Analysis of User Behavior in a Public Library using Bibliomining

PETR HÁJEK[1], JAN STEJSKAL[2]
[1]Institute of System Engineering and Informatics, [2]Institute of Economic Sciences
Faculty of Economics and Administration
University of Pardubice
Studentská 84, 532 10 Pardubice
CZECH REPUBLIC
petr.hajek@upce.cz, jan.stejskal@upce.cz

Abstract: - In the context of the provision of library services, bibliomining is defined as the process of applying data mining techniques to extract patterns of behavior from library databases. The application of bibliomining has hitherto focused on the optimal allocation of budgetary resources based on the past circulation of documents. There is no specific application of bibliomining methods to the recognition of user behavior patterns in the available literature. This paper aims to bridge this gap and presents the application of clustering methods on real data from a public library to obtain behavioral patterns of representative users. The results of this work can then be implemented to achieve more efficient management of public libraries. This paper develops a method of bibliomining, including the characteristics of the various stages of the process. Furthermore, the whole process is applied to research conducted in 2012 in the largest public library in the Czech Republic – the Municipal Library of Prague. The results are interpreted and a proposal for continuation of research is also included.

Key-Words: - public library, user behavior, bibliomining, cluster analysis.

1 Introduction

Data mining has recently gained attention through the realization of a wide range of technical and non-technical issues. In management and marketing, these issues are primarily the forecasting of supply and demand, customer classification or identification of patterns of customer behavior. Data mining can be defined as the process of extracting meaningful and useful patterns and rules from large data sets or databases. Data mining methods mainly include regression, classification, clustering methods, and association rules. These methods are used to obtain knowledge in a form which is understandable to the user, and usually in the form of conditional rules.

Bibliomining is defined in the context of the provision of library services as the process of applying data mining techniques to extract patterns of behavior from library databases [1,2]. This process consists of several steps, namely, target identification, collection and preprocessing of data sets, discovery of knowledge in data, result evaluation, and application of the acquired knowledge in practice. The aim of this process is to obtain behavioral patterns of library users that enable more efficient management of customer relations and continued improvement in the quality of services offered by libraries. Bibliomining is becoming an important tool for supporting library management's decision making processes.

In the past, research in the area of bibliomining has only focused on defining the concept [2] or the technological background of the application, which involves the process of designing a suitable database [3]. Bibliomining applications have hitherto been focused on the optimal allocation of budgetary resources based on the past circulation of documents [4]. What is lacking is a specific application of bibliomining methods based on the recognition of patterns of user behavior. This work aims to bridge this gap and presents the application of cluster methods and methods of extracting attributes of real data of a public library in order to find similarities in the services provided by public libraries based on a representative set of behavior patterns of users. The results of this work can be implemented to achieve more efficient public financing and contracting [15].

The rest of the paper is structured as follows. Section 2 is devoted to the characteristics of library services. There is then a description of data collection and processing. Section 4 characterizes the use of bibliomining methods and the results are interpreted. The paper ends with a summary of the conclusions.

2 Bibliomining

A conceptual framework for bibliomining was defined by Nicholson [2]. It includes five basic elements: operation of the library, bibliographic records, bibliometric data, library services and demographic structure of users.

Operation of the library is represented by the library staff, their patterns of behavior, time allocation, etc. Bibliographic records usually include the name of the item (document), its author, classification, abstract, keywords and data on its availability. Bibliometric data includes citations and cross-references (for more about bibliometrics see [5]). Library services can be divided into searching, circulation of documents (use), librarian assistance, educational and cultural programs, etc. Important information for the demographic structure of users includes place of residence, interests and membership in interest groups, employment and specific job position, education, age, etc.

Several recently proposed groups of methods belong to the area of biblioming. For example web usage mining, where a dataset is obtained from the behavior of users on the Internet or their IP addresses and records in log files [6]. This can track user behavior, how users reach a website, where they are directed next, in what sequence, and how long they send browsing the individual items, etc. Web usage mining has been applied for a research library in the work of Bollen et al. [7]. In [8] log files were analyzed by clusters based on graph structures in order to find communities of common interest.

Based on the textual content of documents it is possible to find a similar of set documents, extract keywords from documents, etc. Text mining is not focused directly on user behavior but on the analysis of text documents. The link to user behavior is achieved through the process of the user searching through a set of text documents. Text mining can therefore contribute to more effective searching and more accurate search results. Combining text mining methods with bibliometrics was performed in [9], where it was possible to remove unrelated links from the summary of document citations.

Bibliomining can be used to realize a wide range of issues associated with the behavior of library users. It can help libraries to reveal the future behavior of users by predicting the circulation of documents. These predictions can then be used to made decisions on the acquisition and allocation of funds. Furthermore, this knowledge can be used to make decisions about future between-borrowing cooperation. The knowledge gained by the circulation of documents can be further used to decide on the location of different types of items in the library, to change the opening hours and the number of librarians, etc.

3 Data Collection and Pre-processing

The survey was carried out in the pilot phase of the project Methodology for measuring the value of library services. It was a purely quantitative questionnaire survey conducted to verify the proposed method of determining empirical data among its respondents. The partial aim was to provide enough empirical data for verification of method of questioning the respondents on their individual preferences of the public services provided by the library, including questions on the valuation of their importance or value.

The basic set of respondents was made up of readers at the Municipal Library of Prague (MLP). The total number of members of the questionnaire panel was 1 061, a total of 620 responses were processed. Among the panel members were randomly selected MKP readers who meet the following criteria: age 15 + and loaned at least one item in 2010.

The panel members were surveyed using an online questionnaire (CAWI type) during October and November 2011. The method of questioning comes from foreign studies [10-12], which use WTA and WTP procedures to determine the respondents' opinions on the value of the services (for an explanation see [10]). Part of the questions was designed separately by the authors of this paper. The return rate of the questionnaires was 35% on average.

More than 88% of readers use the service loan of books and other media without the assistance of a librarian. This assistance, however, is utilized by 52% of users. Other services are used by less than half of the users. Approximately 43% of users spend time in the library reading documents retrieved without the assistance of a librarian. Less than 10% of users utilize the service of copying and printing of documents. A similar situation is observed in the case of reading and downloading of electronic documents. Half of the readers use the library's website to search for documents in the catalogue. About 38% need the services of a librarian to search. None of the users utilized the possibility of processing research documents and only 2% used the services "Ask a Librarian". Assistance of a librarian is mainly used for the purpose of obtaining

information on the availability of books in another branch or library (32%) and instruction on library services (20%). Approximately 17% of users participated in an educational or cultural program in the library, and 7% outside the library. About 41% of users perceive their stay in the library as a form of relaxation and 12% as a place to meet friends. In order to access the Internet, users utilize the library's computers (16%) in approximately the same extent as they use their own computers through Wi-Fi (17%).

The socio-economic and demographic characteristics of the users are as follows. About 80% of the users were women. The sample of users was dominated by persons aged 40 years and over (17% aged 40-49 and 30% aged 50 and over). Regarding education, about 37% had graduated from secondary school and 47% had a university degree. Most users were economically active (55%). In addition, students are represented (21%) and pensioners (16%), housewives (7%) and the unemployed (less than 1%).

The frequency of visits by users and their loan activity in the library can be characterized in the following way. District branches of the library are used more often. Approximately 60% of readers use only or mostly district branches of the library. In contrast, less than 7% of readers only use the central library. The frequency of visits is dominated by the readers who visit the library about once a month (68%). Most users loan 3-6 books per month (58%).

4 Bibliomining Methods Used

The K-means algorithm was used for the identification of services of the public library. This method belongs to the group of unsupervised learning methods as well as clustering methods. Clustering methods are used in data mining to identify certain groups of objects with similar characteristics.

The K-means method is a non-hierarchical clustering method. These methods are preferable for discovering knowledge in databases because they allow more efficient processing and interpretation of large data sets.

In non-hierarchical algorithms (such as K-mode or K-means algorithms) the samples are a predetermined number of clusters. In the case of K-means algorithms initial cluster centers are set first and then the samples, located within a given distance from the center of the cluster, are assigned to the cluster. The key step here is the initial setup of the centers of the clusters (here carried out using a hierarchical single linkage algorithm), which enables the efficient functioning of the K-means algorithm and reduces the possibility of the error (utility) function being stuck at the local minimum. The aim of K-means algorithms is to minimize the utility function J, which has the form:

$$J = \sum_{r=1}^{m} \sum_{i=1}^{n} \left\| \mathbf{p}_i^r - \mathbf{c}_r \right\|^2 \quad (1)$$

where $\mathbf{p}_i^r$ is the sample i belonging to the cluster r and $\mathbf{c}_r$ is the center of the cluster r.

For clustering a K-means algorithm is used for which a fixed number of clusters m=11 is chosen. This number was determined based on the shape of utility function J for a varying number of clusters m. In the area behind m=11 there was no further significant decrease in the utility function J.

5 Results and their Analysis

The results of clustering are presented in own calculated Tables 1 to 3. Table 1 shows the user behavior of typical readers in a particular cluster. This typical reader is represented by the center of the cluster. The first column shows the values with the highest frequency for the entire dataset in order to identify differences between typical readers of individual clusters from that of a single typical reader for the entire data set. A typical reader for the entire data set uses both loan services (home lending) without the assistance of a librarian with searching and with this assistance. Furthermore, she is a woman older than 50 years, economically active with a university education, using mostly a district library, visiting the library once a month with the monthly number of loans in the range of 3 to 4 items. In this way it is also possible to characterize the individual readers in the individual clusters.

Readers in clusters c_1 and c_5 are the least demanding on the library services. These users only use distance loans of book, magazines or CD retrieved either with or without the assistance of a librarian. They are economically active women with a secondary school education between 40 and 49 years (cluster c_1) or university education and age over 50 years (cluster c_5). They only or mostly use the services of a district branch of the library, and once a month to borrow 3-6 books. Cluster c_1 can be characterized as "*borrowers requiring assistance in searching*" and a cluster c_5 as "*independent borrower*s."

Typical readers in other clusters also use other services offered by the library. Cluster c_2 can be characterized as "*users searching for a specific title*." They use both the electronic catalogue and the assistance of a librarian. They are economically

ISBN: 978-1-61804-147-0

active women over 50 years old with a university education who visit the library once a month and use both the district and central libraries to the same degree to borrow 1-2 books. A typical reader in cluster c_3 searches for items without the assistance of a librarian and reads in the library for personal relaxation. He or she is retired and uses the library once a month to borrow 3-4 books. They are therefore people who do not come with a specific idea of what to borrow i.e. "*users looking for rest and relaxation*".

Cluster c_4 is another group of readers who can be described as "*users demanding on the assistance of a librarian*". They are economically active women aged 40-49 years who mostly visit a district library once a month borrowing 5 to 6 items.

Clusters c_6 and c_7 include students who differ in their use of services based on their hitherto level of education. Secondary school students are in cluster c_6, while in cluster c_7 there are university students. Students in cluster c_6 search in the catalogue and on the Internet looking for specific information related to Opencard (city card) services. They mostly use district libraries once a month with a lower frequency of borrowing i.e. 1 to 2 items. They can be described as "*students looking for specific information*". In contrast, students in cluster c_7 use the library for borrowing as well as in-house reading of documents that are retrieved both with and without the assistance of a librarian. They also search in electronic catalogues with and without the assistance of a librarian. They use the library both for personal relaxation and to study their own documents. Furthermore, they use the library computers or their own computers connected to the Internet via Wi-Fi for their studies and other activities. From the point of view of the library they are the most demanding customers and they can be characterized as "*customers demanding on modern technology*". Users in cluster c_9 also use the library for their studies. This is a less demanding user who mainly uses the central branch of the library borrowing to 5-6 items per month. They use only printed documents for their studies, so they can be described as "*studiers of printed documents*".

Users in cluster c_8 are university educated, aged 30-34 years, who search for books and other documents only with the assistance of a librarian, both for distance and for in-house reading. This corresponds to the low frequency of borrowed items. This cluster can be described as "*users needing the assistance of a librarian*."

User behavior in clusters c_{10} and c_{11} is similar; the difference lies only in the use of a librarian to search for the required books. While readers in cluster c_{10} (women with a university education, 30-34 years) make use of this service, 40 to 49 year old men with a secondary school education in cluster c_{11} do not. These clusters also differ in the frequency of use of library services. While the typical user in cluster c_{10} uses the library once a month and has a higher number of loans (5-6), a user in cluster c_{11} borrows fewer books (3-4), but the frequency of visits is the highest of all the clusters (more than once per week). These clusters can be described as "*young employed reader*s" (cluster c_{10}) and "*readers with a high frequency of visits*" (cluster c_{11}).

Table 1: Typical readers in the individual clusters

Serv.	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}
x_1	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
x_2	Y	Y	N	Y	N	N	Y	Y	N	Y	N
x_3	N	N	Y	N	N	N	Y	N	Y	N	N
x_4	N	N	N	N	N	N	Y	Y	N	N	N
x_5	N	N	N	N	N	N	N	N	N	N	N
x_6	N	N	N	N	N	N	N	N	N	N	N
x_7	N	N	N	N	N	N	N	N	N	N	N
x_8	N	N	N	N	N	N	N	N	N	N	N
x_9	N	N	N	N	N	N	N	N	N	N	N
x_{10}	N	N	N	N	N	N	N	N	N	N	N
x_{11}	N	N	N	N	N	N	N	N	N	N	N
x_{12}	N	N	N	N	N	N	N	N	N	N	N
x_{13}	N	N	N	N	N	N	N	N	N	N	N
x_{14}	N	Y	N	N	N	N	Y	N	Y	Y	Y
x_{15}	N	Y	N	N	N	Y	Y	Y	Y	Y	Y
x_{16}	N	Y	Y	Y	N	N	Y	N	N	N	N
x_{17}	N	N	N	N	N	N	N	N	N	N	N
x_{18}	N	N	N	Y	N	Y	Y	N	N	N	N
x_{19}	N	N	N	Y	N	N	Y	N	N	N	N
x_{20}	N	N	N	N	N	N	N	N	N	N	N
x_{21}	N	N	N	N	N	N	N	N	N	N	N
x_{22}	N	N	N	N	N	N	N	N	N	N	N
x_{23}	N	N	Y	N	N	N	Y	N	N	N	N
x_{24}	N	N	N	N	N	N	Y	N	Y	N	N
x_{25}	N	N	N	N	N	N	N	N	N	N	N
x_{26}	N	N	N	N	N	N	Y	N	N	N	N
x_{27}	N	N	N	N	N	N	Y	N	N	N	N
x_{28}	N	N	N	N	N	N	Y	N	N	N	N
x_{29}	N	N	N	N	N	N	N	N	N	N	N
x_{30}	N	N	N	N	N	Y	N	N	N	N	N

Legend: Y stands for yes (using) and N denotes no (not using); x_1 Loan of books, magazines, CD retrieved without the help of a librarian, x_2 Loan of books, magazines, CDs from stock, retrieved with the help of a librarian, x_3 Reading books, newspapers or magazines, etc. studying documents etc. in the library - retrieved without the help of a librarian, x_4 Reading of books, newspapers or magazines, etc. study of documents etc. in the library – retrieved with the help of a librarian, x_5 Copying of documents (without the assistance of a librarian), x_6 Copying of documents (with the assistance of a librarian), x_7 Printing of documents (without the assistance of a librarian), x_8 Printing of documents (with the help of a librarian), x_9 Use of electronic databases, x_{10} Reading e-books, x_{11} Downloading e-books from the website, x_{12} Downloading other electronic documents from the website, x_{13} Search in the paper catalogue, x_{14} Search in the electronic catalogue on a computer, x_{15} Search in the electronic catalogue on the website, x_{16} Search with assistance of a librarian, x_{17} Processing of written research, x_{18} Information on the availability of a book at another branch or library,x_{19} Asking a librarian for instruction on library services, x_{20} Use of the service "Ask a librarian", x_{21} Participation in an educational or cultural program in the library, x_{22} Participation in an educational or cultural

ISBN: 978-1-61804-147-0

program outside the library, x_{23} Spending time in the library for personal relaxation, x_{24} Spending time in the library for studying your own documents, x_{25} Meeting with friends, x_{26} Use of electrical plug sockets for your own computer, x_{27} Use of library computers to access the Internet, x_{28} Use of Wi-Fi connection in the library, x_{29} Use of technical equipment of the library, x_{30} Use of library for Opencard related services.

Table 2: Socio-economic and demographic characteristics of the typical readers

Serv.	c_1	c_2	c_3	c_4	c_5	c_6
x_{31}	fem	fem	fem	fem	fem	fem
x_{32}	40-49	50+	50+	40-49	50+	20-24
x_{33}	SS	Uni	SS	SS	Uni	SS
x_{34}	Eco	Eco	Pen	Eco	Eco	Stu
Serv.	c_7	c_8	c_9	c_{10}	c_{11}	
x_{31}	fem	fem	fem	fem	male	
x_{32}	20-24	30-34	20-24	30-34	40-49	
x_{33}	Uni	Uni	SS	Uni	SS	
x_{34}	Stu	Eco	Stu	Eco	Eco	

Legend: SS is secondary school education with graduation exam, Uni is university education, Eco is economically active, fem denotes female, Pen is pensioner and Stu is student.

Table 3: Characteristics of the typical readers based on the frequency of visits to the library

Service	c_1	c_2	c_3	c_4	c_5	c_6
x_{35}	MB	BC	MB	MB	B	MB
x_{36}	Mon	Mon	Mon	Mon	Mon	Mon
x_{37}	3-4	1-2	3-4	5-6	5-6	1-2
Service	c_7	c_8	c_9	c_{10}	c_{11}	
x_{35}	BC	MB	MC	MB	MC	
x_{36}	Mon	Mon	Mon	Mon	Wk	
x_{37}	3-4	1-2	5-6	5-6	3-4	

Legend: MB is mostly branch library, BC is both branch and central library the same, B is branch library only, MC is mostly central library, Mon is once per month and Wk is once per week.

In terms of frequency, the highest frequency (37% of users in total) is characterized by less demanding readers in clusters c_1 and c_5 i.e. "*borrowers requiring assistance in searching*" and "*independent borrowers*". The following are the clusters c_2 and c_3 i.e. "*users searching for a specific title*" and "*users looking for rest and relaxation*". These clusters are made up of users who are less demanding on additional library services. Other clusters contain less than 10% of users. On the other hand, these clusters use specific library services, such as increased assistance of a librarian (cluster c_4) or technology (cluster c_7). We can therefore say that in addition to the basic services – loans at home, other services are used at high levels only by selected groups of users.

6 Conclusion

Public institutions currently provide public services based on a commission or public order from the central authorities of the state administration. These are mainly decided based on the amount of public funds allocated. Another consideration can be the fulfillment of the electoral program or vested interests. However, neither of these latter aspects is applicable in the case of public libraries. These public services depend primarily on education and further contribute to the elimination of information asymmetry. However, there is no reliable method to determine the appropriate amount of public resources to be allocated and to ensure the effective level of public services.

This paper aims to contribute to libraries determining the extent of public services by determining the typical consumer and not based on financial allocation or evaluation of public services using ROI [13-14]. If it is possible to determine the typical consumer of library services, then it will be possible for the library management to ensure the provision of the appropriate level of public services. This can be used to determine the necessary level of financial resources needed to ensure the standard of public service of a certain kind. This will help donors to better target funds or control their use.

The empirical data and bibliomining methods used in this paper determined the typical features of user groups, including their characteristics. The most frequent clusters in this case were "*borrowers requiring assistance in searching*" and "*independent borrowers*". In terms of frequency then followed "*users searching for a specific title*" and "*users looking for rest and relaxation*". These four user groups included about 63% of all users in the studied sample of respondents.

The sample results can be used in practice to change the attitudes of library management. Specifically, it is possible to assume from the result the following:

- who the typical reader is, i.e. the customer of library services, from which it is possible to deduce what services, in what quantity, quality and time will be required,
- it is possible to use an inverse approach. It will be possible to determine who is not a typical customer and if there is a public interest the library management will change the marketing of the library and try to attract the customer,
- it is possible to determine the customer's anticipation of the public services, thus helping to determine the standard of public services,
- this allows an analysis of the services provided or a specific set of public services (an analysis of services has never been performed in public libraries - what it should involve, there is no taxonomy, etc.)

ISBN: 978-1-61804-147-0

- the results will affect the marketing of public institutions. It will no longer be possible to blame the immateriality of the services and hence the elusiveness of the characteristics or features of the marketing mix
- for the libraries an analysis of the book collections can be performed and their range can be reviewed depending on popularity/need or usefulness to the user.
- development of user needs can be monitored and analyzed over the long-term. Based on the results the library can update the range of public services offered.

The method has many obstacles to its use. One of the fundamental obstacles is the absence of an analysis of the services public libraries provide. The range is too broad and varies from library to library. The results of an analysis of library services would enable the creation of standards in this sector of public services.

The use of the above procedures in practice prevents the need for a large amount of empirical information, the need for its explanatory capability and representativeness. On the other hand, it represents a unique approach to a long-term unresolved issue and provides proof that it is possible to head in this direction.

Acknowledgements
This paper was prepared as part of the research project No. 3052/2012 SOLK entitled "Methodology for measuring the value of library services" funded by the Ministry of Culture of the Czech Republic.

References:

[1] S. Nicholson, The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making, *Information Technology and Libraries*, Vol.22, No.4, 2003, pp.146-151.

[2] S. Nicholson, The Basis for Bibliomining: Frameworks for Bringing together Usage-based Data Mining and Bibliometrics through Data Warehousing in Digital Library Services, *Information Processing and Management*, Vol.42, 2006, pp.785-804.

[3] K. Guenther, Applying Data Mining Principles to Library Data Collection, *Computers in Libraries*, Vol.20, No.4, 2000, pp.60-63.

[4] S.C. Kao, H.C. Hang, C.H. Lin, Decision Support for the Academic Library Acquisition Budget Allocation via Circulation Database Mining, *Information Processing and Management: an International Journal*, Vol.39, No.1, 2003, pp.133-147.

[5] E. Garfield, A.I. Pudovkin, V.S. Istomin, Mapping the Output of Topical Searches in the Web of Knowledge and the Case of Watson-Crick, *Information Technology and Libraries*, Vol.22, No.4, 2003, pp.183-187.

[6] D. Pierrakos, G. Paliouras, C. Papatheodorou, C.D. Spyropoulos, Web Usage Mining as a Tool for Personalization: A Survey, *User Modeling and User-Adapted Interaction Journal*, Vol.13, No.4, 2003, pp.311-372.

[7] J. Bollen, R. Luce, S. Vemulapalli, W. Xu, Usage Analysis for the Identification of Research Trends in Digital Libraries. *D-Lib Magazine*, Vol.9, No.5, 2003.

[8] C. Papatheodorou, S. Kapidakis, M. Sfakakis, A. Vassiliou, Mining User Communities in Digital Libraries, *Information Technology and Libraries*, Vol.22, No.4, 2003, pp.152-157.

[9] R. Kostoff, J. del Rio, J. Humenik, E. Garcia, A. Ramirez, Citation Mining: Integrating Text Mining and Bibliometrics for Research User Profiling, *Journal of the American Society for Information Science and Technology*, Vol. 52, No. 13, 2001, pp.1148-1156.

[10] Ohio, *Value for Money: Southwestern Ohio's Return from Investment in Public Library*, Driscoll & Fleeter, Columbus, OH, 2006.

[11] British Library, *Measuring our Value*, British Library Annual Report 2003/2004, 2004, available online at: http://www.bl.uk/pdf /measuring.pdf.

[12] T. Lynch, *A Study of Taxpayer Return on Investment (ROI) in Florida Public Libraries*. Tallahassee, FL: State Library and Archives of Florida, 2004.

[13] J. R. Matthews, *Measuring for Results – The Dimensions of Pulbic Library Effectiveness*, Libraries Unlimited, London, 2004.

[14] J. Stejskal, P. Kotatkova Stranska, K. Matatkova, P. Hajek, Public Services Value Determining – Case of Public Libraries. In *Proc. from 4th WSEAS World Multiconference on Applied Economics, Business and Development* (AEBD 12), Porto, 2012, pp.140-145.

[15] J. Nemec, L. Vitek, B. Merickova, Contractiong-out at Local Gernment Level: Theory and Selected Evidence from Czech and Slovak Republics, *Public Management Review*, Vo. 7, No. 4, 2005, pp. 638-647.