Academic Libraries Data Analysis Using Data Mining

Yuqing Shi[1, 2, a], Yuelong Zhu[2, b], Jinghong Su[3, c], Lanfang Zhu[1, d], Dongmin Wu [1, e] and Chen Lin [1, f]

[1] Library, Hohai University Nanjing, China

[2] College of Computer and Information Engineering, Hohai University Nanjing, China

[3] Nanjing Hydraulic Research Institute, Nanjing 210029, Jiangsu, China

[a]shiyuqing@hhu.edu.cn, [b]ylzhu@hhu.edu.cn, [c]jhsu@nhri.cn, [d]lfzhu@hhu.edu.cn, [e]dmwu@hhu.edu.cn, [f]chlin@hhu.edu.cn

Keywords: Academic Libraries; Data Detection; Data Mining; Collections management; Data Analysis

Abstract. Nowadays, the requirement to preserve user privacy has avoided the preservation of data that could be used to correlate library data to non library data. This paper used data mining and data detection to identify academic library use patterns and to judge whether students' the number of published papers correlated to academic library use. All academic libraries datasets of this paper were uploaded into a data warehouse, allowing them to be managed, supervised and analyzed. The detection showed patterns of library use by academic department, patterns of book use from 2000 to 2015 15 years and statistics between the number of published papers and academic library use. The technique is demonstrated using data collected at Hohai University in China..

Introduction

Library data are always hard to analyze because the datasets may be very large, and these data come from unconnected sources. Academic libraries are cumulatively desired to supply evidence of their effectiveness in supporting institutional purposes, such as pushing faculty research, accelerating graduate student degree completion and helping publish papers. User assessments and investigations of instruction sessions are always used to afford proof of library efficacies. Whereas these tools should continue to be an important provenience of data, there are limitations to their efficacies in showing the unbiased effect of library collections and services on the academic role of the university. Explicitly, they trust with users offering a precise report of their own use of the academic library and its collections. A selective to feedback and surveys forms is to use actual data, data mining methods and data detection techniques to reveal patterns of use and correlations between user achievement and library activities [1].

Although data analysis is a crucial work of most library study, the data are always contained within a single data set of manageable size [2]. On the other side, data mining and data detection entails the use of a very large dataset or multiple datasets along with methods for analysis that reveal patterns [3]. The topics of data detection and data mining have been discussed in digital library literature for some years [4]. The purpose of data mining is to enhance the quality of the mutual effect between the library and its users. The collected data contain valuable information that can be used to improve library decisions, and can be integrated into the library's strategy. The goal of this paper is to illustrate how time series data mining technology is a good approach to accomplish users' requirements. The value of data detection and data mining for digital libraries is visible [5]. There is important possibility in the data collected by digital libraries, data detection and data mining techniques can use the data to assist with decision making, generate individualized services for specific users and afford evidence of digital library impact on academic research and achievement.

Many libraries collect data and use these data to answer questions such as popularity of various collections, the number of check outs, web site visits and so on [6]. In recent years, many studies from fields of librarianship, statistics data mining, and event prediction have contributed to research in digital libraries [7]. Among these methods, linear and nonlinear time series models are called

black-box models in which prediction is used. The black-box models try to found connection between model outputs and inputs. In this paper, forming partnerships with investigators on campus is a competent way to benefit immediately from the experience of other researchers, both in the management of user privacy concerns as same as in the analysis of data.

The rest of the paper is organized as follows: the next segment introduces data collection and pre-processing and third segment presents reader data analysis and findings. Fourth segment describes the relationship between the number of published papers and academic library use, and finally conclusions are consisted in the last segment.

Data

In this paper, the data collection contains six datasets. Book data: This dataset includes records on over 2.6 million volumes in Hohai University libraries and information details for each book, such as book ID, author, title, catalog date, publication date and so on. Reader data: This dataset contains all 52094 Hohai university staff and student information, including reader ID, reader type, major or department and other related information.. Renewals and Checkouts: This dataset covering the period from May 2000 to October 2015, consisting of 11,324 data points. WanFang Database: This database covering the period from 1990 to 2015. Engineering Village Database: This database is the most comprehensive bibliographic database of scientific and technical engineering research available, covering all engineering disciplines. It includes millions of bibliographic citations and abstracts from thousands of engineering journals and conference proceedings. ISI Web of Science Database: This database covering the period from 1986 to 2015.

Collecting data for readers to Hohai University library were comparatively simple due to it has imprinters at the each trance that require swiping a campus card. The library imprinters keep a log of all campus card swipes. A complete record of all campus card swipes for 2014 and 2015 was uploaded to the data warehouse and cleaned. The search for optimal pattern cluster is performed using Genetic Algorithms.

Reader Data Analysis and Findings

Using the data collected, data detection and data mining techniques were used to analyze readers' behavior, the patterns of use of Hohai University library and the correlation between reader' use of the library and the number of published papers. Via analysis, patterns emerged from the data and some findings are discussed in this paper. The findings are immature because with each finding, more questions are raised, requiring advanced analysis of dataset. It is this process of the cyclical review and iterative queries of the data that makes data detection and data mining such a powerful tool.

One spontaneous question is who uses Hohai University library most? Analysis of library check out patterns by reader type showed a pattern that was anticipated. The percentage of check outs was computed as follows: check out percentage equals the number of check outs by a reader type divided by the total check out number. The percentage of check outs coarsely corresponds to the size of the population, with faculty checking out the minimum number of items, followed by graduate students and then the undergraduates. This challenges the conception that undergraduates are no longer interested in paper characteristics materials, but that conception was also challenged by always observation of students lined up at the circulation desk. What is unexpected is the low percentage of all undergraduates who check out paper characteristics materials. For instance, less than 23.45 per cent of undergraduates checked out library paper characteristics materials in 2015 the first semester. During the same semester, the faculty checked out paper characteristics materials at the rate of just 15.2 per cent.

Different departments use the library in different patterns was found. The assumption is that the humanities departments will use the paper characteristics collections less than the sciences departments, and that assumption was supported by the Huiwen digital library data. The computed percentage equals the number of check outs by users in a department divided by the total number of check outs. With 63 departments and schools at Hohai University, the low percentages are not

surprising. The College of Water Conservancy and Hydropower Engineering takes the top spot by a large margin. At all events, what is surprising is to see Business School in the top three. The Business School of Hohai University has its own library system, and check outs in their system were not included in this research.

From research, some members of the College of Water Conservancy and Hydropower Engineering were always in Hohai library. However, there are also many other people in the Business School besides the ones we see in the library. The computed percentage equals the number of readers in a department who checked out at least one paper characteristics material divided by the total number of people in the department. All of the top sixteen departments support our research that investigators in the sciences are more continual users of Hohai library than researchers in the humanities and social sciences.

The subject distribution of the books was also examined, based on Classification for Library of the Chinese Academy of Sciences, checked out by readers affiliated with the different academic departments. Analysis of various material types was also conducted. It is obvious that the first and second year students are checking out fewer books than the upper level students. Knowing that circulation patterns have changed over time, it is impolitic to forecast patterns by using only data from materials purchased in financial year 2000.

To evaluate the physical use of the academic library, an analysis of imprinter counts was conducted for Hohai University library. Undergraduates use the library in greater numbers than other groups were known, but they are also the largest group. Consequently, the question was not simply about the numbers of each user group, but the percentage of each group using the academic library. The result of data detection shows that graduate students use the academic library in the greatest percentage, followed by undergraduates and, finally, the faculty.

Relationship between the Number of Published Papers and Academic Library Use

In this paper, another research question deal with the number of published papers and academic library use. A Pearson correlation analysis was done collating imprinter and book check out with the number of published papers. In statistics and data detection, the Pearson correlation coefficient is a dimension of the linear dependence correlation between two variables, giving a value between inclusive -1 and +1. It is widely used as a measure of the strength of linear dependence between two variables in the sciences. A greatly uncorrelated data will be close to 0, negatively correlated data approach -1, and positive correlation approaches the value +1.

In this paper, there is little, if any, relationship between data from one semester's book check out or use of Hohai University library and that semester's number of published papers. There is a work-out positive correlation when a semester's the number of published papers is compared with use of Hohai University library, though book check out still explains little relationship. Whereas the low level of positive correlation is disillusion, it is clear that data analysis must be reined.

Detaching undergraduates by class year explains that seniors do have a much stronger relationship between the number of published papers and academic library use. In this paper, it is still true that cumulative academic library use has a stronger relationship than a single semester's use. The other three classes have very weak positive relationship. The correlation is also examined between the usage of Hohai University library and Hohai University the number of published papers by class year using Spearman's rank correlation test, which is the Pearson correlation coefficient between ranked variables. In this paper, the observe results that are consistent with data analysis.

To reveal the data more clearly, a scatter gram is created platting seniors' number of published papers and check out activity. Because the dataset in this paper is very large, seniors are chosen who borrowed between 50 and 1000 books in academic year 2013. The scatter gram explains a large number of students with high number of published papers regardless of academic library check out activity. Anyway, the overall pattern explains that higher GPAs are correlated with higher check-out activity.

Detaching undergraduates by their various schools and college of Arts & Sciences by the areas of mathematics, technology, social sciences and science, engineering and humanities revealed some

interesting findings. Environmental Hydrology, Water Resources, Computer , Law and so on were the ten schools with uniformly positive correlations between the number of published Chinese papers and academic library use, and this held true for individual semesters as same as cumulative academic library use. Specially, For Science Citation Index Expanded there were several minor negative correlations revealed in the analysis.

Summary

Data mining technology is involved with analyzing large sets of data to acquire patterns. Usually, the process demands some cycles of analysis, with each cycle obtaining answers, as same as more puzzles. The study managed heretofore at the university libraries is only the first round of analysis. The initial questions were answered, but as expected, more puzzles arose.

The question of circulation patterns among the different subjects leaving an area for further study. Other academies can benefit from an analog research as a means to improve user services based on Hohai University libraries data patterns. Supervising this study at several universities can reveal similarities that can be generalized to other academies.

Eventually, the relationship between the number of published papers and academic library use is uniformly positive correlations. Academic library use matters more in some subjects. Books also are required in some subjects more than others. Excess classification of the data into particular majors will be necessary to disclose a truer relationship between the number of published papers and academic library use. Additionally, many other possible opportunities different than just achievement library use could conduce to the number of published papers. Consequently, non library support activities and services, such as participation in student organizations and use of tutoring, advising or counseling services, should also be analyzed to identify if they show negative or positive relationships and how those relationships compare to that of academic library utilization.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (No. 51079040/E090101) and Hohai university library '2015' key research projects (No.20150323).

References

[1] T. Herawan, H. Chiroma, P. Vitasari, Z. Abdullah, M. Ismail, and M. Othman: Quality & Quantity Vol. 49-6 (2015), p. 2527-2547.

[2] X.W. Li, H.W. Tan, and A. Rackes: Journal Of Cleaner Production Vol. 106 (2015), p. 97-108.

[3] M. Shoaib, A. Daud, and M.S.H. Khiyal: Arabian Journal for Science and Engineering Vol. 40-6 (2015), p. 1591-1605.

[4] X.L. Sun, J. Kaur, L. Possamai, and F. Menczer: Information Processing & Management Vol. 49-2 (2015), p. 454-464.

[5] D. Espinoza, M. Goycoolea, E. Moreno, and A. Newman: Annals of Operations Research Vol. 206-1 (2013), p. 93-114

[6] P. Vilar, and A. Sauperl: International Journal of Information Management, Vol. 35-5(2015) 551-560.

[7] H.M. Osborne and A. Cox: Program-Electronic Library and Information Systems Vol. 49-1 (2015), p. 23-45