The Contribution of Data Mining in Information Science

Sherry Y. Chen and Xiaohui Liu

Department of Information Systems and Computing, Brunel University

Abstract

Information explosion is a serious challenge for current information institutions. On the other hand, data mining, which is the search for valuable information in large volumes of data, is one of the solutions to face this challenge. In the past several years, data mining has made a significant contribution to the field of information science. This paper examines the impacts of data mining by reviewing existing applications, including personalized environments, electronic commerce, and search engines. For these three types of applications, how data mining can enhance their functions is discussed. The reader of this paper is expected to get an overview of the state of the art research associated with these applications. Furthermore, we identify the limitations of current works and raise several directions for future research.

Keywords: Data Mining; Personalization, Search Engines, Electronic Commerce

Corresponding to *Dr. Sherry Y. Chen, Department of Information Systems and Computing, Brunel University, Uxbridge, Middlesex, UB8 3PH, UK Email: Sherry.Chen@brunel.ac.uk*

1. Introduction

As information institutions transform their role from passive data collection to a more active exploration and exploitation of information, they face a serious challenge: how can they handle a massive amount of data that the institutions generate, collect and store. There is a need to have a technology that can access, analyze, summarize, and interpret information intelligently and automatically. Responding to this challenge, the field of data mining has emerged. Data mining is the process of extracting valuable information from large amounts of data [[20]]. It can discover hidden relationships, patterns and interdependencies and generate rules to predict the correlations, which can help the institutions make critical decisions faster or with a greater degree of confidence [[17]].

In the past decade, data mining changes the discipline of information science, which investigates the properties of information and the methods and techniques used in the acquisition, analysis, organization, dissemination and use of information [[4]]. There is a wide range of data mining techniques, which has been successfully used in the field of information science. This paper is an attempt to illustrate how data mining can contribute to the field of information science by reviewing existing applications. To provide a sound understanding of data mining applications, it is necessary to build a clear link between tasks and applications. Therefore, the paper begins by explaining the key tasks that data mining can achieve, i.e. what can be achieved by data mining (Section 2). It then moves to discuss application domains, i.e. where data mining can be helpful (Section 3). The paper identifies three common application domains, including personalized environments, electronic commerce, and search engines. For each domain, how data mining can enhance the functions will be described. Subsequently, the problems of current research will be addressed, followed by a discussion of directions for future research (Section 4).

2. Tasks: What can be achieved?

Data mining can be used to achieve many types of tasks. Based on the types of knowledge to be discovered, it can be broadly divided into supervised discovery and unsupervised discovery. The former requires the data to be pre-classified. Each item is associated with a unique label, signifying the class in which the item belongs. In

contrast, the latter does not require pre-classification of the data and can form groups that share common characteristics [[36]]. To achieve these two main tasks, four data mining approaches are commonly used: classification, clustering, association rules, and visualization.

2.1. Clustering

Clustering, which is also known as *Exploratory Data Analysis* (EDA, [[46]]), is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups [[41]]. From a data mining perspective, clustering is unsupervised discovery of a hidden data concept. This approach is used in those situations where a training set of pre-classified records is unavailable. In other words, this technique has the advantages of uncovering unanticipated trends, correlations, or patterns, and no assumptions are made about the structure of the data. The management of customers' relationships (Section 3.2.1) is an example of using this approach.

2.2. Classifications

Classification, which is a process of supervised discovery, is an important issue in data mining. It refers to the data mining problem of attempting to discover predictive patterns where a predicted attribute is nominal or categorical. The predicted attribute is called the class. Subsequently, a data item is assigned to one of a predefined set of classes by examining its attributes [[9]]. In other words, the objective of classification is not to explore the data to discover interesting segments, but rather to decide how new items should be classified. One example of classification applications is to provide personalized learning environments based on users' characteristics, needs, and preferences (See Section 3.1).

2.3. Association Rules

Association rules that were first proposed by Agrawal and Srikant [[1]] are mainly used to find out the meaningful relationships between items or features that occur synchronously in databases [[47]]. This approach is useful when one has an idea of different associations that are being sought out. This is because one can find all kinds of correlations in a large data set. It has been widely applied to extract knowledge from web log data [[29]]. In particular, it is very popular among marketing managers and retailers in electronic commerce who want to find associative patterns among products (see Section 3.2.2).

2.4. Visualization

The visualization approach to data mining is based on an assumption that human beings are very good at perceiving structure in visual forms. The basic idea is to present the data in some visual form, allowing the human to gain insight from the data, draw conclusions, and directly interact with the data [[2]]. Since the user is directly involved in the exploration process, shifting and adjusting the exploration goals is automatically done if necessary [[32]]. This approach is especially useful when little is known about the data and the exploration goals are vague. One example of using visualization is author co-citation analysis (See Section 3.3.3).

3. Applications: Where can be Useful?

As the aforementioned discussion, data mining can be used to achieve various types of tasks, such as classification, clustering, association rules, and visualization. These tasks have been implemented in many application domains. The main application domains that data mining can support in the field of information science include personalized environments, electronic commerce, and search engines. Table 1 summarizes the main contributions of data mining in each application.

Section	Application	Contributions
2	Personalised Environments	▪ To adapt content presentation and navigation support based on each individual's characteristics. ▪ To understand users' access patterns by mining the data collected from log files. ▪ To tailor to the users' perceived preferences by matching usage and content profiles.
3	Electronic Commerce	▪ To divide the customers into several segments based on their similar purchasing behavior. ▪ To explore the association structure between the sales of different products. ▪ To discover patterns and predict future values by analyzing time series data.
4	Search Engine	▪ To identify the ranking of the pages by analyzing the interconnections of a series of related pages. ▪ To improve the precision by examining textual content and user's logs. ▪ To recognize the intellectual structure of works by analyzing how authors are cited together.

Table 1: Contributions of Data Mining

3.1. Personalized Environments

Increasingly, personalized applications are attempting to incorporate data mining [[12]]. Personalization refers to the automatic adjustment of information content, structure, and presentation tailored to an individual user [[39]]. One of the major applications that employs personalization is e-learning programs, which adapt content presentation and navigation support to each individual by using the classification approach to identify his/her characteristics, needs, and preferences. For example, Mitchell, Chen, and Macredie (2004) developed a personalized web-based learning environment according to students' cognitive styles [[33]]. In addition,

Esposito, Licchelli, and Semeraro (2004) built student models for an e-learning system based upon the student performance evaluation: good, sufficient or insufficient [[16]].

Prior to the emergence of data mining, most personalized applications adopted two methods: collaborative filtering and content-based filtering methods. In the content-based method, user profiles are created using features extracted from content that user liked or used in the past. Subsequently, new content is provided by matching user's profile to the features of this content [[26]]. The collaborative filtering method builds up profiles of user groups and then using a computational method tries to match current user's profile to similar profiles. Selected data from these profiles are then used to provide recommendations [[43]]. Nevertheless, most web objects are represented by a multimedia type of information so it is difficult to analyze web objects with the content-based filtering methods. In addition, the collaborative filtering method can only be applied within a homogenous type of information domain and suffers from the cold-start problem [[48]]. Other shortcomings include reliance on subjective user profiles that may be prone to biases, or on standing profiles that may become outdated with changing user needs or interests.

Recently it is indicated that usage mining can overcome shortcomings of the aforementioned two approaches to develop more advanced personalized applications [[34]]. In addition, usage mining and traditional personalized approaches can be integrated together or usage mining can be used with content mining. These approaches are described below.

3.1.1 Usage Mining

Usage mining is the process of automatically discovering and interpreting users' access patterns by mining the data collected from users' interaction with the system [[44]]. The extraction of users' interaction data is mainly obtained from log files, which record each individual request [[5]]. This approach is very commonly used for discovering hidden patterns within web access data because web server logs constitute a rich source of data collected in a non-intrusive way (Koutri, et al., forthcoming). Understanding of users' access patterns is useful to

provide effective personalization. Some recent projects, such as Web Personaliser [[34]] and PageGather [[37]], etc., have all concentrated on providing personalization with usage mining.

Web Personalizer [[34]] is a real-time personalized application and is produced based on a framework, which consists of four principal elements: modeling of web pages and users, categorization of web pages and users, mapping between and across web pages and users, and determination of the set of actions to be recommended for personalization. The personalization is provided according to web server logs, which are examined to discover clusters of users that exhibit similar browsing behaviors. These clusters are used to predict the needs of the current user based on their browsing similarities with previous users. A list of hypertext links that have not been visited and are not directly linked from the page is then recommended to the user for browsing the web site.

The PageGather [[37]] is an adaptive web site that semi-automatically improves the organization and presentation by learning from user access patterns. The system takes a Web server access log as input and maps it into a form ready for clustering. Cluster mining is then applied to find collections of related pages at a web site, relying on the visit-coherence assumption. An algorithm is developed to identify candidate link sets to include index pages based on user access logs. The algorithm has five basic steps: (1) process the access log into visits; (2) compute the co-occurrence frequencies between pages and create a similarity matrix; (3) create the graph corresponding to the matrix, and find maximal cliques (or connected components) in the graph; (4) rank the clusters found, and choose which to output; and (5) for each cluster, create a web page consisting of links to the documents in the cluster, and present it to the webmaster for evaluation.

3.1.2 Usage Mining with Collaborative Filtering

One of the examples that usage mining is integrated with traditional personalized approaches is Personalization Expert developed by Lee, Kim, and Rhee (2001). They present a framework that combines the collaborative filtering method with association rule mining to provide personalization. The collaborative filtering task is performed for each domain by using users' web object access information and generates the similarity

information among users that is valid within one specific domain and predicted rating value information of the objects. User similarity information over all the domains is gained by performing linear combination task. Using this information, the similar users can be found and these users' web object access information is to be the input data to discover object association rules. Discovered association rules among web objects are the information source that predicts whether certain web object is current user's favorite object or not. The results show that the proposed framework can provide better recommendation services by analyzing web access patterns of similar users [28].

3.1.3 Usage Mining with Content Mining

In addition to linking with traditional personalized approaches, usage mining can also be brought together with content mining, which extracts useful information from the content of documents. Mobasher et al. (2000b) present such an application, which mines web usage patterns and web content to personalize the user experience, specifically, to recommend new content the user may like to see [[35]]. In this framework, both usage and content profiles are incorporated into the recommendation process. The usage profiles are derived from transaction clusters and the content profiles are based on occurrence patterns of features in page views. The personalized content takes the form of recommended links or products, targeted advertisements, or text and graphics tailored to the user's perceived preferences as determined by the matching usage and content profiles. They found this framework could be applied to perform real-time personalization.

3.2. Electronic Commerce

The widespread use of the web has tremendous impact on the way organizations interact with their partners and customers. Many organizations consider analyzing customers' behavior, developing marketing strategies to create new consuming markets, and discovering hidden loyal customers as the key factors of success. Therefore, new techniques to promote electronic business become essential and data mining is one of the most popular techniques [[9]]. Data mining applications in electronic commerce include customer management, retail business, and time series analysis.

3.2.1 Customer Management

For analyzing customers' behaviors, a frequently used approach is to analyze their usage data in order to discover user interests, and then recommendations can be made based on the usage data extracted. Well-known recommendation systems include online food stores [[45]], music suggestions [[11]], and online bookstores such as Amazon.com [[30]], etc.

Basically, three approaches can be used for making recommendations: (1) collaborative filtering, (2) content-based filtering, and (3) cluster models. The details of the first two approaches have been described in Section 2. The third approach is to create cluster models, which divide the customers into many segments and treat the task as a classification problem. The algorithm's goal is to assign the user to the segment containing the most similar customers. It then uses the purchases and ratings of the customers in the segment to make recommendations. Cluster models can perform much of the computation offline, but recommendation quality is relatively poor [[42]]. A possible solution is to increase the number of segments, but this makes the online user–segment classification expensive.

It seems that the aforementioned approaches have different strengths and weaknesses. Recent studies tend to use a hybrid approach. One of the examples is the study by Wang et al. (2004), who have developed a recommendation system for the cosmetic business. In the system, they segmented the customers by clustering algorithms to discover different behavior groups so that customers in same group have similar purchase behavior. For each group's customers, they used the association rules algorithm to discover their purchase behavior. In addition, they scored each product for each customer who might be interested in it with the collaborative filtering approach and the content-based approach. They found that this approach could not only recommend the right product to the right person, but also recommend the right product to the right person at the right time [[47]].

3.2.2 Retail Business

The major area where retail businesses can benefit from data mining is in the area of market basket analysis [[3]]. The market basket analysis refers to the application of data analysis techniques to databases that store transactions from consumers buying choices of different products. The aim of the analysis is to understand the association structure between the sales of the different products available. Once the associations are found, they may help planning marketing policies. For instance, if there is a relationship between two products over time, then retailers can use this information to contact the customer, decreasing the chance that the customer will purchase the product from a competitor. This is the type of data typically analyzed with association rules.

A recent development is that this approach can be used in detecting patterns in library circulation data. Cunningham and Frank (1999) applied the techniques of data mining to the task of detecting subject classification categories that co-occur in transaction records of books borrowed from a university library. They found that this information can be useful in directing users to additional portions of the collection that may contain documents relevant to their information needs, and in determining a library's physical layout [[14]]. The other new progress is that Chen et al. (2004) propose store-chain association rules for a multi-store environment. The format of the rules is similar to that of the traditional association rules. However, the rules also contain information on store (location) and time where the rules hold. The results of the proposed method may contain rules that are applicable to the entire chain without time restriction or to a subset of stores in specific time intervals [[12]].

3.2.3 Time Series Analysis

In business operations, decision makers need a solid description of future possible developments in order to formally include the future uncertainty in a decision process [[18]]. One of the ways to achieve such a description is by analyzing time series data. A time series is a sequence of observations that is ordered in time, which can be useful for the discovery and use of patterns and prediction of future values [[15]]. Time series analysis is an integral part of effective electronic commerce and is useful for many applications, such as economic forecasting, sales forecasting, budgetary analysis, and stock market analysis.

The main difference between traditional time series analysis and data mining on time series is that the latter can handle a large number of series. Many time series may be collected during normal business operations, so data mining is a useful technique to analyze time series data for electronic commerce. Liu et al. (2001) applied the time series data mining processes in a fast-food restaurant franchise. They found that the adaptation of data mining on time series provides several advantages: (1) to process large amounts of data in an automated fashion, (2) to convert vast amounts of data into information that is useful for inventory planning, labor scheduling, and food preparation planning, and (3) to offer a consistent, reliable and accurate method of forecasting inventory and product depletion rates over set temporal periods commonly used in restaurants. They claimed that such time series data mining could also be applied to other business operations [[31]].

3.3. Search Engines

Data Mining is of increasing importance for search engines. Traditional search engines offer limited assistance to users in locating the relevant information they need. Data mining can help search engines to provide more advanced features. According to current applications, there are three potential advantages: (a) ranking of pages, (b) improvement of precision, and (c) citation analysis. These advantages are described below.

3.3.1 Ranking of Pages

Data mining identifies the ranking of the web pages for a particular topic by analyzing the interconnections of a series of related pages. The PageRank [[6]] and Hyperlink-Induced Topic Search [[22]] apply this approach to find pertinent web pages. In the PageRank, importance of page is calculated based on the number of pages that points to it. This is actually a measure based on the number of backlinks to a page. A backlink is a link pointing to a page, rather than pointing out from a page. This measure is used to prioritize pages returned from a traditional search engine using keyword searching. Google applies this measure to rank the search results. The

benefit is that central, important, and authoritative web pages are given preferences. However, the problem is that it only examines the forward direction. In addition, a much larger set of linked documents is required.

HITS is different from the PageRank in that it examines both backward and forward direction. HITS identifies pages with most in-links and most out-links among a set of web pages in the same domain. The pages with most in-links are defined as *authorities*, and the pages with most out-links as *hubs*. There are two main steps: (a) a *sampling* component, which constructs a focused collection of several thousand web pages likely to be rich in relevant authorities; (b) a *weight-propagation* component, which determines numerical estimates of hub and authority weights by an iterative procedure. As the result, pages with highest weights are returned as hubs and authorities for the research topic. The benefit of this approach is that it can provide a densely linked community of related authorities and hubs.

3.3.2 Improvement of Precision

The problem of HITS is that it is a purely link structure-based computation, ignoring the textual content. Therefore, the precision is low. On a narrowly focused topic, it frequently returns resources for a more general topic. IBM Almaden Research Centre continued refinements of HITS to develop CLEVER search engine [[8]]. The main refinement is to promote the precision by combining content with link information, breaking large hub pages into smaller units, and computing relevance weight for pages. Following the CLEVER, Focused Crawling [[7]] is another further enhancement of HITS. The improvement is that Focused Crawling selectively seeks out pages that are relevant to a predefined set of topics. The topics are specified not using keywords, but using exemplary documents. Rather than collecting and indexing all accessible documents to be able to answer all possible ad hoc queries, a Focused Crawler analyzes its crawl boundary to find the links that are likely to be most relevant for the crawl, and avoids irrelevant regions of the web.

User's log is the other source that can be used to improve the precision. Zhang and Dong (2002) developed a Chinese image search engine named eeFind by using Matrix Analysis on Search Engine Log (MASEL). The basic idea of MASEL is to use the query log to find relations among users, queries, and clicks on results. The relation between pages chosen after a query and the query itself provides valuable information. After a query, a user usually performs a click to view one result page. Each click is considered a positive recommendation of that

result so that the system will provide further results based on the recommendation [[49]]. Guan and Wang (2003) used a similar approach to produce a search engine named Nstar. Like other search engines, Nstar needs users to provide keywords to define the scope of the search, and then to create a set of potential relevant results. Subsequently, a sample-based mining method is applied to extract further information. The sample-based mining method extracts information based on a sample specified by the users. The system then looks for similar parts from other pages automatically based on the pattern and style of the sample. The method is based on the observation that many Web sites are organized by the same institute, and thus the pages therein commonly exhibit very similar stylistic properties [[19]].

3.3.3 Citation Analyses

Author co-citation analysis (ACA) has been widely used as a method for analyzing the intellectual structure of science studies. It can be used to identify authors from the same or similar research fields. He and Hui (2002) used data mining to perform such an analysis and developed a data warehouse named Web Citation Database, which contained citation indices of Web publications that can be mined to help document retrieval. Hierarchical clustering is used as the mining technique for author clustering. This technique begins by considering each author to be a cluster. In the mining process, two clusters are combined together until, at the end, all of the authors are in a single cluster. Subsequently, multidimensional scaling is adopted for displaying author cluster maps, in which cluster represents a research area and authors within the same cluster are the experts of the same research area [[21]]. Moreover, other research applied visualization techniques to present ACA. Chen and Paul (2001) used 3D virtual landscape to represent author co-citation structures. The most influential scientists in the knowledge domain appear near the intellectual structure's center. In contrast, researchers who have unique expertise tend to appear in peripheral areas. The virtual landscape also lets users access further details regarding a particular author in the intellectual structure, such as a list of the author's most-cited papers, abstracts, and even the full content of that author's articles [[10]].

Co-citation analysis can also support the automatic indexing, which can autonomously locate articles, extract citations, identify citations to the same article that occur in different formats, and identify the context of citations in the body of articles. One of the examples is Research Index system (formerly known as CiteSeer), which works by downloading papers from the Web and converting them to text. It then parses the papers to extract the citations and the context in which the citations are made in the body of the paper. Furthermore, references are automatically linked to the documents they cite, creating a fully-fledged citation index. In other words, the system provides relevant papers to users by using common citations to make an estimate of document similarity [[27]].

4. Problems and Future Works

The above three application domains demonstrate that data mining is a very useful technology and opens new opportunities for data analyses. However, there are still some difficulties, which need to be aware of and should be investigated in further works:

4.1. Data Cleaning

It is important to know that the effective implementation of data mining methods depends on a large extent on the quality of the data. Therefore, a step of data cleaning before analysis is usually needed. Data cleaning deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. It becomes especially necessary when multiple data sources need to be integrated. This is because the sources often contain redundant data in different representations. Although the large number of tools indicates both the importance and difficulties of data cleaning, so far only a little research has appeared on data cleaning [[38]].

4.2. Interdisciplinary Research

Data mining is an interdisciplinary study, which embraces several current information technologies such as information visualization, machine learning, and soft computing. One of the significant problems for the interdisciplinary research is the range and level of domain of expertise that are present among potential users so it can be difficult to provide access mechanisms appropriate to all [[25]]. Therefore, there is a need to develop tools that can help to conduct good communications among experts from the fields of different technologies. In addition, further works are needed to integrate these technologies and this integration may offer more methodologies to investigate the problems of data mining.

4.3. Multimedia Mining

Multimedia mining is engaged to mine unstructured information and knowledge from multimedia sources. This is a challenging field due to the fact that Multimedia databases are widespread. There are tools for managing and searching within such collections, but the need for tools to extract hidden useful knowledge embedded within multimedia data is becoming critical for many applications. However, multimedia mining has received less attention than text mining [[23]], and it opens a new window for future research to explore. Today, we need tools for discovering relationships between data items or segments within images, classifying images based on their content, extracting patterns from sound, categorizing speech and music, recognizing and tracking objects in video streams, relations between different multimedia components, and cross-media object relations.

4.4. Assessment of Effectiveness

As described in Section 3, data mining can enhance the functions of various types of applications in the field of information science. However, the effectiveness of these applications is still a question. It is due to the fact that most previous evaluation concerned the comparison of different techniques, instead of assessing the effectiveness from user points of view. Certainly, there is a need to conduct more empirical studies to verify the effectiveness and evaluate the performance of these applications based on the users' position. Better understanding of users'

needs can help designers to develop effective applications that will take the advantages of the features of data mining into account at the same time focusing on users' requirements. In addition, such an evaluation can shed light on efficient use of data mining techniques and can potentially result in better end-to-end system performance, which in turn has a direct positive impact on the user experience.

5. Conclusions

Within the past 10 years, there has been significant progress in the field of information science. Some of this progress represents improvements in existing techniques. One of these techniques is data mining, which can search for interesting relationships and global patterns from various types of resources. These relationships and patterns represent valuable knowledge about the objects and this is reflected by many applications in the field of information science.

In this paper, we have given some background to data mining techniques. This knowledge is useful in selecting appropriate approaches for a specific application. The survey of applications presented in this paper provides additional insight into the contribution of data mining in information science. Interested readers may further explore applications in a specific area through the references listed in this paper. Understanding the scope and limitations of current applications can be very useful in developing new applications.

Data mining is widely used in many applications. The paper focuses on three main application domains, including electronic commerce, personalized environments, and search engines. It should be noted that data mining has also been applied to other application domains, such as bioinformatics, digital libraries, and web-based learning, etc. It is another direction for future research to investigate what major functions are required for each application domain and to develop concrete criteria for the evaluation of their effectiveness. These works can be integrated together to generate guidelines, which can be used for commercial and research communities to

select suitable data mining techniques. The ultimate goal is to enhance the functions and performance of these applications by exploiting the full potential of data mining techniques.

Acknowledgement

This study is in part funded by the Engineering and Physical Sciences Research Council in UK (Grant Reference: GR/R57737/01) and the Arts and Humanities Research Board in UK (Grant Reference: MRG/AN9183/APN16300).

6. References:

[1] Agrawal, R., & Srikant, R. Fast algorithms for mining association rules. P*roceedings of the 20th international conference on very large databases* (1994), Santiago, Chile.

[2] Ankerst, M.. Visual Data Mining with Pixel-oriented Visualization Techniques. *Proceedings of ACM SIGKDD Workshop on Visual Data Mining* (2001).

[3] Bigus, J. Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support. New York: McGraw-Hill (1996).

[4] Borko, H. (1968) Information Science: What is it?. American Documentation. 19(1), 3-5.

[5] Bose, I. & Mahapatra, R. K.. Business data mining - a machine learning perspective. *Information & Management.* 39(3) (2001) 211-225.

[6] Brin, S. & Page, L. The Anatomy of a Lage-Scale Hypertextual Web Search Engine. *Proceedings of the 7th International World-Wide Web Conference* (1998), pp.107-117.

[7] Chakrabarti, S. Berg, M. van den, Dom, B. Focused Crawling: A New Approach for Topic-Specific Resource Discovery, *Proceedings of the 8th International World Wide Web Conference* (1999).

[8] Chakrabarti, S., Dom, B. E., Kumar, S.R., Raghavan, P., Rajagopalan, S., Tomkins, A., Gibson, D. & Kleinberg, J.M. Mining the web's link structure. Computer. 32 (1999) 60-67.

[9] Changchien, S., & Lu, T. Mining association rules procedure to support on-line recommendation by customers and products fragmentation. *Expert Systems with Applications*. 20(4) (2001) 325-335

[10] Chen, C. and Paul, R.J. Visualising a knowledge domain's intellectual structure. Computer, 34(3) (2001) 65-71.

[11] Chen, H. C. and Chen, A.L.P. A music recommendation system based on music data grouping and user interests. Proceedings of the CIKM'01, Atlanta, Georgia (2001), pp. 231–238.

[12] Chen, Y. L., Tang, K., Shen, R. J. & Hu, Y. H. Market basket analysis in a multiple store environment. Decision Support Systems, forthcoming.

[13] Cooley, R., Mobasher, B., & Srivastava, J. (2000). Web Mining: Information and Patterns Discovery on the World Wide Web. *Proceeding of the 9th IEEE International Conference on Tools with Artificial Intelligence* (1997), pp. 558-567,

[14] Cunningham, S. and Frank, E. Market Basket Analysis of Library Circulation Data. *Proceedings of the Sixth International Conference on Neural Information Processing*, 2 (1999) 825-830.

[15] Depenya, F. & Gil-Allana, L. A. (2003) Testing of Nonstationary Cycles in Financial Time Series Data. (Available: http://www.unav.es/econom/investigacion/working/wp1503.pdf)

[16] Esposito , F., Licchelli, O., Semeraro, G. Discovering Student Models in e-learning Systems. *Journal of Universal Computer Science*, 10(1) (2004), 47-57

[17] Gargano, M. L. & Ragged, B. G. Data mining – a powerful information creating tool. *OCLC systems services*. 15(2) (1999) 81-90.

[18] Groenendijk, P., Lucas, A. & de Vries, C.. A hybrid joint moment ratio test for financial time series, Technical report, Vrije Universiteit, Amsterdam (1998)

[19] Guan, T. & Wong, K. F. Nstar: an interactive tool for local web search. *Information & Management*, 41(2) (2003) 213-225

[20] Hand D.J., Mannila H., and Smyth P. *Principles of data mining*, MIT Press (2001).

[21] He, Y. and Hui, S. C. Mining a Web Citation Database for author co-citation analysis, *Information Processing & Management*, 38(4) (2002) 491-508

[22] Kleinberg, J. Authoritative sources in a hyperlinked environment. *Journal of the ACM*. 46(5) (1999), 604-632.

[23] Kosala, R., & Blockeel, H. Web mining research: A survey. *ACM SIGKDD Explorations*, 2(1) (2000) 1-15

[24] Koutri, M. Avouris, N., Daskalaki, S. A survey on web usage mining techniques for web-based adaptive hypermedia systems. In S. Y. Chen and G.D.Magoulas (Eds), Adaptable and Adaptive Hypermedia Systems, Idea Group Inc. (forthcoming)

[25] Kuonen, D. Challenges in Bioinformatics for Statistical Data Miners. *Bulletin of the Swiss Statistical Society*, 46 (2003) 10-17

[26] Lang, K.. Newsweeder. Learning to filter netnews. *Proceedings of the 12th International Conference on Machine Learning* (1995), Tahoe City, California.

[27] Lawrence, S. Giles, C. L., Bollacker, K. D. Digital Libraries and Autonomous Citation Indexing. *IEEE Computer* 32(6) (1999) 67-71

[28] Lee, C. H., Kim, Y. H. & Rhee, P. K. Web personalization expert with combining collaborative filtering and association rule mining technique. *Expert Systems with Applications*. 21(3) (2001), 131-137

[29] Lee, K. C., Kim, J. S., Chung, Kwon, S. J. Fuzzy cognitive map approach to web-mining inference amplification. *Expert Systems with Applications*. 22(3) (2002). 197-211.

[30] Linden, G., Smith, B. & York, J. Amazon.com Recommendations: Item to Item Collaborative Filtering. *IEEE Internet Computing.* 7(1) (2003) 76-80

[31] Liu, L. Bhattacharyya, S., Sclove, S. L. Chen, R. & Lattyak, W. J. Data mining on time series: an illustration using fast-food restaurant franchise data. *Computational Statistics & Data Analysis*, 37(4) (2001) 455-476.

[32] Lopez, N., Kreuseler, M. & Schumann, H. A Scalable Framework for Information Visualization,º *IEEE Transactions on Visualization and Computer Graphics*, 8(1) (2002) 39-51.

[33] Mitchell, T., Chen, S. Y. and Macredie, R. D. Adapting Hypermedia to Cognitive Styles: Is it necessary?. *Proceedings of the Workshop on Individual Differences in Adaptive Hypermedia* (2004).

[34] Mobasher, B., Cooley, R., & Srivastava, J. Automatic personalization based on web usage mining. *Communications of the ACM*, 43(8) (2000a) 142-151.

[35] Mobasher, D., Dai, H., Luo, T., Sun, Y. & Zhu, J.. Integrating web usage and content mining for more effective personalization. *Proceedings of the International Conference on E-Commerce and Web Technologies* (2000b).

[36] Nolan, J. R. Computer systems that learn: an empirical study of the effect of noise on the performance of three classification methods. *Expert Systems with Applications.* 23(1) (2002) 39-47

[37] Perkowitz, M. and Etzioni, O. Towards adaptive web sites: Conceptual framework and case study. Artificial Intelligence Journal, 118(1/2) (2000) 245-275

[38] Rahm, E.; Do, H. Data Cleaning: Problems and Current Approaches, IEEE Bulletin of the Technical Committee on Data Engineering. 23(4) (2000) 3-14

[39] Ramakrishnan, N. Perugini, S. The Partial Evaluation Approach to Information Personalization, Technical Report cs.IR/0108003, Computing Research Repository (CoRR), (2001)

[40] Rantzau, R. (1997) Extended Concepts for Association Rule Discovery. Available at: http://elib.uni-stuttgart.de/opus/volltexte/2000/721/pdf/DIP_1554.pdf

[41] Roussinov, D. and Zhao, J. L. Automatic discovery of similarity relationships through Web mining. *Decision Support Systems*. 35(1) (2003), 149-166

[42] Schafer, J.B., Konstan, J.A., and Reidl, J. E-Commerce Recommendation Applications. *Data Mining and Knowledge Discovery*, 5(1/2) (2001) 115-153.

[43] Shardanand, U., & Maes, P. Social information filtering: algorithms for automating `Word of Mouth'. *Proceedings of the Conference on Human Factors in Computing Systems-CHI'95* (1995).

[44] Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. N.. Web usage mining: Discovery and applications of usage patterns from Web data. *SIGKDD Explorations*, 1(2) (2000) 12-23.

[45] Svensson, M., Laaksolahti, J. Höök, K., and Waern, A. A Recipe Based Online Food Store, In *Proceedings of the 2000 International Conference on Intelligent User Interfaces* (IUI 2000), New Orleans, Louisiana, USA, pp. 260 263

[46] Tukey, J. W. Exploratory data analysis. Reading, MA: Addison-Wesley Publishing Company (1977).

[47] Wang, Y. Chuang, Y., Hsu, M. and Keh, H. A personalized recommender system for the cosmetic business. *Expert Systems with Applications*, 26(3) (2004) 427-434

[48] Yu, P. S. Data mining and personalization technologies. *Proceedings of 6th International Conference on Advanced Systems for Advanced Applications* (1999) p. 6-13

[49] Zhang, D. and Dong, Y. A Novel Web Usage Mining Approach For Search Engine. Computer Networks 39(3) (2002) 303-310