Application and Empirical Research of Data Mining Technology in Decision Optimization of Chinese Publishing Enterprise

Shuo Liu
School of Economics and Management
Beijing Institute of Graphic Communication, BIGC
Beijing, China
liushuo621@163.com

***Abstract*—As the age of big data is coming and the market is becoming increasingly competitive, it's important for enterprise to discover the knowledge and rules existing in big data for decision optimization. Publishing schedule and association of publications have become two important factors which must be considered during the publishing decision making process. This paper firstly analyzes the feasibility of data mining used in Chinese publishing industry. Then, a scientific decision making process based on data mining is established. Finally, we collect a mass of book sales data to show the practical application of data mining in publishing decision support.**

Keywords-data mining; OLAP; association analysis; decision support

I. INTRODUCTION

The press and publication industry of China has been developing rapidly in resent year; high-techs such as internet and digital technology give impetus to publishing industrial transformation and upgrading. Almost all of the publishing enterprises have completed the management informationization and e-commerce nowadays. So, huge amounts of data including management, publishing and book sales is emerging and saved, which means that the age of big data for publishing industry has come. How to use the mass data and dig the potential value hiding in the mass data has become an urgent issue for the publishing enterprises in China. This paper firstly analyzes the feasibility of data mining technology used in publishing enterprises of China; and then proposes two important factors that "publishing schedule" and "association of publications" should be considered during publishing decision making process. A scientific decision making process based on data mining technology is established afterward. Finally, we collect a mass of book sales data to show the practical application of data mining technology including OLAP and association analysis in publishing decision support.

II. THE FEASIBILITY OF DATA MINING USED IN PUBLISHING INDUSTRY

Data mining is a process of knowledge discovery based on mass data involving various algorithms, computer science and information technology. [1] Online Analytical Processing (OLAP) is a kind of data analysis technique which is similar but different from data mining. OLAP enables the data analyst and manager to analyze mass data in multiple dimensions in order to optimize the decision making. [2] Although data mining and OLAP are different, their goals are both for decision support, so they are always used for decision support simultaneously.

By the definition of data mining and OLAP, we can infer that there are two necessary conditions for the application of data mining and OLAP, one is mass data the other is application software of data mining algorithm.

Management information system and e-commerce have enhanced the information level of publishing industry in China. Business management informationization and network marketing are emerging and saving huge amounts of data moment by moment in publishing enterprises of China; especially the professional e-commerce platforms such as Amazon save and manage the mass data of publication sales. The mass data of business management and sales provides data source of data mining and OLAP. The publishing enterprises in China have met the requirement of mass data for data mining and OLAP.

Since the enormous data and complex algorithm of data mining and OLAP, it's impossible to complete the data organization and analysis by person, so the application software of data mining and OLAP is necessary. Because data warehouse is designed for data analysis, so data warehouse is the best data source of data mining and OLAP actually. But building a data warehouse always requires heavy investment; the heavy investment is unacceptable for most publishing enterprises in China now. Fortunately, most database system being used in enterprise such as Access, SQL Server even Excel can establish the data set for analysis, and office software such as Excel and free data mining software such as Weka can complete the process of OLAP and data mining. So, publishing enterprises in China have met the requirement of application software for data mining and OLAP.

In conclusion, it's feasible for Chinese publishing enterprises to use data mining technology for decision support and decision optimization.

III. THE SCIENTIFIC DECISION MAKING PROCESS BASED ON DATA MINING

Based on the important role of data mining, the traditional decision making process of publishing should be improved. The scientific publishing decision making process based on data mining and OLAP will include the following steps. Firstly, the decision maker should define the problem according to the decision- making demand; and then extract data from inside and outside the company. The second step is data selection, preprocessing and transformation, after that the data set for data mining and OLAP will be generated. The third step is to select appropriate methods to carry out the data mining or OLAP, and then we may acquire knowledge or rules hiding in the data. The finally step is to make decision based on the knowledge or rules acquired in last step. If the knowledge and rules can meet the decision-making demand, the process will end, if not, the process will return to the first step to redefine the problem and reopen the process. The publishing decision making process based on data mining and OLAP is shown as Figure 1.

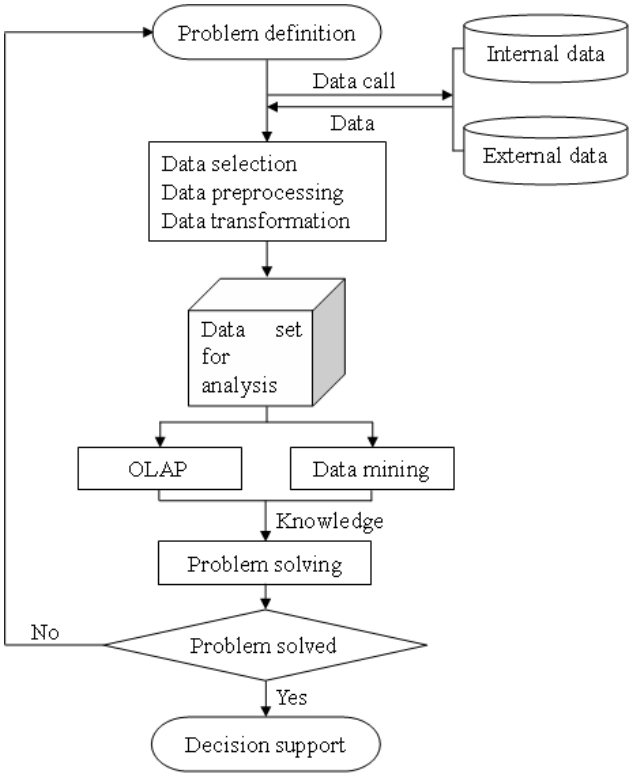

Figure 1. The publishing decision making process based on data mining and OLAP.

IV. EMPIRICAL RESEARCH

In this part, we collect a mass of sales data of a book firstly, and then analyze the data from multiple dimensions based on OLAP in order to decide the publishing schedule of its second edition or similar books. Finally, we use the results of association analysis analyzed by professional e-commerce platform to decide the "publishing combination" and optimize the traditional publishing decision.

A. The Data of EmpiricalRresearch

We optionally select a bestseller titled "C Primer Plus (5th edition)" to be the research object. This book was published by POSTS & TELECOM PRESS in 2005. We collect the sales data of the book from 2013/11/1 to 2014/3/31 including 240 sales records of 151 days. Part of the data after preprocessing is shown in Table 1 and all the sales data was collected from the professional e-commerce platform "dangdang.tmall.com".

TABLE I. THE BOOK SALES DATA FOR OLAP

No.	Sales records			
	Book	***Price***	***Sales***	***Time***
1	C Primer Plus (5th edition)	42.4	1	2013/11/1 7:35
2	C Primer Plus (5th edition)	42.4	1	2013/11/1 9:25
3	C Primer Plus (5th edition)	42.4	1	2013/11/1 16:26
4	C Primer Plus (5th edition)	42.4	1	2013/11/1 19:26
5	C Primer Plus (5th edition)	42.4	1	2013/11/2 11:23
…	…	…	…	…
236	C Primer Plus (5th edition)	42.4	1	2014/3/20 21:19
237	C Primer Plus (5th edition)	42.4	1	2014/3/24 12:30
238	C Primer Plus (5th edition)	42.4	1	2014/3/24 18:46
239	C Primer Plus (5th edition)	42.4	1	2014/3/25 17:06
240	C Primer Plus (5th edition)	42.4	1	2014/3/30 21:33

a. We collected the sales data from "http://detail.tmall.com/item.htm?spm=a1z10.3.w4011-3223502060.20.ZU3Xsl&id=17570263964&rn=75964382be9379515362f0faa958148f"

B. OLAP for Multidimensional Data Analysis

The purpose of this OLAP research is to find the best publishing schedule for the books which are similar to "C Primer Plus (5th edition)" or the new edition of "C Primer Plus (5th edition)". Based on the purpose, we will analyze the book sales data "Sales" from "Book" and "Time" dimension.

We will process the OLAP through pivot table provided by Excel. The each item of sales data in Table 1 is accurate to second, and it's too accurate to make decision based on these original data. So we will roll up the data to observe the sale rules of different time periods.

Firstly, we roll up the data to observe the rules of each month. The OLAP result of "Sales" accurate to month is shown as Table 2 and Figure 2. From Table 2 and Figure 2, we obviously observe that the best sale performance of book "C Primer Plus (5th edition)" happened in November, the sales volume was 107.

TABLE II. THE RULES OF BOOK SALES IN EACH MONTH ANALYZED BY OLAP

No.	Rules of book sales in different months		
	Book	***Time***	***Sales***
1	C Primer Plus (5th edition)	Nov.	107
2	C Primer Plus (5th edition)	Dec.	63
3	C Primer Plus (5th edition)	Jan.	22
4	C Primer Plus (5th edition)	Feb.	41
5	C Primer Plus (5th edition)	Mar.	31

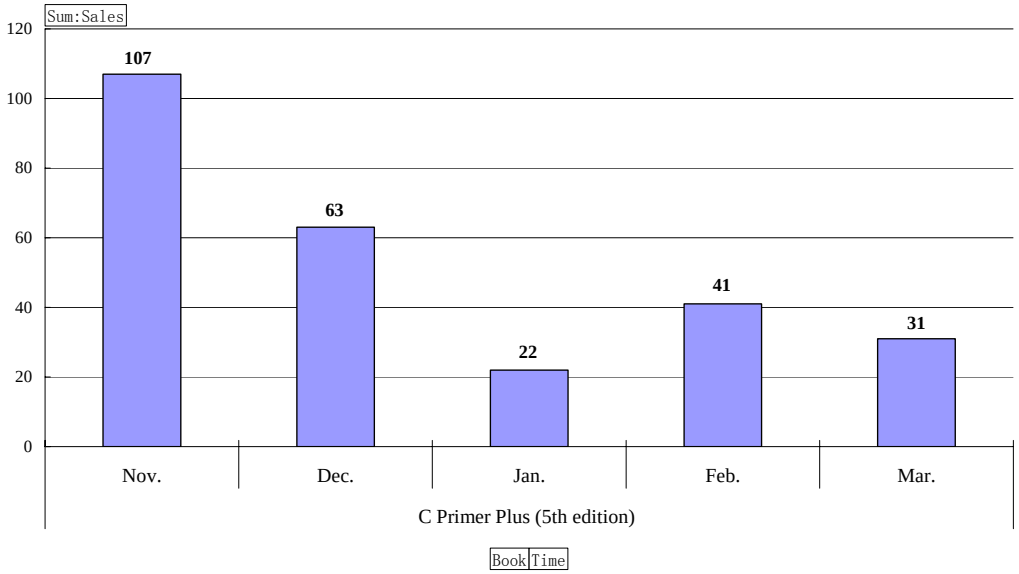

Figure 2. The rules of book sales in each month analyzed by OLAP.

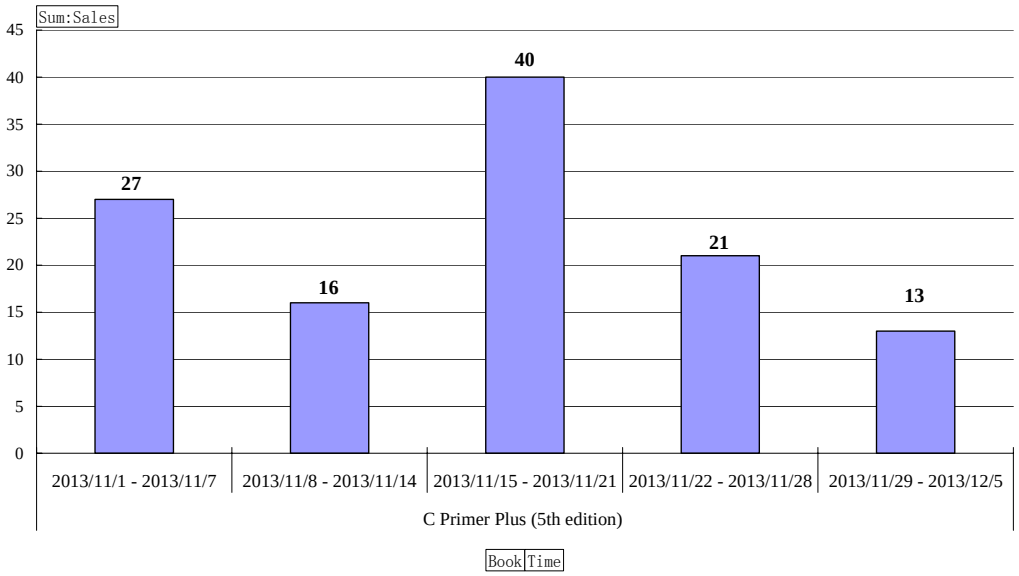

Figure 3. The rules of book sales in each weak of the best sales performence month Novenber analyzed by OLAP.

Further, we drill down the sales data in Nov. to see which week in this month had the best sales performance. The OLAP result of "Sales" accurate to week in Nov. is shown as Table 3 and Figure 3. From Table 3 and Figure 3, we obviously observe that the best sale performance of book "C Primer Plus (5th edition)" happened in the third week that from 15th to 21st in Nov., the sales volume was 40.

TABLE III. THE RULES OF BOOK SALES IN EACH WEAK OF THE BEST SALES PERFORMENCE MONTH NOVENBER ANALYZED BY OLAP

Week.	**Rules of book sales in each week in Nov.**		
	Book	***Time***	***Sales***
1	C Primer Plus (5th edition)	2013/11/1 - 2013/11/7	27
2	C Primer Plus (5th edition)	2013/11/8 - 2013/11/14	16
3	C Primer Plus (5th edition)	2013/11/15 - 2013/11/21	40
4	C Primer Plus (5th edition)	2013/11/22 - 2013/11/28	21
5	C Primer Plus (5th edition)	2013/11/29 - 2013/12/5	13

According to this result, the publishing enterprise can decide the publishing schedule of the similar books or the new edition during this week. The scientific decision of publishing schedule will facilitate book sales effectively.

C. Association Analysis for Publishing Combination

The purpose of association analysis is to find the "book combination" to publish relevant books simultaneously. The association analysis requires a mass of consumer lists to find the books bought together. Publishing enterprises can carry out association analysis through commercial software such as "Clementine" or free software such as "Weka". No matter which method is used, it will take much time and effort to complete the association analysis. Fortunately, most works of association analysis have been done by the professional e-commerce platforms. The professional e-commerce platform "Tmall" recommends the books always bought together with "C Primer Plus (5th edition)" such as "Witty Algorithms in C", "Coding Faster: Getting More Productive with Microsoft Visual Studio" and so on. Now the association analysis has become regular service in professional e-commerce platform. So, if the publishing enterprise publishes the "book combination" based on association analysis, the sales performance will improve effectively.

V. CONCLUSION

Facing the increasingly competitive market, it's necessary for publishing enterprises to change the traditional way of decision making. As the age of big data is coming, decision making based on data mining will be much more scientific than before. Informatization of press and publishing industry has endowed the publishing enterprises with the technological conditions for data mining and OLAP. The empirical research shows the application of OLAP and association analysis in publishing decision making based on the process proposed in this paper. According to the empirical research results, we propose the best publishing schedule and publishing combination of one book, and the publishing decision based on data mining technology will improve the market performance effectively.

ACKNOWLEDGMENT

This work is supported by the "Institute Level Projects Funded by Beijing Institute of Graphic Communication" (Project No. : E-b-2012-11): "Application Research of Data Mining Technology in Chinese Publishing Industry".

REFERENCES

[1] L. Kai Ji, "Data Warehouse and Data Mining," Peking University Press, 2008.

[2] H. Yu Jie and Zh. Jun Chao, "Practices Tutorial of Data Warehouse and OLAP," Tsinghua University Press, 2008.