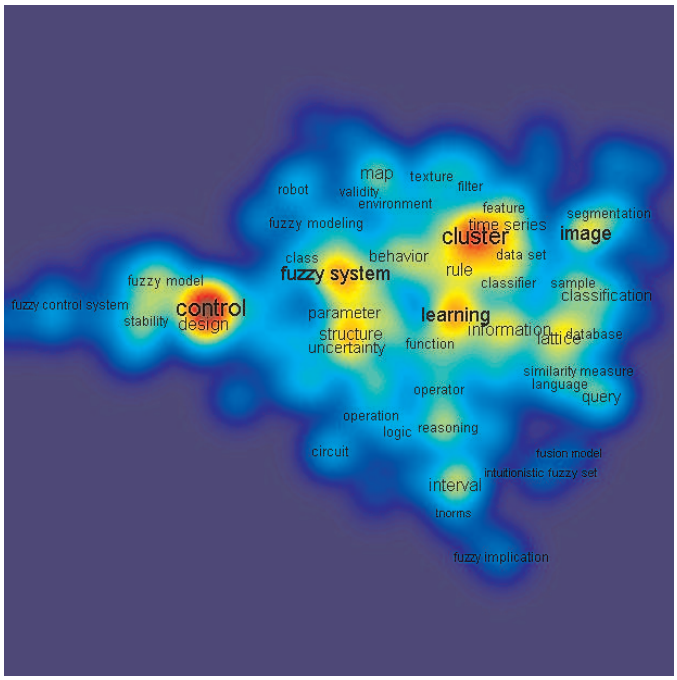


Fig. 1. Concept density map of the fuzzy systems research field in 2006.

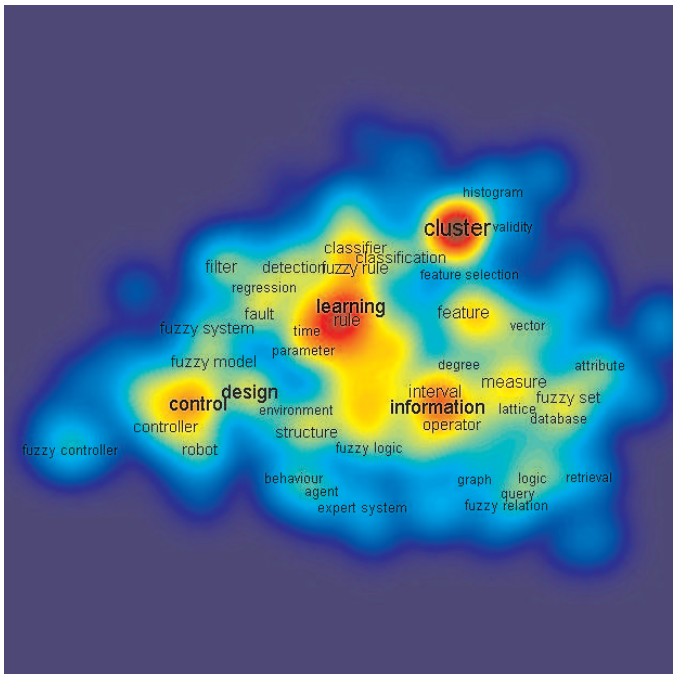


Fig. 2. Concept density map of the fuzzy systems research field in 2007.

the concepts occurring in the abstract were identified. This was done using a simple thesaurus of the FS field that we had constructed ourselves (for more details, see [11]). In the next step, co-occurrences of concepts were counted. Based on concepts' co-occurrence frequencies, the associations between pairs of concepts were calculated. The association between two concepts was quantified using a measure called association strength. The construction of concept maps was accomplished using a new method called VOS (*visualization of similarities* [12]), that provides a two-dimensional space in which the concepts are located in such a way that the distance between any pair of concepts reflects their association strength as accurately as possible.

To display a concept map, we use so-called concept density maps where colors are used to indicate the density of concepts. Dark blue indicates the lowest density and dark red indicates the highest density. Concept density maps are especially useful to get a quick overview of the various clusters of concepts within a concept map. The approach that we took to calculate concept densities is discussed in [13].

The concept density maps for 2006 and 2007 can be seen in Figure 1 and Figure 2, respectively. The map for 2006 contains 207 concepts while the map for 2007 contains 273 concepts. It can be seen that the concepts in the fuzzy systems field are grouped into a number of clusters in the map for 2006 and the map for 2007. In both maps, two clusters of concepts that received quite a lot of attention correspond to two important topics in the fuzzy systems field, namely *fuzzy control* and *fuzzy clustering and classification*. Examples of other clusters of concepts that can be found in the maps correspond to the topics of *fuzzy sets* and *fuzzy information processing*. By comparing both maps, it turns out that the structure of the fuzzy systems field has been fairly stable during the last two years. However, it seems that the focus has shifted from the topic of fuzzy control in 2006 to the topic of fuzzy clustering and classification in 2007. This shift in focus is supported by the fact that approximately 80 out of the 250 reviewers of the FUZZ-IEEE 2007 have listed the topic "Fuzzy data analysis—clustering, classifiers, pattern recognition, bio-informatics" as an area of expertise. The next most popular topics are only listed by approximately 40 reviewers. So, there is evidence to suggest it is the most active subarea of the fuzzy systems field at the moment.

B. Visualising Bayesian Networks

Bayesian networks have proven to be a well-suited technique [5], [22], [43], [44], [51] for sophisticated data analysis. The graphical structure of a Bayesian network encodes the structure of a set of conditional probability distributions. To provide the user with an intuitive visualization method, the probabilistic structure can be interpreted as a collection of association rules [2] for which we present an intuitive visualization method (for details, see [49], [50]).

Consider that we are dealing with the task of data analysis in a business intelligence workflow. The aim is to infer concept descriptions that help a domain expert identifying suspicious subsets in the modelled data. Considering the setting of some automobile manufacturer, the domain expert could be given a network structure as shown in Figure 3. This network provides a first insight into the corresponding application domain. For example, one can infer that temperature and the road surface conditions have some (at least statistical) impact on the type of failure.

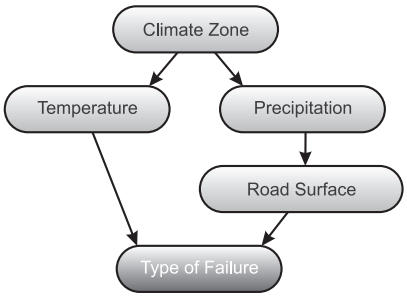


Fig. 3. A simplified Bayesian network as it may result from the analysis of data collected by an automobile manufacturer.

Unfortunately, there is no pictorial information available *which* road conditions have *what kind* of impact on *which* type of failure. However, this information can be easily retrieved in form of conditional probabilities from the underlying data set, given the network structure. This becomes clear if the sentence above is rephrased: "Given a specific road surface condition, what is the failure probability of a randomly selected vehicle?"

The answer to such a question can be given as an association rule. Its antecedent attributes induce a partition of the entire data set into subsets of items sharing the same values for the antecedent attributes. Each of these item sets is then again subdivided according to the different class attribute values. The resulting item sets (each representing an association rule) can be plotted in a chart as follows: (1) represent each item set as a circle, the size being proportional to the number of its items. (2) calculate appropriate x- and y-coordinates as values of association rule evaluation measures chosen in advance. (3) colorize the circles with a distinct color for each class attribute value.

If we choose recall and lift [55] as rule evaluation measures for the x- and y-axis, respectively, then the result could be a chart as shown in Figure 4. A user would use the following heuristics to interpret this chart: "Large circles in the upper right-hand side corner are good candidates for a closer inspection."

The data set under consideration for Figure 4 contained approximately 300,000 cars that exposed a many-valued class variable, hence the different colors and shades of the circles in Figure 4. Although there was no explanation for the sets 2, the subset 1 represented roughly 900 cars with increased failure rates which could be explained by the respective values of the attributes Mileage and RoadSurface and therefore helped to assess the problem.

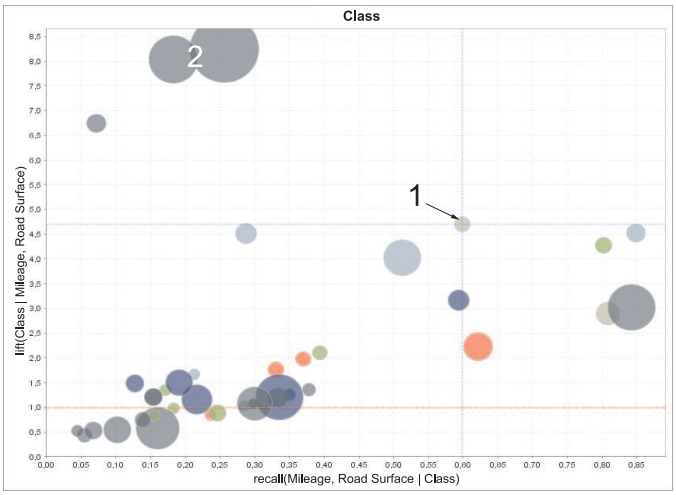


Fig. 4. Although it was not possible to find a reasonable description of the vehicles contained in subsets 2, the attribute values specifying subset 1 were identified to have a causal impact on the class variable.

III. Approximate Matching in Frequent Pattern Mining

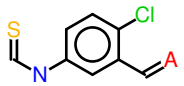
Frequent pattern mining consists in finding all patterns of a given type (most common are sets, sequences, trees and graphs) that have a user-specified minimum support in a (structured) data set. For example, in classical frequent item set mining, which was originally developed for market basket analysis, one tries to find all sets of items (e.g. products, service options, special equipment items) that appear in a given minimum number of transactions. In frequent subgraph mining—which has applications in biochemistry, web mining, and program flow analysis—a database of (attributed) graphs (e.g. a set of molecules or data flow graphs of procedures) is given and one tries to find all subgraphs that occur in a user-specified minimum number of these graphs.

Due to the rapid development of frequent pattern mining, the basic, mathematical problem can be considered solved. For all standard types of problems efficient algorithms have been proposed and implemented. An overview for frequent item set mining can be found in [20], where implementations of improved and extended versions of the best known algorithms Apriori [3], Eclat [56], FP-growth [21], but also several other approaches where compared against each other. For the most general task of frequent graph mining, efficient algorithms have been developed by transferring the core ideas to graph structures. Examples of such algorithms include MolFea [27], FSG [28], MoFa [6], gSpan [53], CloseGraph [54], FFSM [24], and Gaston [41]. The core ingredient of these methods is a canonical form of a graph, which is used to avoid redundant search (see [7] for the basic idea and a unified view).

However, that the mathematical problem can be considered solved and that there are efficient implementations does not mean that there is no work left to be done. The problem is that in applications special objectives and domain-specific background knowledge have to be taken into account to make the search efficient and the output meaningful.

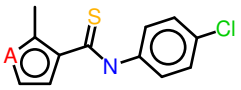
A. Approximate Matching

One of the issues arising in applications is that the mathematical problem is based on exact matching. That is, the frequent pattern has to appear exactly in the elements of the database (transactions, sequences, graphs). However, for many applications this is too strict a requirement. For example, frequent item set mining can be used to find patterns in alarm sequences in telecommunication networks [30] using



	A=O	A=N
CA:	5.5%	3.7%
CI/CM:	0.0%	0.0%

Fig. 5. First example of a wildcard fragment from the NCI HIV data set.



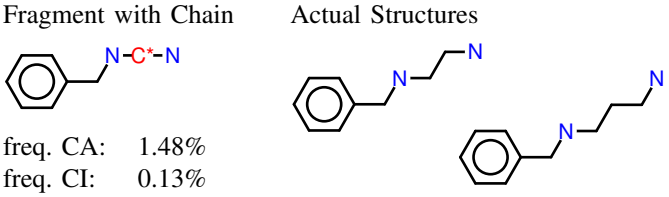
	A=O	A=S
CA:	5.5%	0.01%
CI/CM:	0.0%	0.0%

Fig. 6. Second example of a wildcard fragment from the NCI HIV data set.

a time windowing technique, by which the continuous alarm sequence is split into pseudo-transactions. Unfortunately, such alarms often get delayed, lost, or repeated due to noise, transmission errors, failing links etc. If alarms do not get through or are delayed, they can be missing from the transaction (time window) its associated items (alarms) occur in. If one uses exact matching in this case, the support of some item sets, which could be frequent if the items did not get lost, may be smaller than the user-specified minimum. This leads to a possible loss of potentially interesting frequent item sets. Therefore specific algorithms that can find approximate, fault-tolerant, "fuzzy" item sets are needed. Examples of such algorithms include [10], [45], and [52], which already provide first feasible solutions, but more research is needed to improve their efficiency and practical usefulness.

Another example is molecular fragment mining, where a database of chemical compounds (represented as attributed graphs) is mined for frequent or (if activity information is present) discriminative fragments, which can give hints about potential pharmacophores. Here exact matching is not always appropriate, because what a (bio)chemist is mainly interested in is the chemical behavior of a compound, not necessarily its exact composition, as slightly different structures can behave in a very similar way. For example, a benzene ring with only carbon atoms and a ring in which one carbon is been replaced by nitrogen often exhibit (almost) the same properties. Hence it is desirable to be able to mine fragments with "wildcard" atoms [23]. Examples are the two molecular fragments shown in Figures 5 and 6 (the letter A denotes the wildcard atom) that can be found in the National Cancer Institute's HIV antiviral screening data set (NCI HIV data set, CA: confirmed active, CI/CM: only moderately active or inactive). Especially the second fragment demonstrates the advantages of such an approach: the fragment in which A is a sulphur atom would never be found without wildcard matching, since its support as an exact pattern is too low. However, it may be just this variant that possesses the best properties for drug development.

Another example of approximate matching are chains of carbon atoms. Often a chemist is only interested in the presence of a carbon chain, but not its exact length, since a chain, regardless of its length, allows for a rotation of the molecular parts that are connected by the chain. An example fragment from [35], which can be found in the National Cancer Institute's Cancer data set, is shown in Figure 7: the fragment shown on the left matches both structures on the right.



freq. CA: 1.48%
freq. CI: 0.13%

Fig. 7. An example of a fuzzy carbon chain from the NCI cancer data set.

B. Incorporating Background Knowledge

Another problem in frequent pattern mining is the size of the output. For low minimum support values the number of found frequent patterns often enough exceeds the number of elements in the database to analyse. In order to tackle this problem, it can be useful to restrict the output to patterns with certain meaningful properties, thus filtering out uninteresting patterns. For example, if the application area is molecular fragment mining, then treating rings (e.g. benzene rings) as units—that is, requiring that a ring is either present as a whole or not at all—not only reduces the size of the output, but also improves its interpretability and speeds up the mining process considerably [8], [23]. The loss, if there is any, is negligible, since chemists usually see rings as chemical units and thus are often not interested in fragments containing only a few bonds of a ring (of which it may not even be clear whether they are always part of a ring or in some molecules part of a branch).

In the same way, chemical background knowledge is needed in order to use the technique of wildcard atoms (see above) effectively. Different elements do not always behave in a similar way. Therefore one needs chemical background knowledge what elements in what contexts (e.g. in rings, when certain neighbors are present/absent) can be considered similar. This background knowledge then has to be used in the search to make it efficient and its output meaningful for a user.

Generally, filtering out the uninteresting patterns in order to relieve a user from scanning a large number of potentially interesting patterns manually is one of the greatest challenges of frequent pattern mining. Such filtering can hardly be achieved by relying on domain-independent measures (like statistical indicators), because, as is well known, not everything that is statistically significant is interesting or even relevant. For example, mining hospital data will reveal that 100% of all pregnant patients are female and 0% male—a statistically extremely significant, but entirely obvious and thus irrelevant result.

IV. Industrial Applications

The success of fuzzy systems in industrial applications is mainly due to fuzzy control applications. After early success stories like the Danish cement kiln (1982) and the Sendai Subway (1986) we have grown used to fuzzy features in appliances and consumer electronics like washing machines and digital cameras, for instance.

The application of fuzzy systems in data analysis is frequently reported in research papers, however, fuzzy technology is noticeably absent from the portfolio of all large data mining software vendors. In [40] Nauck argues that this is due to a lack of research software that can function as a reference implementation.

Instead of fuzzy systems, data mining software has focussed on neural networks and decision trees — two methods that are pervasive in modern data analysis suites provided by big names like, for example, SAS, IBM or SPSS. Oracle Data Mining provides support vector machines instead considering them to be a superset of neural networks [42]. Neural networks also enjoyed success in large scale applications like the credit card fraud detection system Falcon [14].

Bayesian networks (probabilistic graphical models) have been around for a long time in form of general purpose development environments like Hugin or Netica. Recently interest in Bayesian networks has grown in the area of operational risk management [47] leading even to the appearance of specialized companies [1]. BT reports an application of automatically generated Bayesian Networks for the analyis of customer surveys [15], [39]. Volkswagen is using a related graphical model (Markov networks) for item planning [17].

When we look at reported industrial analytical applications we see that rule-learners based on machine learning or AI approaches (e.g. decision trees), neural networks and Bayesian networks are well established in in that space because vendors of industry-standard software platforms have decided to include them into their portfolios. The question now is: where is the niche for fuzzy systems?

In the following we look at a few examples where fuzzy systems and some other soft-computing related methods were used in a data analysis context and then summarize the trends and possibilities arising from these approaches. We begin with a short description of the graphical model for item planning used a Volkswagen. This is a very complex application and illustrates the need for more sophisticated methods than simply relating some inputs directly to some outputs as we see in most fuzzy data analysis scenarios, like, fuzzy classifiers, for example. After that we look a explanation generation, the automation of data analysis and pro-active knowledge discovery.

A. Graphical Model for Industrial Planning

Complex products like automobiles are usually assembled from a number of prefabricated modules and parts. Many of these components are produced in specialized facilities not necessarily located at the final assembly site. An on-time delivery failure of only one of these components can severely lower production efficiency. In order to efficiently plan the logistical processes, it is essential to give acceptable parts demand estimations at an early stage of planning.

At the Volkswagen Group an item planning system is used to support parts demand planning and calculation as well as capacity management across all vehicle trademarks. Since the parts from which a car is assembled can only be combined w.r.t. numerous technical and marketing constraints, it is crucial to assess these constraints when new evidence (such as a shortage of a component or an increased number of orders of a specific type of car) arises.

The implementation is based on graphical models ([5], [29], [46]) — more specifically on Markov Networks — that decompose the enormous entirety of possible car configurations into subspaces that can be tackled more efficiently. The decomposition is driven by the exploitation of independence constraints which are extracted from an underlying relation that is defined by a rule base representing the expert knowledge of the vehicle production from different points of view: There may be technical constraints stating for example, which transmission types can be combined with which engine models. Even though some car configurations are technically possible, the sales and marketing department may want to exclude them due to the lack of an appropriate customer target group. Therefore, these rules not only prune the entire number of car combinations but also impress an structure on the remaining "allowed" combinations that is exploited when inducing the graphical decomposition model.

Since the main planning task consists in integrating new or modified knowledge, one can distinguish the following tasks w.r.t. the affected model component.

1) Qualitative Model Change: The above-mentioned qualitative component of the model is derived from the given rule base. Thus, a modification of the rules call for an *update* of the underlying decomposition structure [17]. Update operations allow experts to refine rules to better fit the technical constraints or extend the number of configurations when new parts become available. From the marketing point of view, it is possible to react on market changes if, for example, one offers the customer a wider variety of customization options to choose from.

2) Quantitative Model Change: A quantitative model change refers to a reassessment of the numerical parameters [18], i.e. the relative frequencies of car combinations. It is a necessary consequent of a qualitative model change, since the newly introduced configurations require proper initial probabilities. The main task, however, is to *revise* the numerical parameters in order to adapt the model to new technical or marketing-related requirements. When a shortage of one or more parts occurs, the revision process serves as a method of answering the question, what impact there will be on the production pipeline. Another scenario would be an increased ordering of a special car trademark. To be able to deliver the requested number of cars, it is crucial to know which subparts of the production line have to adapt to the raised demand. While the revision process takes as input entire user-modified marginal distributions one can consider a specialization of this process: initializing one or more attributes with a fixed value (i.e., changing the respective distributions to deterministic ones that assign the entire probability mass to a single value). This method is called *focusing* and coincides with the well-known concept of evidence propagation ([9], [25]). It is used to evaluate fictive settings and answer what-if-questions to

experts of both technical and marketing domains.

The described system was launched back in 2001 by the Corporate IT, Sales and Logistics departments of the Volkswagen Group and is rolled out to all trademarks since 2004. While the system design is maintained by Volkswagen itself, the entire model change and prediction engine is developed solely by ISC Gebhardt [19]. The chosen model representation by Markov networks, the reduction of calculation time and the corresponding update and revision operations turned out to be essential prerequisites for advanced parts demand planning in the presence of huge numbers of possible configurations under numerous technical constraints.

B. Explanation Generation

In [36] Nauck describes Intelligent Travel Estimation and Management System ITEMS, a web-based software system that predicts, manages, visualizes and explains travel patterns of a mobile workforce. ITEMS is a tool for service industries like telecommunications, gas, water, electricity etc. that have to schedule jobs for large mobile workforces. Successful scheduling requires suitable estimates of inter-job times that are mainly determined by travel time for which routing information is typically not available. ITEMS has been developed and used by BT to estimate travel times of its engineers. ITEMS has a learning component that constantly builds new (crisp) regression models for travel time prediction.

In addition ITEMS contains an explanation facility based on decision trees and neuro-fuzzy systems (NEFCLASS [37]) that display rule-based information about individual journeys. The rules derived from travel data explain, for example, why a certain journey may have been late. The information of those rules can be used by managers to improve the overall system behavior. For example, if an automatically generated rule would reveal that travel between two specific areas takes usually longer than predicted at a certain time of day the scheduler can be advised to avoid scheduling journeys between those two areas at that time of day.

The purpose of the explanatory rules is to provide resource managers with a tool to investigate workforce travel patterns. The rules provide a summary of the actual data and highlight influential variables. In order to be useful the rules must be simple and sparse. It must also be possible to create the rule base on the fly in a very short time. The user would select a specific set of journeys for which he requires explanations. The system must then create the rules completely automatically without user interaction.

The user can decide if he prefers crisp rules from a decision tree or fuzzy rules generated by NEFCLASS. In [36] an interpretability index is introduced and it is shown that the fuzzy rules typically score higher on this index and should therefore be easier to interpret. Incidentally, the accuracy of the fuzzy rules are also slightly higher than the decision tree. Furthermore, it was important for this application to be able to use numerical and symbolic variables in the same rule.

C. Automation of Data Analysis

Analysts in businesses are typically domain experts and rarely highly trained data analysis experts. Such business users have difficulties using specialized data mining software that is typically aimed at specialists. For this reason BT's research organization ran a project on the automation of data analysis in order to find ways of empowering business users who have to analyse data. The result of this research program is the SPIDA platform (Soft Computing Tool for Intelligent Data Analysis) [38], [48]. SPIDA makes state-of-the-art data analysis techniques available to non-experts by employing a wizard that selects appropriate data analysis methods given soft high-level requirements. The wizard also configures and runs the chosen methods automatically.

SPIDA uses soft constraints for the selection of an appropriate data analysis method. These constraints represent the user's requirements regarding the analysis problem in terms of the actual problem (like prediction, clustering or finding dependencies) and preferences regarding the solution.

Requirements can potentially be defined at any level of abstraction. The wizard uses expert knowledge in terms of a fuzzy rule base to map high-level requirements onto required properties of data analysis methods which will then be matched to actual properties of analysis methods. In order to compute the degree of match a new measure for the compatibility of fuzzy requirements with fuzzy properties was introduced [48] that can also be applied to other problems in the area of multi-criteria decision making.

Although SPIDA offers fuzzy methods for data analysis (in addition to neural networks, decision trees, support vector machines etc.) the main use of fuzzy technology appears behind the scenes to control and automate the data analysis process. The main advantage of using fuzzy methods in this context is the ease of implementation and the advantage of using soft constraints which appear less complex to users.

Another example for an automated analytical application based on soft computing has also been developed at BT [15], [39]. The software iCSat is used by BT to analyse customer survey data and to identify drivers of customer satisfaction. iCSat automatically learns the structure and probabilities of a Bayesian network from data and represents it as a series of simple bar charts the user can manipulate.

The business analysts using iCSat are used to working with visualizations like bar charts to represent distributions of key variables. iCSat allows users to modify bar charts by simple mouse operations to specify inputs to a Bayesian network. In contrast to standard Bayesian tools the user can also enter soft evidence instead of simply selecting only one value for a variable. The result of the propagation process is visualized in a second set of bar charts with two sets of bars in each chart. This allows the user to instantly compare the distributions in the data against predictions by the Bayesian network given the inputs the user has provided.

The fact that the software uses Bayesian networks is completely hidden from the user. The graphical representation of a network is deliberately not shown. Because networks are

generated automatically from data the directions of edges in the networks are determined by heuristics. Users would be prone to misinterpret the directions of edges in the network and confuse probabilistic correlations with causal influences.

D. Pro-active Knowledge Discovery

For businesses it is a strategic issue to discover new information and using it before others do. This requires the ability to detect, assess and respond to new trends and events rapidly and intelligently. The main goal of data mining is to find information in data that is interesting, novel and potentially useful [16]. This definition suggests some element of "surprise" related to the results of a data mining process. Otherwise the results were to be expected and therefore not really novel and possibly less interesting. However, data mining today is typically a goal-oriented assumption-driven activity. Data mining software allows users to build predictive models which in turn may lead to interesting and useful predictions. But data mining also intrinsically assumes that the domain under consideration is constant.

However, this assumption is typically not correct. Businesses, customer behavior, market development, etc. is in a constant flux. The churn prediction model that I have built today may be obsolete tomorrow because of changing consumer attitudes or market conditions. From this perspective the question: "Which patterns exist?" as it is answered by state-of-the-art data mining technology, is replaced by the question: "How do patterns change?" as described in [4] where a framework for rule change mining is presented.

Boettcher et al. [4] describe an approach for association rule mining on time-stamped data that looks at how association rules change over time. They have developed measures for describing features of trends for rule support and confidence. By measuring steepness, recentness, reliability and inconsistency of rule trends the authors can define an interestingness measure on the discovered rules.

The framework can continuously and automatically mine time-stamped data. Patterns (association rules) are evaluated against their interestingness and displayed in a ranked list. The framework uses a fuzzy rule base to combine the four trend measure into a single interestingness score. This approach also allows to customize the interestingness measure to certain domains where, for example, steep trends may be expected, but recent trend changes are more interesting.

Like in the previous subsection fuzzy methods are here used behind the scenes of an analytical platform to improve its usability. Instead of browsing patterns ordered by four abstract trend measures an intuitive single ranking criterion is computed by a fuzzy rule base.

This framework for rule change mining represents an important trend for analytical applications. Data analysis should become automatic and pro-active and report only interesting novel findings. For this to work we need ways of assessing interestingness which is an intrinsically subjective and fuzzy concept. Fuzzy methods can play an important role in this context by applying them as an enabling technology behind the scenes of analytical applications.

V. Trends and Opportunities

Soft computing methods in data analysis have very much focused on structured data and have competed with statistical and AI (machine learning) approaches. Especially fuzzy approaches for data analysis are typically compared against "classical" solutions, claim a more or less minute performance improvement and further claim advantages in terms of being interpretable in form of intuitive fuzzy rules. However, from our experience in various industrial applications we learned that business users are typically not interested in interpretability aspects or derived fuzzy rules. Aspects of accuracy, speed, maintainability, and especially integration into operational systems are much more important.

Emphasising the intuitiveness and interpretability of fuzzy rule-based data analysis approaches is relevant to some extent for convincing experienced analysts who may gain some convenience compared to using other approaches. This aspect of fuzzy data analysis is very similar to the advantages of fuzzy control, which mainly appeals to control engineers (and not to end users), because of the opportunity of quickly, cheaply and intuitively designing a control application.

In order to make an impact on industrial data analysis soft computing and fuzzy methods in particular should focus on areas where they can play out their strengths. In the remainder of this section we will briefly look at some areas where we think fuzzy methods can make an impact.

Cluster Analysis: In cluster analysis fuzzy methods have an actual analytical advantage in delaying the assignment of an element to a cluster from the time of cluster formation until the time of decision making. Fuzzy cluster analysis therefore will continue to play an important role in analysing structured data. For businesses clustering time-stamped data comprising numerical and non-numerical features is important for detecting, for example, changing or emerging customer behavior. To support this demand, methods for automatically finding the correct number of clusters and methods for treating non-numerical data are required.

Information fusion, question answering, search, and text mining: Fuzzy methods need to focus more on unstructured and semi-structured data, i.e. simple text and text with mark-up (like XML, etc.). We are already beginning to see fuzzy methods that allow the combination of different information hierarchies [31] or fuzzy grammars [32] that are applicable to the classification of text. These methods are important for managing the ever-faster growing body of documents, e-mails, images, music, videos etc. that we insist on storing on our hard drives, in business databases, or on the Internet ("digital obesity" [33]). Businesses need these methods to analyse customer communications and to web-enable consumer-oriented business processes. The future search engine that will replace Google will have to have the ability of identifying documents that actually answer a particular question instead of simply

containing a list of keywords. Fuzzy grammars and soft information fusion are likely to play an important role in this.

Inexact search and approximate matching approach as mentioned in Section III are also important areas where fuzzy techniques are indispensible. Fuzzy databases are a well-studied example, but fuzzy matching approaches are required to extend typical search strategies as we find them in search engines, rule mining, knowledge discovery etc.

Explanation and summarization: An important aspect of data analysis in business environment is to represent the results in an appropriate way. This can be done by intuitive visualizations or by textual explanations that can be understood by domain experts. Lack of these technologies are one of the causes of business users applying typically only very simple analytical tools. For example, consider trend analysis. A domain expert may be interested in monitoring attributes in his data to find out which display significant trends. A typical naive approach is to fit linear regressions and to display attributes whose slope exceed a given threshold. In reality, what the user is probably looking for are trends that are very steep, have changed direction recently, display a growing or diminishing slope etc. A more sophisticated trend analyser that generates suitable linguistic summaries would be much more useful. An approach for generating linguistic summaries for trends has been described by Kacprzyk et al. [26]. By turning the approach around linguistic summaries can act as a search language and help identifying or filtering out analysis results. Generation of explanations or summaries is relevant in many areas: what type of clusters have been detected, which attributes particularly influence a prediction, what is the quality of the data etc.

Behind the scenes: Successful projects carried out at BT's research center have shown that fuzzy methods can be very useful in automating and supporting data analysis procedures. Fuzzy methods have been used to select, configure and execute data analysis algorithms [38] or to simplify the display of analysis results [4]. The ever-growing demand for data analysis in areas like business intelligence and customer relationship management result in the necessity to make analytics available on all levels of an enterprise. That requires integration and automation of algorithms and simplification of user interfaces including visualization of results. Fuzzy methods can play an important role in this area.

Knowledge Fusion: Most analytical projects will not be successful without the application of domain knowledge which is often incomplete and vague. Domain knowledge guides the analysis process and filters out interesting results from uninteresting ones. Analysis results themselves add to the body of knowledge which must be appropriately updated or revised. Analysis results can be inconclusive, incomplete, unreliable etc. and these phenomena need to be taken into account. It is a business reality that data cannot always be aggregated into a data warehouse before it can be analysed. Data stream mining and distributed data mining will become more important which means methods for fusing analysis results will be required. Both fuzzy and probabilistic methods are required for this type of analytical applications. We already see implementations emerging, for example, in the third generation telecare systems that use distributed sensor networks and fuzzy data analysis combined with domain knowledge to monitor the long-term well-being of individuals requiring care [34].

VI. Conclusions

Soft computing methods are valuable assets in intelligent data analysis. Neural networks or probabilistic graphical models have been used for some time already and found their way into industry-standard software platforms. Fuzzy methods have yet to follow them on this route. We have discussed how visualisation and automation concepts can support data analysis tasks and that fuzzy methods can play an important role by providing them with key features like representing similarities or heuristic expert knowledge. In addition, fuzzy methods can provide genuine advantages in areas like approximate matching or cluster analysis.

Fuzzy data analysis research has focused a lot on simple analytical scenarios like classification and function approximation. However, many realistic data analysis scenarios are much more complex, like the examples from Volkswagen and BT illustrate. While fuzzy methods are already used "behind the scenes" to support complex analytical scenarios more research activities should concentrate on developing fuzzy methods for complex real-world analytics.

Acknowledgement

We would like to thank Jörg Gebhardt for valuable details of the item planning application at Volkswagen.

References

[1] Agena, "Bayesian Network and Simulation Software for Risk Analysis and Decision Support," 2007. http://www.agenarisk.com

[2] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules Between Sets of Items in Large Databases," *Proc. Conf. on Management of Data*, 207–216. ACM Press, New York, NY, USA, 1993.

[3] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Verkamo, "Fast Discovery of Association Rules," *Advances in Knowledge Discovery and Data Mining*, pp. 307–328. AAAI Press / MIT Press, Cambridge, CA, USA, 1996.

[4] M. Boettcher, D. Nauck, C. Borgelt, and R. Kruse, "A Framework for Discovering Interesting Business Changes from Data," *BT Technology Journal*, vol. 24, no. 2, pp. 219–228, British Telecom, London, United Kingdom, 2006.

[5] C. Borgelt and R. Kruse, *Graphical Models — Methods for Data Analysis and Mining*. John Wiley & Sons, United Kingdom, 2002.

[6] C. Borgelt and M.R. Berthold, "Mining Molecular Fragments: Finding Relevant Substructures of Molecules," *Proc. IEEE Int. Conf. on Data Mining (ICDM 2002, Maebashi, Japan)*, pp. 51–58. IEEE Press, Piscataway, NJ, USA, 2002.

[7] C. Borgelt, "Canonical Forms for Frequent Graph Mining," *Proc. 30th Ann. Conf. German Classification Society (GfKl 2006, Berlin, Germany)*, pp. 337–349. Springer-Verlag, Heidelberg, Germany, 2006.

[8] C. Borgelt, "Combining Ring Extensions and Canonical Form Pruning," *Proc. 4th Int. Workshop on Mining and Learning with Graphs (MLG 2006, Berlin, Germany)*. ECML/PKDD Organization Committee, Berlin, Germany, 2006.

[9] E. Castillo, J.M. Gutiérrez, and A.S. Hadi, *Expert Systems and Probabilistic Network*. Springer-Verlag, Berlin, Germany 1997.

[10] Y. Cheng, U. Fayyad, and P.S. Bradley, "Efficient Discovery of Error-Tolerant Frequent Itemsets in High Dimensions," *Proc. 7th ACM SIGMOD Int. Conf. on Knowledge Discovery and Data Mining (KDD'01, San Francisco, CA)*, pp. 194–203. ACM Press, New York, NY, USA, 2001.

[11] N.J. van Eck, L. Waltman, J. van den Berg, and U. Kaymak, "Visualizing the WCCI 2006 Knowledge Domain," *Proc. IEEE Int. Conf. Fuzzy Systems (Vancouver, Canada)*, pp. 7862–7869. IEEE Press, Piscataway, NJ, USA, 2006.

[12] N.J. van Eck and L. Waltman, "VOS: A New Method for Visualizing Similarities Between Objects," *Proc. 30th Ann. Conf. German Classification Society*, Springer-Verlag, Berlin, Germany, 2006.

[13] N.J. van Eck, F. Frasincar, and J. van den Berg, "Visualizing Concept Associations Using Concept Density Maps," *Proc. 10th Int. Conf. Information Visualisation (IV 2006, London, UK)*, pp. 270–275. IEEE Press, Piscataway, NJ, USA, 2006.

[14] FairIsaac, "Falcon Fraud Manager," 2007. http://www.fairisaac.com/Fairisaac/Solutions/Product+Index/Falcon+Fraud+Manager

[15] J. Fatah, D. Nauck, and M. Boettcher, "Modelling Customer Satisfaction Using Bayesian Networks," *Proc. 11th Int. Conf. Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2006, Paris, France), Industrial Track*, pp. 37–44. Paris, France, 2006.

[16] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds., *Advances in Knowledge Discovery and Data Mining*. MIT Press, Menlo Park, CA, USA, 1996.

[17] J. Gebhardt, H. Detmer, and A.L. Madsen, "Predicting Parts Demand in the Automotive Industry — An Application of Probabilistic Graphical Models," *Proc. Int. Joint Conf. on Uncertainty in Artificial Intelligence (UAI'03, Acapulco, Mexico), Bayesian Modelling Applications Workshop*, Morgan Kaufman, San Mateo, CA, USA, 2003.

[18] J. Gebhardt, C. Borgelt, R. Kruse, and H. Detmer, "Knowledge Revision in Markov Networks," *Journal on Mathware and Soft Computing, Special Issue "From Modelling to Knowledge Extraction"*, vol. 11, no. 2–3, pp. 93–107, 2004.

[19] ISC Gebhardt, "Intelligent Systems Consulting," 2007. http://www.isc-gebhardt.de

[20] B. Goethals and M. Zaki, Eds., *Proc. 1st and 2nd IEEE ICDM Workshop on Frequent Itemset Mining Implementations*, CEUR Workshop Proceedings 90 and 126, 2003/2004. http://www.ceur-ws.org/Vol-90/ http://www.ceur-ws.org/Vol-126/

[21] J. Han, H. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation," *Proc. ACM Conf. Management of Data (SIGMOD'00, Dallas, TX)*, pp. 1–12. ACM Press, New York, NY, USA, 2000.

[22] D. Heckerman, "A Tutorial on Learning with Bayesian Networks," Microsoft Research, Advanced Technology Division, Redmond, WA, Tech. Rep. MSR-TR-95-06, 1995, revised 1996.

[23] H. Hofer, C. Borgelt, and M.R. Berthold, "Large Scale Mining of Molecular Fragments with Wildcards," *Intelligent Data Analysis*, vol. 8, pp. 495–504. IOS Press, Amsterdam, Netherlands, 2004.

[24] J. Huan, W. Wang, and J. Prins, "Efficient Mining of Frequent Subgraphs in the Presence of Isomorphism," *Proc. 3rd IEEE Int. Conf. on Data Mining (ICDM 2003, Melbourne, FL)*, pp. 549–552. IEEE Press, Piscataway, NJ, USA, 2003.

[25] F.V. Jensen, *An Introduction to Bayesian Networks*. UCL Press, London, United Kingdom, 1996.

[26] J. Kacprzyk, A. Wilbik, and S. Zadrozny, "Linguistic Summaries of Time Series via a Quantifier based Aggregation using the Sugeno Integral," *Proc. IEEE Int. Conf. on Fuzzy Systems (Vancouver, Canada)*, pp. 713–719. IEEE Press, Piscataway, NJ, USA 2006.

[27] S. Kramer, L. de Raedt, and C. Helma, "Molecular Feature Mining in HIV Data," *Proc. 7th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2001, San Francisco, CA)*, pp. 136–143. ACM Press, New York, NY, USA, 2001.

[28] M. Kuramochi and G. Karypis, "Frequent Subgraph Discovery," *Proc. 1st IEEE Int. Conf. on Data Mining (ICDM 2001, San Jose, CA)*, pp. 313–320. IEEE Press, Piscataway, NJ, USA, 2001.

[29] S.L. Lauritzen and D.J. Spiegelhalter, "Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems," *Journal of the Royal Statistical Society, Series B*, vol. 2, no. 50, pp. 157–224, 1988.

[30] H. Mannila, H. Toivonen, and A.I. Verkamo, "Discovery of Frequent Episodes in Event Sequences," *Tech. Rep. C-1997-15*. University of Helsinki, Finland, 1997.

[31] T.P. Martin and Y. Shen, "Soft Mapping Between Hierarchical Classifications," *Proc. 11th Int. Conf. Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2006, Paris, France)*, pp. 1027–1032. Paris, France, 2006.

[32] T.P. Martin and B. Azvine, "Evolution of Fuzzy Grammars to Aid Instance Matching," *Proc. IEEE Int. Symp. on Evolving Fuzzy Systems*, pp. 163–168. Ambleside, United Kingdom, 2006.

[33] T.P. Martin, B. Azvine, and Y. Shen, "Digital Obesity and the Need for Fuzzy Categories," *Proc. 6th Ann. Workshop on Computational Intelligence*, pp. 79–84. Leeds, United Kingdom, 2006.

[34] T.P. Martin, B. Majeed, B.-S. Lee, and N. Clarke, "Fuzzy Ambient Intelligence for Next Generation Telecare," *Proc. IEEE Int. Conf. on Fuzzy Systems (Vancouver, Canada)*, pp. 4285–4292. IEEE Press, Piscataway, NJ, USA 2006.

[35] T. Meinl, C. Borgelt, and M.R. Berthold, "Mining Fragments with Fuzzy Chains in Molecular Databases," *Proc. 2nd Int. Workshop on Mining Graphs, Trees, and Sequences (MGTS 2004 at PKDD 2004, Pisa, Italy)*, pp. 49–60. ECML/PKDD Organization Committee, Pisa, Italy, 2004.

[36] D. Nauck, "Measuring Interpretability in Rule-based Classification Systems," *Proc. IEEE Int. Conf. on Fuzzy Systems (St. Louis, MO)*, pp. 196–201. IEEE Press, Piscataway, NJ, USA, 2003.

[37] ——, "Fuzzy Data Analysis with NEFCLASS," *Int. J. Approximate Reasoning*, vol. 32, pp. 103–130, 2003.

[38] D. Nauck, M. Spott, and B. Azvine, "SPIDA — A Novel Data Analysis Tool," *BT Technology Journal*, vol. 21, no. 4, pp. 104–112, 2003.

[39] D. Nauck, D. Ruta, M. Spott, and B. Azvine, "A Tool for Intelligent Customer Analytics," *Proc. IEEE Int. Conf. Intelligent Systems (London, UK)*. IEEE Press, Piscataway, NJ, USA, 2007.

[40] D. Nauck, "GNU Fuzzy," *Proc. IEEE Int. Conf. on Fuzzy Systems (London, UK)*. IEEE Press, Piscataway, NJ, USA, 2007.

[41] S. Nijssen and J.N. Kok, "A Quickstart in Frequent Structure Mining Can Make a Difference," *Proc. 10th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD2004, Seattle, WA)*, pp. 647–652. ACM Press, New York, NY, USA, 2004.

[42] Oracle, "Oracle 10g Data Mining FAQ," 2005. http://www.oracle.com/technology/products/bi/odm/odm_10g_faq.html

[43] J. Pearl, "Aspects of Graphical Models Connected with Causality," *Proc. 49th Session of the Int. Statistics Institute*, 1993.

[44] J. Pearl and S. Russel, "Bayesian Networks," *Technical Report R-216*, University of Californa, Los Angeles, CA, USA, 1994.

[45] J. Pei, A.K.H. Tung, and J. Han, "Fault-Tolerant Frequent Pattern Mining: Problems and Challenges," *Proc. ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMK'01, Santa Babara, CA)*. Santa Barbara, CA, USA, 2001.

[46] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.

[47] S. Ramamurthy, H. Arora, and A. Ghosh, 2005. http://www.hugin.com/cases/Finance/Infosys/oprisk,

[48] M. Spott and D. Nauck, "On Choosing an Appropriate Data Analysis Algorithm," *Proc. IEEE Int. Conf. on Fuzzy Systems (Reno, NV)*, pp. 597–602. IEEE Press, Piscataway, NJ, USA, 2005.

[49] M. Steinbrecher and R. Kruse, "Visualization of Local Dependencies of Possibilistic Network Structures," *Proc. Int. Symp. of Fuzzy and Rough Sets (ISFUROS'06)*, pp. 77–80. UCLV, Santa Clara, Cuba, 2006.

[50] ——, "An Alternative Interpretation of Probabilistic Potentials for Exploratory Data Analysis," *Proc. 1st Joint Conf. German Statistical Society*, 2007.

[51] T. Verma and J. Pearl, "An Algorithm for Deciding if a Set of Observed Independencies has a Causal Explanation," *Proc. 8th Conf. on Uncertainty in Artificial Intelligence*, pp. 323–330. Morgan Kaufmann, San Francisco, CA, USA, 1992.

[52] X. Wang, C. Borgelt, and R. Kruse, "Mining Fuzzy Frequent Item Sets," *Proc. 11th Int. Fuzzy Systems Association World Congress (IFSA'05, Beijing, China)*, pp. 528–533. Tsinghua University Press and Springer-Verlag, Beijing, China, and Heidelberg, Germany, 2005.

[53] X. Yan and J. Han, "gSpan: Graph-Based Substructure Pattern Mining," *Proc. 2nd IEEE Int. Conf. on Data Mining (ICDM 2003, Maebashi, Japan)*, pp. 721–724. IEEE Press, Piscataway, NJ, USA, 2002.

[54] ——, "Closegraph: Mining Closed Frequent Graph Patterns," *Proc. 9th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2003, Washington, DC)*, pp. 286–295. ACM Press, New York, NY, USA, 2003.

[55] Y.Y. Yao and N. Zhong, "An Analysis of Quantitative Measures Associated with Rules," *Methodologies for Knowledge Discovery and Data Mining*. Springer-Verlag, Berlin, Germany 1999.

[56] M. Zaki, S. Parthasarathy, M. Ogihara, and W. Li, "New Algorithms for Fast Discovery of Association Rules," *Proc. 3rd Int. Conf. on Knowledge Discovery and Data Mining (KDD'97, Newport Beach, CA)*, pp. 283–296. AAAI Press, Menlo Park, CA, USA, 1997.