Application of text mining in library book procurement

Ding Cong

Ding Cong Library, Jinan University, Guangzhou, China

Abstract. With the diversification of readers' preferences, the continuous growth of book price and the shortage fund for purchasing books, it becomes increasingly important to explore reads' real needs and spend limited money on those needed books. In this context, Chinese Word Segmentation technology combined with Chinese Library Classification are used to explore readers' reading preferences by using news, management and finance disciplines' borrowing data of Jinan University in the year of 2015. Results showed that readers showed remarkable preferences for books of a specific subcategory or theme. Meanwhile the hot borrowed books often gathered in a few core publishing house, with the core publishing house books accounting for 70%, 63% and 76% of the total sampling books for news, management and finance discipline respectively. All in all, this research is of great significance for the book procurement job of the high school library.

Keywords. Book procurement, Book borrow, Chinese Library Classification, Chinese word segmentation.

1 Introduction

The function of university library is to serve teaching and research of university. Book procurement play an important part in enriching the collection of resource and improving the level of teaching and research. In recent years, a series changes have taken place in the internal and external environment of the library, which include the diversification of readers' preferences, the growing variety of books ,the continuous growth of book prices and the shortage fund for purchasing books. Under such circumstances, how to use the limited funds to purchase readers' needed books become a tough problem placed in front of the majority library staff. In order to resolve the above problems, the library community have done a series of fruitful research in how to find a reasonable book procurement method. Summing up, the main research aspects of this task are as follows.

Firstly, book procurement principle obtained by setting mathematical models based on library collection data, circulation data and reader information data. For example, based on Bayesian network theory, Yu Yali and Zhou Qingsong[1-2] obtain the procurement proportion of social science, art and literature books by establishing a model using library collection data, readers' circulation data. Peng Jun[3] proposed a second-choice algorithm and corresponding decision-making formula for book procurement. There are many such studies. Most of them borrow some principles from mathematics or economics, such as particle swarm algorithm, genetic algorithm and neural network algorithm. These kinds of research process prominent scientific nature. But they are complex in modeling and data collecting and thus lack utility. And meantime, it is not objective in determine the weight of each factor because of lack of objective standard.

Secondly, principles based on core press and core authors. The first case is Google Scholar based method. The research process are roughly as follows. Firstly, core bibliography are acquired by searching Google Scholar database to get the citation situation of these books coming from publisher. Secondly, core authors and core press are obtained by analysis and screening these core bibliography. Mu Weiguo's[4] paper is such case. The second case is library collection data based method. Combining Bradfort's law and twenty-eight rule, core authors and core press are acquired by the statistics large number of library collection books. The papers of Wang Jingfen[5-6] are such case. The guiding significance of such research to book procurement are great. However, the acquisition of core press and core authors are often time-consuming because of large amount of labor statistics work.

Thirdly, principles based on Patron-Driven Acquisitions(PDA) mode. The realization of such research are achieved by using the book recommendation system and the OPAC online catalog. In the first place, the bibliography information provided by book sellers and publishers are imported into search system. And then certain trigger mechanism which assign a threshold value are set. If searching number of certain book item exceed the threshold value then book procurement order will be trigger automatically[7-9]. This method process great significance because it is based on users' real needs. However, as a passive means to acquire readers' needs, it requires readers' active participation which in itself is a problem.

Fourthly, principles based on data mining. Most of these studies are purely theoretical research and lack empirical analysis. The common mode of these studies are about how to collect data from library collection data, circulation data and reader information data, how to clean up, transform and integrate data and how to put forward corresponding data mining principles and methods using these data[10-13]. This kind of research mostly use the relevant data accumulated by library to explore readers' needs actively. In this respect, it has certain advantage over PDA mode and avoids the problem of readers' low participation.

Undoubtedly, text mining falls into the category of data mining. As a actively method for exploring readers' needs, the tough problem of Chinese Word Segmentation(CWS) must be solved first. The ICTCLAS CWS system of Chinese Academy of Sciences is adopted in this paper. The hierarchical Hidden Markov Model(HMM) is used in this system, which can realize the tagging of words, the recognition of new words and the recognition of names, and support both simplified and traditional CWS with 98.45% accuracy. And meanwhile the system provides API which can be used to realize personal CWS needs.

In this paper, text mining method is used to explore the borrowing data of news discipline, management discipline and financial discipline of Jinan University Library(hereinafter referred to as the library). The branch class of the three disciplines' book borrow number, disciplines' reading preferences and disciplines books' core press are explored deeply. The analysis results can be used to guide book procurement process, formulating reasonable book procurement strategy.

2 Data sources and research methods

Data sources. The data for text mining come from the library's annual borrow ranking data accumulated in the year of 2015. Each book ranking record contains title, author, call number, press, borrowing times and other relevant data.

Research methods. In order to illustrate data mining process and meantime simplify the discussion process. Only the borrow data of three key disciplines(news, management and finance) of Jinan university library are taken to explore the law reflected in these borrowed data. The discussion for other disciplines can be imitated. The research ideas of this paper are as follows.

Firstly, filtering data with the call number starting with G21(journalism), C93(management) and F8(finance) from the borrowing database. The matching number of records for each type are 409,252 and 1974 respectively.

Secondly, the borrowing distribution situation of the branch class of the three disciplines are analyzed. Take G21 as example, as the branch class of G21, the borrowing distribution of G212(news gathering and reporting) and G211(organization and management) will be examined in detail. The basis for dividing discipline sub class is Chinese Library Classification. During the analysis, the method of traversing borrowing data is used in program to judge the sub classification of each book.

Thirdly, book titles for each of the three discipline's borrowed books are collected to form text string respectively. And then Ansj java package implementing the ICTCLAS algorithm is used to execute the Chinese word segmentation process for the above three book title string. Finally, statistical results are artificially filtered and the synonymy words are merged to obtain the frequency distribution of the segmented term.

Fourthly, by using the method mentioned above, the information of the publishing house of these books are segmented by CLC and the frequency distribution of these publishing house are counted.

3 Data analysis

3.1 Analysis of Book borrowing distribution of the branch class of news, management, and finance discipline

According to Chinese Library Classification, further subdivision for news, management and finance disciplines are done. News disciplines can be subdivided to G210, G211,G212 and several other subcategories. Management discipline is divided into C93, C93-0, C931 and several other subcategories. Financial discipline can be divided into four subcategories of F81, F82,F83 and F84. And then book borrowing data for these subcategories are counted respectively. The results are shown in Figure 1.

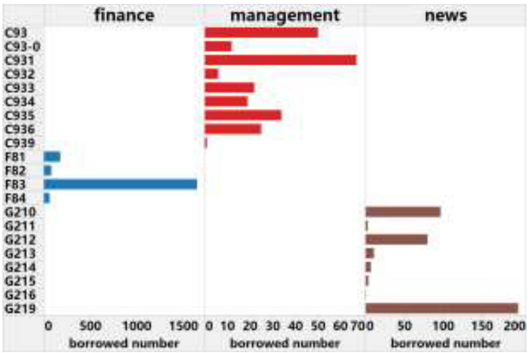

Fig. 1. Book Borrowing Distribution of the Subcategories of News, Management and Finance Disciplines

Note. G210-Journalism, G211-Organization and Management, G212-News Coverage, G213-Editing, G214-Journalists, G215-Newspaper publishing, G216-Various Newspaper, G219-Journalism in the World, C93-0-Management Theory and Methodology, C931-Management Technology and Methodology, C932-Consulting, C-933-Leadership, C93-Decision Science, C935-Management Planning and Control, C936-Management Histology, C939-Application Management, C93-Management, F81-Finance/State Finance, F82-Money, F83-Finance/Banking, F84-Insurance

As can be seen from Figure 1, the needs for book under the subcategories of news, management, finance disciplines have significance difference. Specifically, subcategories of Journalism in the World, Journalism and News Coverage rank the top three in the borrowing list. Whereas the borrowing number of management discipline are relative average with Management Technology and Methodology and Management ranking the top two list. Subcategory of Finance/Banking under finance discipline occupy an absolute advantage. The number of books borrowed under this class reached 1648, accounting for 83% in the total borrowed number. It is self-evidence that when formulating book procurement plans the above subcategories under the three disciplines should be focused. And meanwhile library collection distribution and subject development trend should also be considered to formulate a reasonable book procurement strategy.

3.2 Book title word frequency analysis for the borrowed books of news, management and finance discipline

Word frequency analysis is the concept of bibliometrics, which refers to the use of occurrences of key word capable of revealing the core content of literature to determine the hotspots and development trends of that field[14]. This paper uses this concept and apply it to analyze word frequency of book title of borrowed books. Word frequency of the book title of borrowed books are acquired according to research methods described above. And then for the three disciplines, word term with the occurrences above 5 times, 5 times and 10 times are filtered respectively and bubble chart is drawn into Fig 2 and Fig 3.

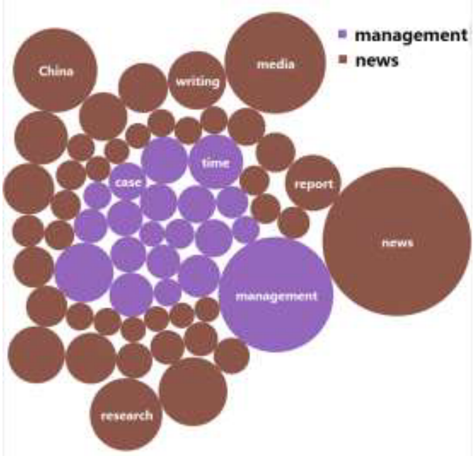

Fig. 2. Book titles' frequency distribution for news, management disciplines

Figure 2 and 3 use different colors to represent different disciplines with the size of the filled circle representing the occurrences times of that key word. It is possible to explore a reader's reading preference from these figure. For instance, the key word China appears in word segmentation results of the borrowed book title of news discipline which may indicate that readers of news discipline probably show great interest in news event in China. Again, there are high frequency of practitioner, qualifications and exam in the word segmentation results of finance discipline. This may indicate that readers own great willingness for obtaining qualification certificate in finance field. Of course, in actual practice, sampling survey must be done to confirm those potential reading preference. Once those reading preference are confirmed, books of such kinds can be increased accordingly.

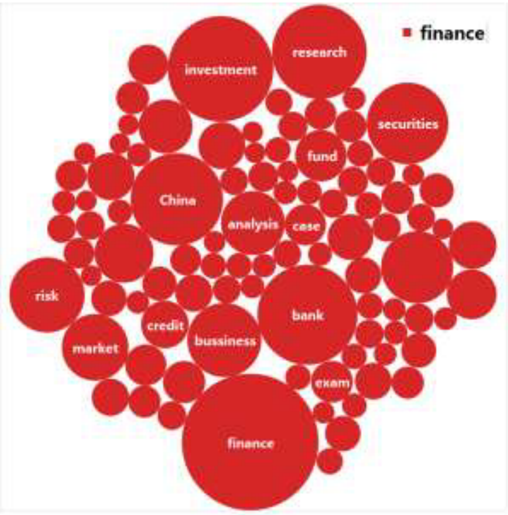

Fig. 3. Book titles' frequency distribution for finance discipline

3.3 Word frequency analysis for press of borrowed books in news, management and finance disciplines

Unlike other research methods, this paper uses text mining strategy for determining core press. To avoid including books with low borrow times into the statistics analysis process, this paper choose only those book borrow data with the borrow times more than 6 times. After statistics, the number of book borrow records meeting the above criteria are 57, 38 and 297 respectively. And text miming results are shown in figure 4.

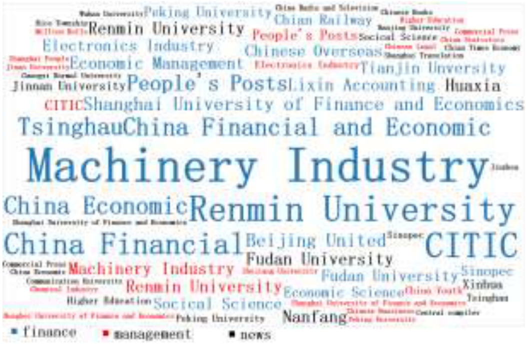

Fig. 4. Book publishers' frequency distribution for news, management and finance disciplines

Figure 4 represents different discipline with different color with the font size representing the occurrences times. After statistics, the press distribution of borrowed books of the three major disciplines are summarized as follows.

Firstly, there are 28 publishing houses in the journalism discipline, including NanFang Daily Publishing House, HuaXia Publishing House, Fudan University Press, Renmin University Press, Jinnan University Press, Xinhua Publishing House, Peking University Press, Tsinghua University Press, Higher Education Press, Social Sciences Literature Publishing House, China Petrochemical Press and etc. The mentioned 11 press have a total of 40 books on the list, accounting for 70% of the total sample.

Secondly, there are a total 20 publishing houses for the total 38 books sample of management discipline. Among them, Machinery Industry Press, Renmin University Press, CITIC Publishing House, People's Post Press, China Youth Publishing House and Electronics Industry Publishing House have a total of 24 books on the list, accounting for 63% of the total sample.

Thirdly, there are 70 press for the 297 books of finance discipline. Among them, Machinery Industry Press, Renmin University Press, CITIC Publishing House, China Financial Publishing House, China Financial and Economic Publishing House, China Economic Publishing House, Tsinghua University Press and other 10 press have a total 226 books on the list, accounting for 76% of the total sample. Notably, Machinery Industry Press, Renmin University Press and CITIC Press have 46, 56 and 24 books on the list respectively, occupying nearly one-third of the total sample.

4. Research conclusion and countermeasures

Based on the above analysis, the following conclusions can be drawn.

Firstly, it is possible to refine readers' demand based on the refinement borrowing data by Chinese Library Classification. In the process of making book purchase strategy, the borrow ratio for refined subcategories books can be referred. Meantime the actual library collection situation and development trend for a specific discipline should also be taking into consideration when making such book procurement strategy.

Secondly, readers' reading preferences can be explored from the word segmentation results of those borrowed books' title. These excavated information can be confirmed by way of sampling survey, which make the purpose of sampling survey greatly enhanced and is of great benefit for PDA-based book procurement.

Thirdly, publishing houses for borrowed book only distributed in a few of core publishing house. Books from these press usually possess high quality and are welcomed by readers. In the actual book procurement, it is a good practice to build on books from core press and at the same time supplement some high quality books from peripheral publishing house.

There are still several problems in this research. First of all, due to lack of reader information, there is an ambiguity in the direction of borrowing information which in in turn limits the use of reader information for accurate needs mining. Secondly, only single variable , such as call number, title or publishing house, not several variables combined, is used for data mining. This study can be enriched from the following aspects.

Firstly, acquiring reader information for borrowed books from OPAC and integrating those information to borrowing records. It is possible to utilize those information to analyze the reading preference of a specific reader or a group of reader(faculty or student) and formulate corresponding book procurement strategy aiming at those readers.

Secondly, integrating various data to explore readers' reading preferences. For example, according to the above analysis results, management books and finance books published by Machinery Industry Publishing House are welcomed. However, according to one's intuitive impression, Machinery Industry Press should be classified into engineering book publisher. Actually, if combined with author data, it is not difficult to find that a large part of the list books published by Machinery Industry House are English original translation. What management or finance readers really needs are the Chinese version for high level foreign original books.

References

1. Yu Yali. Application of Bayesian Network Theory in Library Book Procurement. Library and Information Service, 2015(7):126-127
2. Zhou Qingsong. Application of Bayesian Net in Book Procurement of Library. Yunan: Master of Yunan University, 2012:1-40
3. PengJun, Lumin, Ya Fayi et al. The Construction of Book Procurement Decision-making System in University Library Based on Quadratic Choice Algorithm. Practice Research, 2009, 32(9): 74-77
4. Mu Weiguo. Research on Book Purchasing Strategy Based on Core Bibliography and Popular Collection Analysis. Library Journal, 2015(8):44-50
5. Wang Jingfen, Gao Qi, Huang Jing et al. The Application of Core Publishing House in Purchasing Original Books. Library Journal, 2014(2): 64-66
6. Wang Jingfen, Gao Qi, Liang Weibo. The Application of the Core Authors in the Book Purchasing of Maritime. Library Journal, 2015(3):81-83
7. Jia Zhaoxia, Li Liwen. Acquisition Mode of Chinese Books in University Libraries Based on Purchasing Decision-making. Library Work and Study, 2014(4): 57-60
8. Hou Zhijiang, Hou Lingjuan. A New Method of Book Purchasing Based on OPAC Log User Behavior Analysis. Library Development, 2015(1):70-72
9. Tang Jisheng. A Review of Research on Patron-Driven Acquisitions(PDA) in China. Research on Library Science, 2015(2):22-28
10. Zhang Cunlu, Huang Peiqing, Wang Ziping. The Application of Data Mining in Book Procurement. Information Science, 2004, 22(5): 581-583
11. Song Yu. Research on Book Procurement Model Based on Data Mining. Research on Library Science, 2014(17): 53-55
12. Chi Chunjia, Mao Zhiyong. Research on Assistant Decision-making in Formulating University Book Purchasing Plan Based on Data Mining. Journal of Modern Information, 2009, 29(7): 108-110
13. Wu Jin, Mao Zhongxing. On the Book Purchasing Mode of University Libraries in the Background of Big Data. Journal of Changzhou University(Social Science Edition), 2014, 15(5): 133-135
14. Cun JieWang, Qian Qian. Analysis of Research Focus and Research Methods in the Field of Knowledge Management During the Past Decade. Information Science, 2014, 32(10): 157-157