The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0264-0473.htm

Using data mining to improve digital library services

Ana Kovacevic
Faculty of Security Studies, University of Belgrade, Belgrade, Serbia
Vladan Devedzic
Department of Software Engineering, FON – School of Business Administration, University of Belgrade, Belgrade, Serbia, and
Viktor Pocajt
Faculty of Technology and Metallurgy, University of Belgrade, Belgrade, Serbia

Received 2 September 2009
Revised 11 November 2009
Accepted 16 November 2009

Abstract

Purpose – This paper aims to propose a solution for recommending digital library services based on data mining techniques (clustering and predictive classification).

Design/methodology/approach – Data mining techniques are used to recommend digital library services based on the user's profile and search history. First, similar users were clustered together, based on their profiles and search behavior. Then predictive classification for recommending appropriate services to them was used. It has been shown that users in the same cluster have a high probability of accepting similar services or their patterns.

Findings – The results indicate that *k*-means clustering and Naive Bayes classification may be used to improve the accuracy of service recommendation. The overall accuracy is satisfying, while average accuracy depends on the specific service. The results were better for frequently occurring services.

Research limitations/implications – Datasets were used from the KOBSON digital library. Only clustering and predictive classification was applied. If the correlation between the service and the institution were higher, it would have better accuracy.

Originality/value – The paper applied different and efficient data mining techniques for clustering digital library users based on their profiles and their search behavior, i.e. users' interaction with library services, and obtain user patterns with respect to the library services they use. A digital library may apply this approach to offer appropriate services to new users more easily. The recommendations will be based on library items that similar users have already found useful.

Keywords Digital libraries, Databases, Serbia, Data handling, Service delivery

Paper type Research paper

1. Introduction

Although we are today overwhelmed with data, techniques for finding appropriate information are mostly based on syntax search or low-level multimedia features. For improving search results in interaction with digital libraries (DLs), other more intelligent techniques should be used, based on both top-down knowledge creation (e.g. ontologies, user modeling) and bottom-up automated knowledge extraction (e.g. data mining, web mining) (Chen, 2003).

Valuable information extracted from the collection of DL data can be integrated into the library's strategy, and can be used to improve library search (Chang and Chen, 2006). For an effective design of systems and particularly to help users to find

The Electronic Library
Vol. 28 No. 6, 2010
pp. 829-843

0264-0473
DOI 10.1108/02640471011093525

information more easily, it is crucial to understand how people perform searches. This is especially important in continuous development of technologies. By exploring users' behavior we try to understand better the users themselves and their information needs and provide them with better user-oriented applications. To achieve this goal, we can anticipate a specific user's needs and problems in advance, by using experience of other similar users.

The idea of a recommender system is to help users by advising them on relevant products/information by predicting in advance their interest in a product; this prediction is based on various types of information, e.g. users' past purchases and product features (Huang *et al.*, 2002). For a DL, user recommendations may be very helpful (Geisler *et al.*, 2001, Liao *et al.*, 2009).

To help DL users obtain useful information more easily, we can use data mining techniques. Since data mining techniques are very popular, many researchers have applied them in various domains. However, few are focused on the domain of DLs. Our main objective is to use data mining techniques to recommend specific services to DL users.

We have developed the REKOB system to support the users of a specific digital library called KOBSON (http://kobson.nb.rs). In REKOB, we apply different and efficient data mining techniques for clustering DL users based on their profiles and their search behavior. We do not apply data mining to the library documents, but to its services; thereafter we recommend an appropriate service to a new user. By services, we assume online journal services such as Science Direct, Springer, and Blackwell among others.

The paper is organized as follows: a review of related work is discussed in section 2; section 3 describes the REKOB system architecture; experimental results and evaluation are provided in section 4; and finally, we draw our conclusions and plans for future work in section 5.

2. Related work

Data mining is an iterative process of searching for new, previously hidden information in large volumes of data (Kantardzic, 2003; Devedzic, 2001). We apply data mining according to the CRISP-DM standard (Cross-industry Standard Process for Data Mining, www.crisp-dm.org) (Chapman *et al.*, 2000); it specifies the following phases in the process of data mining (see Figure 1):

- *Business understanding*. After defining the project objectives and requirements, formulate data mining problem definition and prepare initial strategy for achieving these goals.
- *Data understanding*. After collecting the data, analyze it to familiarize with it and discover initial insight; also evaluate the quality of data.
- *Data preparation*. Prepare the final data set from the initial raw data that will be used in the process. Select cases and variables that are appropriate for analyses and perform the necessary data transformations.
- *Modeling*. Apply appropriate modeling techniques and calibrate the model to optimize the results. The results show patterns (i.e. the model) discovered for the data analyzed. If necessary, loop back to the preparation phase to bring the form of the data in line with the specific requirements for particular data mining techniques.

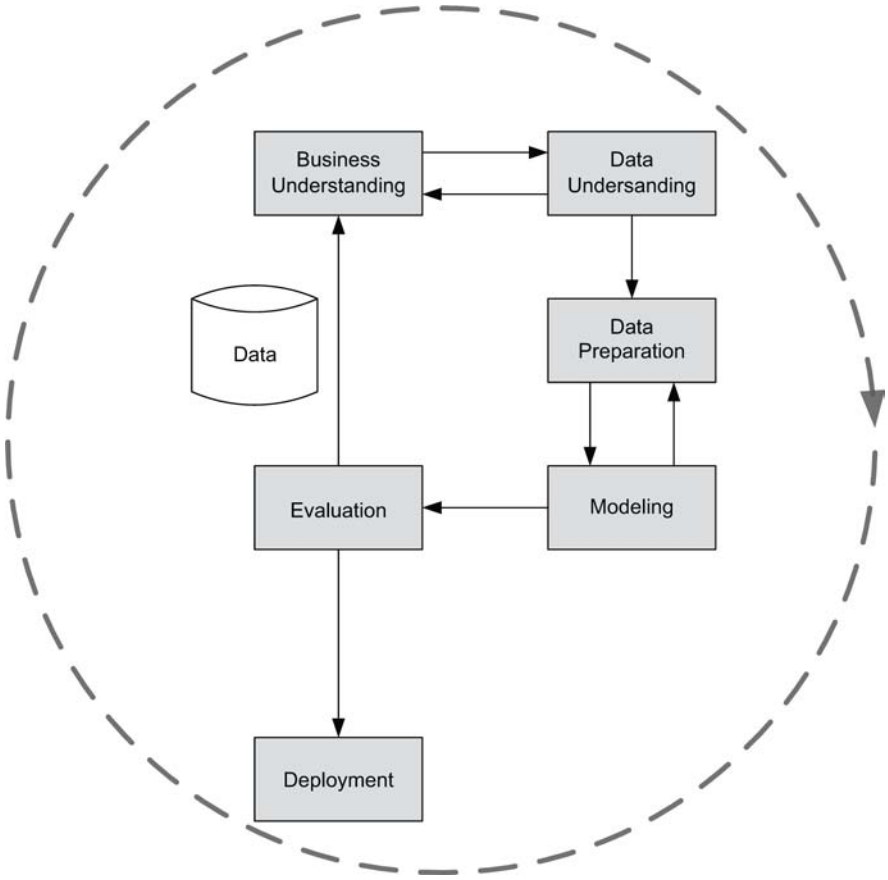

Figure 1.
Iterative process of data mining

- *Evaluation*. Evaluate the model (the patterns discovered) from the quality and effectiveness perspectives before deploying, and discover whether the model achieves the objectives set for it in the business-understanding phase.
- *Deployment*. make use of the model (patterns) created.

Note that it is very important to preprocess the data before continuing with the data mining process (Larose, 2005), otherwise the end-results will be misleading.

Recommender systems are implemented in various domains. In the commercial world, the best known example of a recommender system is Amazon.com (www.amazon.com), a web-based book seller, where personalized recommendation is made to its customers by combining three approaches: traditional collaborative filtering, cluster models, and search-based methods (Linden *et al.*, 2003). Gao *et al.* (2005) proposed a recommender approach to stock data analysis based on common user interests and a neural network. Although the experimental platform used a stock data set, their approach can also be useful in other numerical applications and in text-based digital libraries.

Qin *et al.* (2008) clustered partial preference relations as a means for agent prediction of users' preferences. New preferences for a user may be predicted by examining preferences of other users in the same cluster. They experimented on commonly used dataset for testing collaborative filtering technology: the MovieLens dataset (The MovieLens dataset is an existing dataset of movie ratings available at: http://grouplens.org). In the DL domain, Huang *et al.* (2002) used a graph-based recommender system that combines the content-based and collaborative approaches. Tsai and Chen (2008) recommend items from an e-library environment using adaptive resonance theory and data mining techniques. They cluster users by using adaptive resonance theory and then apply Apriori algorithm to generate rules for recommending appropriate literature.

3. REKOB

Our data mining system – REKOB – supports new users of the KOBSON library. In fact, KOBSON is the major provider of digital learning resources in Serbia (Consortium of Serbian Libraries) (Kosanovic, 2002), but in our research, we also use the term KOBSON to denote the corresponding DL in a technical sense. KOBSON provides Serbian students, teachers, and researchers access to foreign journals and learning resources. The DL services that KOBSON supports include Science Direct (SD, www.sciencedirect.com), Springer (SP, www.springerlink.com), Blackwell (BW, www.blackwell-synergy.com), Proquest (PQ, http://proquest.umi.com), JS (JSTORE, www.jstor.com), etc.

3.1 Overview of the data mining process

We have defined the data mining problem and made an initial plan for recommending KOBSON services to new users by analyzing the process with domain experts and identifying their need to improve KOBSON's user-oriented services. Our objective was to help new users of the KOBSON DL (as well as the ones who have a problem in finding relevant resources) in finding appropriate information by recommending them the service that similar users have found the most useful for them. The recommendation can be generated starting from the user's profile information and from the similarity of the user's behavior (when interacting with the KOBSON DL) with that of other users.

The data mining process is performed as illustrated in Figure 2. Users' search history data (from the KOBSON proxy server log) and their profiles (from the KOBSON database) are collected and then analyzed. Appropriate data from the log are parsed using regular expressions. From the log file, we extract users who have downloaded at least one paper from the DL. The parsed log data and user profiles are then loaded into the previously created database tables (Oracle 10 g rel. 10.2, www.oracle.com), as shown in Figure 2. In the next step, data preparation, these data are transformed and normalized into a format suitable for clustering the users.

After the data preparation is completed, the data mining is performed. The results of the data mining are user patterns – clusters of similar users generated based on the users' profile information and search behavior. This is done using the k-means data-mining algorithm (Campos *et al.*, 2003), further discussed in the Appendix. The user patterns are generated in the form of clustering rules (see section 3.5 and Figure 3 for details).

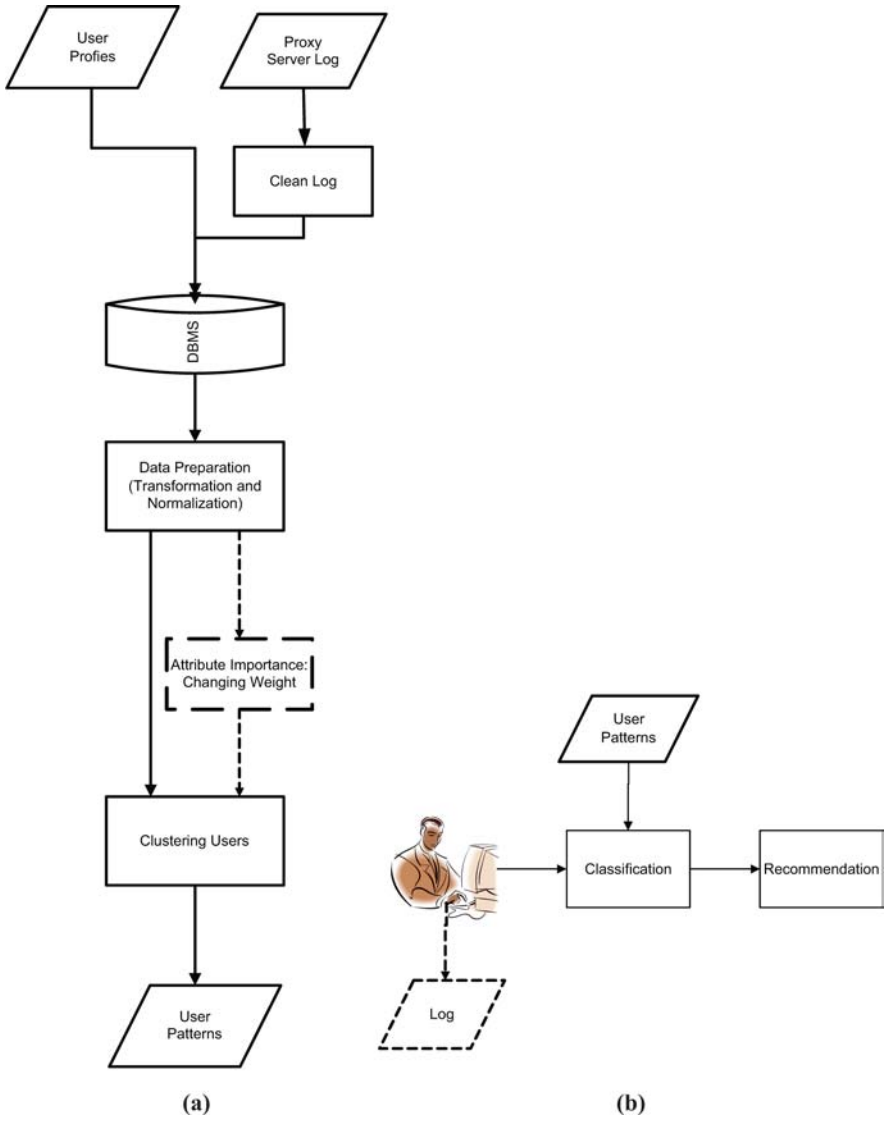

Figure 2.
The REKOB system data mining (a) processes (b) recommendation

In order to perform the classification of new users efficiently, the most significant attributes of data in each cluster should be identified. This is especially important in case, when there are a large number of attributes, to reduce them to the most significant ones. For discovering the most significant attributes, the minimum description length algorithm (MDL, 2008) is used (it is also further detailed in the Appendix). In REKOB, MDL is used to assign a higher weight to a specific, more significant attribute.

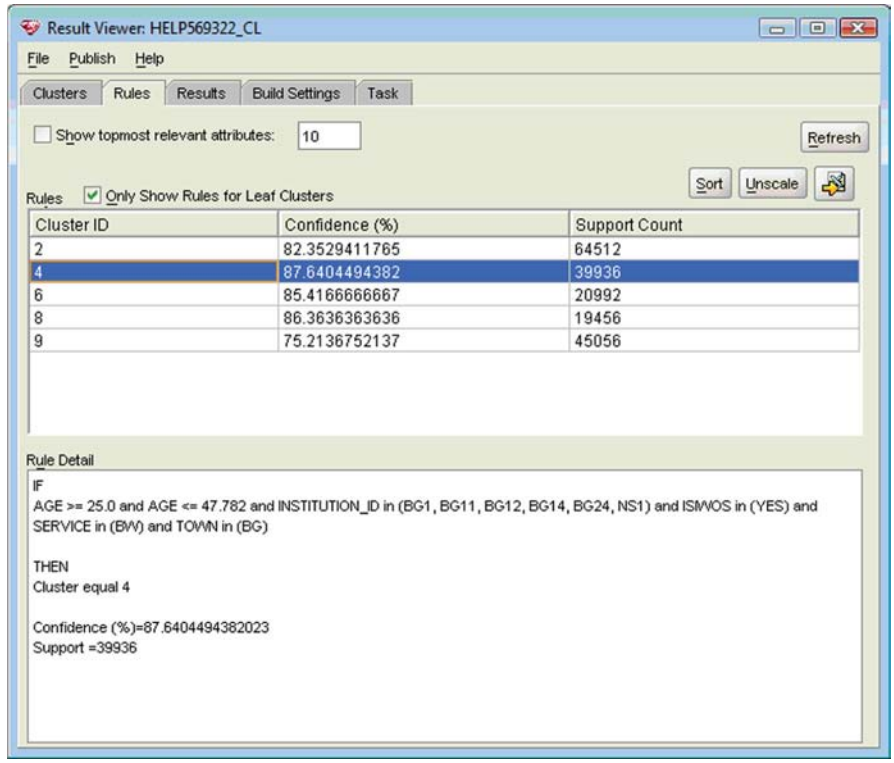

Figure 3.
An example of a clustering rule, with the corresponding confidence and support factors

3.2 Recommendations

Once the clusters (user patterns) are created, new users can be matched to them and the user category (the most similar cluster) can be determined (see Figure 2). Finally, a suitable recommendation for a new user is created based on the Naive Bayes classification algorithm (Larose, 2006). This algorithm is presented in more detail in the Appendix. Note that information about the new user's interaction is also logged into the system log. Likewise, information about all users is periodically refreshed in the log and the data mining is run again in order to update the user patterns.

3.3 Data

In the data mining process outlined in Figure 2, we used real-world data from the KOBSON DL (see Figure 2, top). To be able to use the resources and services of the KOBSON DL, a user has to register first. The user profile data collected through the registration procedure (user id, name, institution, age, gender and e-mail) is slightly modified for the data mining process. To protect the users' privacy, we replaced their names and institutions with codes. Hence, only authorized individuals will be aware of the users' identity and the institution he/she is with.

In addition to the users' demographic and administrative data collected during the registration process, we also used data collected from the proxy server (KOBSON,

2005), and related to the download of articles, using specific services from the DL (see the beginning of section 3). In the period 2004-2006, the number of downloaded articles per year was in the range 650,000-850,000.

We partitioned the data into a training data set and a test data set. We used the training data set to discover the user patterns, and then conducted pattern evaluations using the test data set. The results obtained with the test data are an indicator of how the user patterns will perform with the new data. We built and tested several models (i.e. several collections of clusters (user patterns)), using different techniques, and made a choice based of the performance on the test data set.

3.4 Data preparation

Preparing the data for data mining required to parse the data from the KOBSON proxy server log (see Table I) by using regular expressions and to extract only those users who had downloaded at least one paper.

After loading these selected data into the database (Oracle 10 g rel. 2), we obtained two relevant tables in the database: user (see Table II) and service (see Table III). The user table contains demographic and administrative information about the users of the KOBSON DL. The serivce table contains the data (extracted from the log, Table I) that describes instances of the DL usage patterns, which include the user, the service used, the date, and time of access, and the files downloaded. The initial version of the service table that we obtained contained about 230,000 records.

In the further data preparation, we join the user table and the service table and group by User_id and Service. The resulting table is the download table (see Table IV)

Attribute	Value
IP address	XXX.XXX.XXX.XXX
Agent	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)
Userid	User1
Date:time	[22/Sep/2004:14:18:16 + 0100]
Method	GET
URL	service/doc1.pdf
Protocol	HTTP/1.1
Status	200
Size	119562

Table I.
A record from the proxy server log

Attribute	Data type and precision
User_id	VARCHAR2 (9)
Name	VARCHAR2 (30)
Surname	VARCHAR2 (30)
Institution_id	NUMBER (6)
Town	VARCHAR2 (30)
Age	NUMBER (3)
Gender	VARCHAR2 (1)
Position	VARCHAR2 (20)
E-mail	VARCHAR2 (50)

Table II.
The user table in the database

and it is used in further DM process. User_id, Institution_id, Town, Gender, Age are attributes from the user table. The service attribute denotes the service used by the concrete user. The Rec_Num attribute stands for the number of papers downloaded by the specific user and by using the specific service.

3.5 Discovering user patterns

We used Oracle Data Miner10.2 (www.oracle.com/technology/products/bi/odm/odminer.html) in the data mining process. In particular, we relied on using OracleDataMiner JavaApi, but it is also possible to generate PL/SQL code for the specific model. First, we explored the data visually and statistically in order to familiarize better with the data and acquire the knowledge necessary to guide the data mining process. To do this, we used histograms (where the data distribution can be seen more easily). For example, if we want to analyze the age attribute, and we set the number of bins to five, the histogram will show five bars (or groups), age values for each group, the number of cases in each bin, and the corresponding percentage (see Figure 4).

Normalization of numerical data has been done by using min/max techniques in which all the data is normalized to the values between 0 and 1. In the process of indentifying similar users by applying the enhanced *k*-means and other data mining techniques (as described previously and in the Appendix), we set the maximum number of clusters to five and the maximum number of the algorithm iterations to ten. Euclidean distance has been used for the distance function.

After such a clustering, we extracted rules to describe the clusters, with the corresponding confidence and support factors (see the Appendix). An example of such a rule with the corresponding confidence is shown in Figure 3. Support count shows the number of cases for which the rule is true, and confidence (%) is the probability that a case described by this rule will actually be assigned to the cluster. In plain English, the meaning of the clustering rule shown in Figure 3 is:

Table III.
The service table in the database

Attribute	Data type and precision
User_id	VARCHAR2 (9)
Service	VARCHAR2 (10)
Date	DATE
File	VARCHAR2 (250)

Table IV.
The download table in the database

Attribute	Data type and precision
User_id	VARCHAR2 (20)
Institution_id	VARCHAR2 (20)
Town	VARCHAR2 (2)
Rec_num	NUMBER (9)
Gender	VARCHAR2 (1)
Age	NUMBER (3)
Service	VARCHAR2 (20)

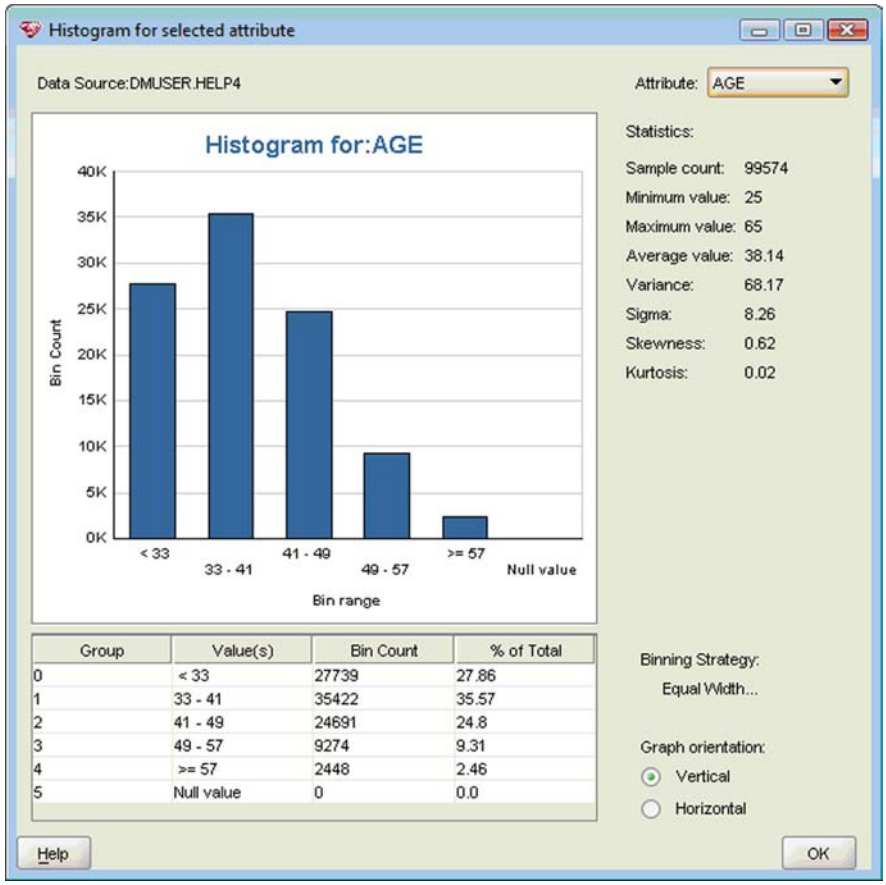

Figure 4.
A histogram for the age attribute

If:

- the user is a man, and he is between 25 and 47.74 years old;
- he is from a specific institution (e.g. BG1, BG11, BG12, BG14, NS1); and
- he is from the town "BG".

Then:

- he would use the BW service with the confidence of 80.48780 (support: 16,896 cases).

Note that this rule (as well as all other clustering rules discovered) relates users (see Table II) and KOBSON services they used (see Table III). In other words, the user patterns (clusters) actually describe information of the "which users use what KOBSON services" kind. This information is essential for recommending services to the new users, once they are matched to the user patterns (clusters).

In order to characterize the user patterns more effectively, we explored the most significant attributes of the KOBSON services. Finding the most significant attributes,

is especially important in case where there is a lot of attributes, and it is too expensive (and almost useless) to consider all of them within the data mining process. Figure 5 shows that the institution's code (institution_id) is the most important attribute for the target service, whereas gender is not so important (what may have been expected). In other words, Figure 5 reflects the importance of certain attributes of a specific service for recommending that service to the user. Accordingly, we may use this for weight factors, so institution_id will have a higher weight than the other attributes (see Figure 5) in clustering.

4. Evaluation

As can be seen in Figure 1, the crucial step before deploying the model is its evaluation.

Classification algorithms may be tested with new records, based on the classification model built. The results obtained with the test data, can be used as an indicator, of how well the model will work with new data. We used 60 per cent of our data for building the model and the remaining 40 per cent for testing it. The data are divided in two sets by random selection of cases using the split transformation.

The test indicated that the Naïve Bayes classification has 62.2 per cent better predicting confidence than the "naive" rule. The "naive" rule is usually used as the baseline for evaluating the performance of classification (Campos *et al.*, 2005), (Hamm, 2007). Using the naive rule method, a new record will be classified as a member of a

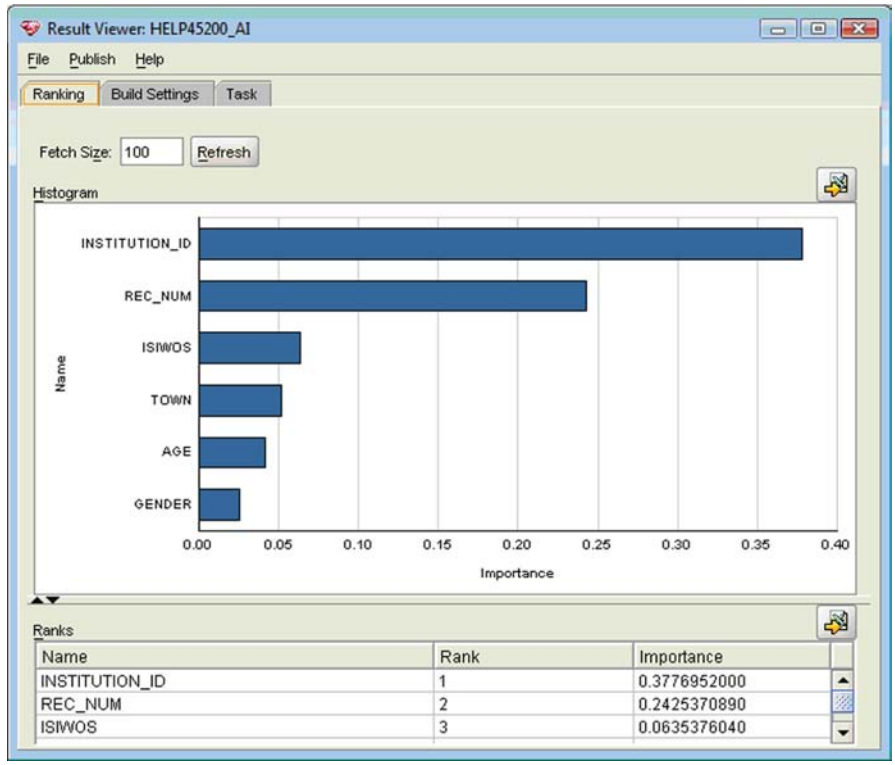

Figure 5.
Example of attribute importance

major class. So in our case, the naive rule would classify that all users would use the SD service (Science Direct), as a dominant service. The Decision Tree with clustering achieved a predictive confidence of 44 per cent better than the naive rule.

Furthermore, when we tried to test the classification using the Naïve Bayes algorithm without clustering the users first, we achieved a worse result – the predicting confidence was 43.49 per cent. Therefore, we chose Naïve Bayes with clustering for further matching. The rules for classification have been created, deployed, and added to the application.

We can further evaluate our model using the confusion matrix (see Figure 6). The confusion matrix is made on the test data, when the output value (in this case, the service the user has used) is known and is compared to the values predicted by the model. In the confusion matrix, the rows show the actual data, whereas the columns are the predictions made by the classification model. The confusion matrix measures the probability that the model predicts correct and incorrect values. Moreover, by analyzing the confusion matrix we can find out the types of errors that the model is likely to make. The sum of the values in the matrix is equal to the number of scored records in the input data table.

In the confusion matrix shown in Figure 6, it may be seen that the most frequent services, such as SD (Science Direct), PQ (ProQuest), and BW (Blackwell) are predicted

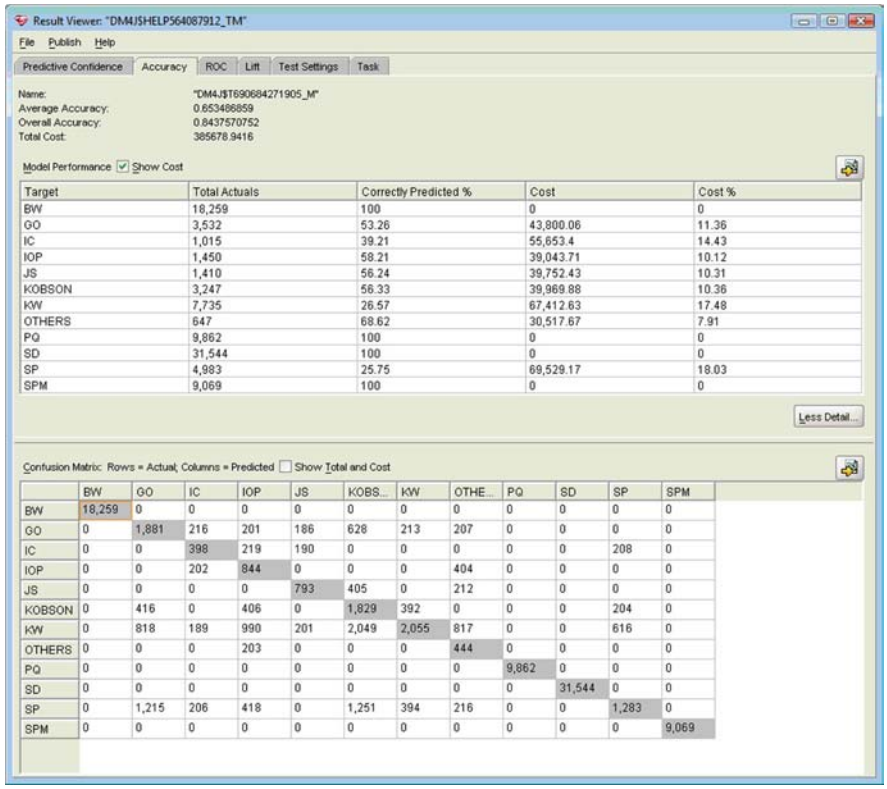

Figure 6.
Accuracy, cost and confusion matrix of the model

correctly, but there were errors in predicting less frequent services. For example, all of the 31.544 records for the SD service was predicted correctly, whereas for the JS (JStore) service 793 records were predicted correctly and 577 incorrectly. The correctness of its prediction is $793/(577 + 793) = 0.5788$ (or 57.88 per cent).

The overall accuracy is the simple accuracy of the model, whereas the average accuracy is the average per-class accuracy achieved at a specific probability threshold that is greater than the accuracy achieved at all other possible thresholds (ODM, 2005). The overall accuracy is 84, whereas the average accuracy is lower, 65. The prediction is the most likely target value for this case, and the probability is the confidence in that prediction.

5. Conclusion and future work

In this paper we have proposed a solution for recommending digital library users a service from the library, based not only on statistical significance of service usage, but also considering the users' profiles. Our main research was focused on helping users to find relevant material more easily. We achieved it by using data mining techniques on historical data and by recommending the services that similar users would choose. We first clustered the users based on their profiles together with their search behavior. It has been shown that the users in the same cluster have high preference for using similar services. The results show that the *k*-means clustering and the Naïve Bayes classification can be used together to improve the service recommendation. Finally, we applied our model to test data in order to evaluate its accuracy. It has been shown that the overall accuracy is satisfactory, especially in frequently occurring services. In the near future, we plan to add an effective visual representation for recommending specific services to the users and to discover most significant problems that the users encounter by using text mining techniques to analyze the users' e-mails or free text interviews.

References

Campos, M., Milenova, B. and Mccracken, M. (2003), *Enhanced* K-*means Clustering*, United States Patent Application 20030212520, available at: www.freepatentsonline.com/y2003/0212520.html (accessed 26 August 2009).

Campos, M., Stengard, P. and Milenova, B. (2005), "Data-centric automated data mining", *Fourth International Conference on Machine Learning and Applications*, IEEE Computer Society, Los Angeles, CA, pp. 97-104.

Chang, C.C. and Chen, R.S. (2006), "Using data mining technology to solve classification problems – a case study of campus digital library", *The Electronic Library*, Vol. 24 No. 3, pp. 307-21.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinart, T., Shearer, C. and Wirth, R. (2000), *CRISP–DM Step-by-step Data Mining Guide*, available at: www.crisp-dm.org/CRISPWP-0800.pdf (accessed 26 August 2009).

Chen, H. (2003), *Towards Building Digital Library as an Institution of Knowledge*, NSF Post Digital Library Futures Workshop, Chatham, MA, available at www.sis.pitt.edu/%7Edlwkshop/paper_chen.html (accessed 26 August 2009).

Devedzic, V. (2001), "Knowledge discovery and data mining in databases", in Chang, S.K. (Ed.), *Handbook of Software Engineering and Knowledge Engineering Vol. 1 – Fundamentals*, World Scientific Publishing, Singapore, pp. 615-37.

Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P. (1996), "From data mining to knowledge discovery: an overview", in Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (Eds), *Advances in Knowledge Discovery and Data Mining*, American Association for Artificial Intelligence, Menlo Park, CA, pp. 1-34.

Gao, K., Wang, Y.-C. and Wang, Z.Q. (2005), "Similar interest clustering and partial back-propagation-based recommendation in digital library", *Library Hi Tech*, Vol. 23 No. 4, pp. 587-97.

Geisler, G., McArthur, D. and Giersch, S. (2001), "Developing recommendation services for a digital library with uncertain and changing data", *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital libraries, Roanoke, VI*, ACM, New York, NY, pp. 199-200.

Hamm, C.K. (2007), *Oracle Data Mining Gold from Your Warehouse*, Rampant TechPress, Kittrell, NC.

Huang, Z., Chung, W., Ong, T.-H. and Chen, H. (2002), "A graph-based recommender system for digital library", *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries, Portland, OR*, ACM, New York, NY, pp. 65-73.

Kantardzic, M. (2003), *Data Mining: Concepts, Models, Methods, and Algorithms*, John Wiley & Sons, Hoboken, NJ.

KOBSON (2005), internal data of the project on the evaluation of the Serbian authors publishing productivity.

Kosanovic, B. (2002), "Koordinirana nabavka inostranih izvora naučno-tehničkih informacija u Srbiji: stanje i perspektive", *Infoteka*, Vol. 3 Nos 1-2, pp. 55-64.

Larose, D. (2005), *Discovering Knowledge in Data: An Introduction to Data Mining*, John Wiley & Sons, Hoboken, NJ.

Larose, D. (2006), *Data Mining: Methods and Models*, John Wiley & Sons, Hoboken, NJ.

Liao, S., Kao, K., Liao, I., Chen, H. and Huang, S. (2009), "PORE: a personal ontology recommender system for digital libraries", *The Electronic Library*, Vol. 27 No. 3, pp. 496-508.

Linden, G., Smith, B. and York, J. (2003), "Amazon.com recommendations: item-to-item collaborative filtering", *Internet Computing*, Vol. 7 No. 1, pp. 76-80.

MDL (2008), *Minimum Description Length, Oracle® Data Mining Concepts*, available at: http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/algo_mdl.htm#CHDGBDAJ (accessed 26 August 2009).

ODM (2005), *Oracle Data Mining Concepts 10g Release 2*, available at: http://download.oracle.com/docs/html/B14339_01/4descriptive.htm#i1005741 (accessed 26 August 2009).

Qin, M., Buffett, S. and Fleming, M.W. (2008), "Predicting user preferences via similarity-based clustering", *Proceedings of the Twenty-first Canadian Conference on Artificial Intelligence, Windsor, Ontario*, Springer-Verlag, Heidelberg, pp. 222-33.

Rissanen, J. (1978), "Modeling by shortest data description", *Automatica*, Vol. 14, pp. 465-71.

Tsai, C.S. and Chen, M.Y. (2008), "Using adaptive resonance theory and data-mining techniques for materials recommendation based on the e-library environment", *The Electronic Library*, Vol. 26 No. 3, pp. 287-302.

Appendix

Clustering

For grouping similar users together, we employed an unsupervised learning technique-clustering. The data within a cluster are more similar to one another than to those in other clusters. We cluster users according to their personal profiles and search behavior

(extracted from the log) by using the enhanced *k*-means algorithm (Campos *et al.*, 2003). It is a hierarchical top-down clustering algorithm, which generates clusters according to a user-specified criterion. The algorithm randomly defines initial centroids, which approximate centers of gravity, and uses a distance measure to calculate the distance between centroids and data objects (Hamm, 2007). It assigns probabilities to the generated clusters, by assuming that, the data in each cluster follows an isometric Gaussian distribution.

The rules that define when a data item fits into a cluster are specified as If-Then assertions of the form: "IF A implies B, THEN the cluster is Cluster1". Typically, these clustering rules are uncertain, and there are two certainty measures associated with each rule – confidence and support. For each rule, the confidence is the conditional probability of B given A, and the support for the rule is the estimation of the number of cases for which the rule is true.

The minimum description length (MDL) was first introduced by Rissanen (1978). The idea is that the simplest, most compact representation of data is the best and most probable explanation of the data. In the MDL algorithm, each attribute is considered as a simple predictive model of the target class and attributes are ranked according to their significance in predicting a target (MDL, 2008).

Classification

Classification is the task of mapping data into predefined classes (Fayyad *et al.*, 1996). Classification task begins with training data for which the target values are known. Classification of a collection consists of dividing the items that make up the collection into categories or classes (ODM, 2005). The model is built on historical data, and the goal is to predict the target class for each record of new data. In a specific case, the most likely target class is given as a prediction, and an appropriate probability indicates the confidence level of the prediction. Confusion matrix is used to display the type of errors the model is likely to make and is used for model evaluation.

Different classification algorithms use different techniques for finding relations between the predictor attributes' values and the target attributes' values in the build data. There is no perfect algorithm for all problems; the usefulness of an algorithm depends on the size of the dataset, the types of patterns that may exist in the data, the type of data, the goal of analysis, and many other factors (Hamm, 2007).

Generally, classification may be done for a binary target or for a multiclass target. (Binary targets have only two values, whereas multiclass targets have more than two values.) In a multiclass target, we may rank top choices for each case. In REKOB, the target has been multiclass. The data may be split in two sets, one for building the classification model and the other for testing the built model, where the model accuracy in predicting the target value is measured. In REKOB, we used the Naive Bayes classification algorithm (Larose, 2006). Because it is very fast and the attributes are independent from each other. Also, it is suitable for a multiclassification problem and for small number of predictor attributes. The Naïve Bayes algorithm is based on the Bayes' Theorem that states that the probability of event A occurring given that event B has occurred (P(A|B)) is proportional to the probability of event B occurring given that event A has occurred multiplied by the probability of event A occurring ((P(B|A)P(A)) (ODM, 2005). Naive Bayes algorithm makes the assumption, that each attribute is conditionally independent of the others. Even if this assumption of independence is violated in practice, it does not significantly degrade the model's predictive accuracy. In addition to the Naïve Bayes algorithm, we also tested the Decision Tree algorithm (Larose, 2005) for our classification case, but we achieved better results using the Naïve Bayes algorithm.

About the authors

Ana Kovacevic received her BSc and Msc degrees in Electrical Engineering from the University of Belgrade, where she also received her PhD in 2010. She is currently a Lecturer at the

University of Belgrade, Faculty of Security Studies. Her research interests include data mining, text mining, digital libraries, databases, visualization, and multimedia. She is a member of the GOOD OLD AI research network. Ana Kovacevic can be contacted at: kana@rcub.bg.ac.rs

Vladan Devedzic is a Professor of Computer Science with the Department of Software Engineering, FON – School of Business Administration, University of Belgrade. His long-term professional goal is to bring together ideas from the broad fields of intelligent systems and software engineering. His current professional and research interests include knowledge modeling, ontologies, semantic web, intelligent reasoning techniques, software engineering, and application of artificial intelligence to education and healthcare.

Viktor Pocajt received his BSc degrees in mechanical engineering (1988) and chemical engineering (1992), MBA (1994), and PhD (1999), all from the University of Belgrade. His main professional orientation is development and implementation of specialized e-business solutions and web-based information systems, applications and services for the metal working industry and environmental engineering. With more than 20 years of experience in metal working industry, research, IT development, applied computing, and consulting, he managed the implementation of several ERP systems, as well as development of a number of e-business solutions, specialized Web-oriented software systems and databases for the global market. Viktor has lectured on more than 70 specialized seminars and courses, and authored and co-authored about 40 scientific papers. He is currently a Professor at the Faculty of Technology and Metallurgy, University of Belgrade, Serbia.