Market Basket Analysis of Library Circulation Data

Sally Jo Cunningham, Eibe Frank
Department of Computer Science
University of Waikato
Hamilton, New Zealand
{sallyjo, eibe}@cs.waikato.ac.nz

Abstract: "Market Basket Analysis" algorithms have recently seen widespread use in analyzing consumer purchasing patterns—specifically, in detecting products that are frequently purchased together. We apply the Apriori market basket analysis tool to the task of detecting subject classification categories that co-occur in transaction records of books borrowed from a university library. This information can be useful in directing users to additional portions of the collection that may contain documents relevant to their information need, and in determining a library's physical layout. These results can also provide insight into the degree of "scatter" that the classification scheme induces in a particular collection of documents.

1. Introduction

Barcodes are ubiquitous today; practically every product comes to the consumer with a zebra-striped sticker on the back. The software supporting these barcode based purchasing/ordering systems produces vast amounts of sales data, typically captured in "baskets" (records in which the items purchased by a given consumer at a given time are grouped together). This data was quickly recognized by the business world as having immense potential value in marketing. In particular, commercial organizations are interested in discovering "association rules" that identify patterns of purchases, such that the presence of one item in a basket will imply the presence of one or more additional items. A hypothetical example of such a rule might be that shoppers who purchase toothpaste are also likely to buy bananas on the same trip to the grocery store. This "market basket analysis" (MBA) result can then be used to suggest combinations of products for special promotions or sales, devise a more effective store layout, and give insight into brand loyalty and co-branding.

This paper presents a novel application domain for MBA: modelling library circulation. Over 20,000 checkout transactions were captured from the University of Waikato library. This data was processed by a local implementation of the Apriori MBA algorithm (Agrawal and Srikant, 1994), with the goal of detecting commonalities in the way that patrons borrow books across two or more subject classification categories. In MBA terminology, the circulation records describe baskets (documents checked out at the same time by the same patron), for a set of consumers (library patrons). The association rules induced by Apriori identify basket co-purchase patterns (Library of Congress subject categories that tend to co-occur when several books are borrowed at a time). The induced rules are potentially useful in investigating the degree to which the classification scheme physically distributes documents for this user group, and can be useful in informing shelf layout and in directing users to other, possibly relevant, portions of the library.

Related work

MBA techniques have sparked a great deal of interest from the business world in recent years, and several commercial MBA analysis packages have recently been announced. The Apriori algorithm was the first induction tool for the discovery of association rules in large databases (Agrawal, et al, 1993); modifications have been proposed to improve its efficiency (Agrawal and Srikant, 1994; Srikant and Agrawal, 1996), and a variety of similar algorithms have been proposed that induce more expressive rules—for example, by mining over multiple abstraction levels (Han and Fu, 1995) or be extracting generalized rules (Srikant and Agrawal, 1995).

The bulk of previous work in the analysis of library circulation data has been directed at characterizing reading patterns of a particular set of users (for example, Antebellum New Yorkers in Zboray, 1991); measuring the usage seen by individual documents or sub-collections, to support collection evaluation (eg, Day and Revil, 1995; Eldredge, 1998); or creating predictive models of the total circulation level for a given library (eg, Barr and Sichel, 1991; Naylor and Walsh, 1994). The circulation analysis work most closely related to the present paper examined patterns in browsing across subject classifications, but the methodology involved creating summarizations of physical distances between shelves containing circulating items, rather than inducing patterns of co-occurrence in the classification categories themselves (Losee, 1993). Evaluations of document subject classification schemes have generally been based on subjective or qualitative criteria. Quantitative studies have examined the "scatter" imposed by moving documents from one classification scheme to another (Boyce et al, 1990), or used information theoretic measures to compare the information commonalities between documents placed physically adjacent to one another under a particular scheme (Losee, 1992). This present work is novel in its data mining approach to detecting patterns in library circulation data, and its focus on subject co-occurrences at the user transaction level.

This paper is organized as follows: Section 2 describes the pre-processing required for the circulation data and presents a brief explanation of the Library of Congress (LC) classification scheme; Section 3 gives details of the Apriori algorithm; Section 4 presents and discusses the association rules induced by Apriori from the circulation data; and Section 5 presents our conclusions.

2. Data Pre-processing

Raw circulation data was acquired for approximately 50K (52,518) documents borrowed from the University of Waikato library in the first few months of 1998. Each line of data contained a timestamp, anonymized patron id, document id, and a "call number" (indicating the subject classification code and shelf location). Figure 1 presents four lines of raw data, from February 18, 1998.

```
18-FEB-1999 10:09:39.55 498799 0001004356083 BD#450.D38 $1966
18-FEB-1999 10:09:45.35 498799 0001000982700 BF#698.F746 $1976
18-FEB-1999 10:16:52.37 426113 0001004933931 QA"76.9.A73G63 $1993
18-FEB-1999 10:18:52.81 426713 0001005193725 PL$6465.Z77W55 NO.1
```

Figure 1. Raw circulation data

Timestamp and user id data were used to group multiple lines of data into "baskets" representing multiple items checked out at the same time. The date and id were then discarded.

For MBA, the crucial decision in analyzing data is choosing the level of concept abstraction at which to mine associations. Regularities may be difficult to detect or non-existent at the lowest level (here, the individual book id) because the co-occurrence matrix between items is too sparse. The common data preparation technique is to represent individual transaction items at a more abstract level, such that interesting associations begin to appear. In this application, the lowest abstraction level—the document id—was clearly too primitive a concept to yield interesting results, as individual books in the library have very low circulation rates (most are checked out only a few times a year, if they are borrowed at all). The concept level that provided the most acceptable degree of support for induction without over-generalization was the second level of the LC classification category, as indicated in the document call number.

The LC classification scheme organizes documents hierarchically into 21 categories of knowledge (labelled A-H, J-N, P-V, W, and Z). These broad classifications are further divided into narrower subclasses by adding one or two additional letters, and then finally

assigned a numeric classification range that most precisely characterizes the content and coverage of the document. Figure 2 presents a portion of the Science ("Q") portion of the LC scheme. Additional "Cutter" extensions are added by libraries to the LC classification (letters and numbers appearing after the period in the call number); these cutter numbers orthogonally to the hierarchical LC system, and are primarily used to create a linear bookshelf ordering for documents that share an LC classification.

Q	
1-390	Science (General)
1-295	General
300-390	Cybernetics
350-390	Information theory
QA	
1-939	Mathematics
1-43	General
47-59	Tables
71-90	Instruments and machines
75-76.95	Calculating machines
75.5-76.95	Electronic computers. Computer science

Figure 2. Portion of the "Science" LC classification

After data pre-processing, then, the data in Figure 1 was transformed to:

BD, BF
QA, PL

The original 52K lines of raw data were reduced to approximately 20K "baskets"; of these, 4308 included references to more than one subject heading.

3. The Apriori Algorithm

The following description of the Apriori algorithm is drawn from (Agrawal et al, 1993; Chen et al, 1996):

Let $I=\{i_1, i_2, \ldots, i_m\}$ denote the set of items, represented as literals. As noted in Section 2, the granularity of the data representation may vary across mining applications, so that an item might represent a very specific object (for example a 300mll container or Brand X cream), or a more general object from the domain's concept hierarchy (for example, a dairy product). Let D be a set of transactions ("baskets"), where each transaction T is a set of items such that $T \subseteq I$. The number of items of each type that were purchased in a given transaction is not included in this representation, and is not used to by the Apriori algorithm; instead, the presence of an item literal in a transaction indicates simply that one or more of those items were purchased.

Let X be a set of items. A transaction T is said to contain X if and only if $X \subseteq T$. An association rule is defined as an implication of the form $X \Rightarrow Y$, where $X \subset Y$, $Y \subset I$, and $X \cap Y = \varnothing$. The goal is not to induce all possible association rules over a given set D of transactions, but only those rules whose *support* and *confidence* exceed user-supplied thresholds. A rule $X \Rightarrow Y$ is said to have support s over the transaction set D if s% of the transactions in D contain $X \cup Y$, and to have confidence c if c% of the transactions in D that contain X also contain Y.

The Apriori algorithm decomposes the process of inducing association rules into two sub-problems:

1. Find all sets of items that have transaction support above the minimum threshold s. These are termed *large itemsets*.

2. Use the large itemsets to generate association rules. For every large itemset l, all non-empty subsets of l are listed; each such subset a is represented by a rule of the form $a \Rightarrow (l - a)$ if the ratio of support(l) to support(a) exceeds the minimum confidence threshold.

Apriori iteratively discovers large itemsets, considering first itemsets that will generate a single item in the association rule lhs, then rules with two items in the lhs, and so forth. In the second and later iterations, the pass begins with a seed set of candidate itemsets that were found to be large in the previous

pass. During each iteration these seeds are used to generate new potentially large itemsets; at the end of the Apriori determines which are indeed large, and these itemsets form the seed for the next pass. The algorithm iterates until no new large itemsets are identified.

A number of modifications to the Apriori algorithm have been proposed, primarily to improve the efficiency of the rule generation process (eg, Agrawal and Srikant, 1994). The Apriori implementation used to analyze library circulation data is based on the original Apriori algorithm (Agrawal et al, 1993). The size of the dataset did not warrant the implementation of efficiency enhancements.

4. Induction Results

Association rules are output by Apriori in the following form:

<premise> <no. of transactions covered by premise>
⇒
<conclusion> <no. of transactions covered by premise and conclusion>
<(confidence value for rule)>

The association rules induced by Apriori from the library circulation data are presented in Figure 3. For clarity's sake, semantically redundant rules were removed from the output set (that is, only one rule was retained from the pair $X \Rightarrow Y$ and $Y \Rightarrow X$). Note that the association rules are not transitive; the presence of rules $X \Rightarrow Y$ and $Y \Rightarrow Z$ does not necessarily imply that $X \Rightarrow Z$ will attain sufficient confidence to exceed the threshold.

Minimum support: 0.01
Minimum confidence: 0.01

1. HF 565 ⇒ HD 265 (0.47)
2. RC 241 ⇒ TMC 109 (0.45)
3. HC 314 ⇒ HD 141 (0.45)
4. LA 100 ⇒ LB 44 (0.44)
5. HG 127 ⇒ HF 51 (0.4)
6. HG 127 ⇒ HD 51 (0.4)
7. LC 220 ⇒ LB 84 (0.38)
8. HB 209 ⇒ HD 78 (0.37)
9. PN 197 ⇒ PR 55 (0.28)
10. HV 215 ⇒ HQ 60 (0.28)
11. HB 209 ⇒ HC 55 (0.26)
12. GN 285 ⇒ DU 67 (0.24)
13. HC 314 ⇒ HF 71 (0.23)
14. KUQ 238 ⇒ HD 53 (0.22)
15. HM 338 ⇒ HD 71 (0.21)
16. BF 319 ⇒ HM 67 (0.21)
17. BF 319 ⇒ LB 54 (0.17)
18. BF 319 ⇒ HQ 50 (0.16)
19. HM 338 ⇒ HQ 45 (0.13)
20. HM 338 ⇒ HF 44 (0.13)
21. HQ 374 ⇒ HD 45 (0.12)

Figure 3. Apriori output

Three subject headings appearing in Figure 3 are non-standard: RC, here denoting a "reading collection" document in the Education sub-library, rather the LC category for internal medicine; TMC, indicating that the document is part of the teaching materials collection in the Education sub-library; and KUQ, an extended LC heading for New Zealand law documents.

Discussion

All of the rules in Figure 3 have a very simple structure, containing only a single subject heading in the premise and in the conclusion. Library "baskets" tend to be much smaller and simpler than baskets reported for other domains, such as grocery stores—in this dataset, the average number of items borrowed at a time was about two—and consequently the potential support for more complex association rules was lacking.

Library classification schemes are designed to bring similar documents together, to group documents in the hope that this physical co-

location by subject will facilitate browsing for patrons. A successful classification scheme will minimize the amount of time spent browsing for or retrieving documents by reducing the number of "stops" that a patron must make in the library stacks. A scheme that is ill adapted to a given collection will "scatter" like documents across subjects (Losee, 1993).

Note that for 15 of the 21 of the rules generated, the subjects in both the premise and conclusion fall under the same top-level (single letter) LC category. At the highest level, then, for the majority of patterns patrons are finding the documents that they borrow in classifications that are semantically relatively "close" to each other. The six exceptions (rules 2, 12, 14, 16, 17, 18) indicate interesting interdisciplinary interests among the library patrons—for example, rule 12 links anthropology (GN) with Oceania (DU), a combination that is not unexpected given that anthropology and Maori/Pacific development courses at this university tend to highlight of the local region. These instances of subject scatter are likely to be idiosyncratic to this particular university and user base, rather than indicating a flaw in the classification scheme itself.

From the viewpoint of physical library layout, the shelving plan groups the browsing points together well: the association rules indicate that patrons are travelling relatively short distances between the spots from which they retrieve books. Only two rules (14, 17) include classifications that are housed on different floors of the library, and only two rules (16, 20) include classifications that are distanced more than half the length of a single floor. The induced rules do not indicate any major deficiencies in the shelving arrangements on or across floors.

4. Conclusions

We have demonstrated that MBA techniques can be applied to the analysis of library circulation data—specifically, to detect subject classification headings that co-occur in circulation transactions. The induced patterns provide insight into the degree of physical and conceptual scatter imposed on patrons when browsing or retrieving documents.

One application for these results is the construction of browsing hints for patrons—perhaps in the form of notices attached to the shelves—informing patrons of the other subject headings that are associationally linked to the heading that they have located. This type of user- and usage-based feedback can allow a library to adaptively support the information seeking/browsing behavior of its patrons at low cost, and without making major changes in the cataloging or physical arrangement of the collection.

Of course, the association rules can only indicate borrowing patterns that are present in the data—that are either supported by classification proximity, or that patrons exhibit despite any classification or library layout deficiencies. This technique does not supplant previous circulation analysis methods, but rather provides additional insight into the issues that influence collection utlization.

References

Agrawal, R., Imielinski, T., and Swami, A. (1993) "Mining association rules between sets of items in large databases", Proceedings of the ACM SIGMOD Conference on Management of Data, Washington, D.C., pp. 207-216.

Agrawal, R., and Srikant, R. (1994) "Fast algorithm for mining association rules", Proceedings of the 20th VLDB Conference, Santiago (Chile), pp. 478-479.

Barr, A., and Sichel, H.S. (1991) "A bivariate model to predict library circulation", Journal of the American Society for Information Science 42(8), pp. 546-553.

Boyce, R., Douglass, J.S., and Rabalais, J. (1990) "Measurement of subject scatter in the Superintendent of Documents classification", Government Publications Review 17(4), pp. 333-339.

Chen, Ming-Syan, Han, Jiawei, and Yu, Philip S. (1996) Data Mining: An Overview from a Database Perspective. IEEE Transactions on Knowledge and Data Engineering 8(6), pp. 866-883.

Day, M., and Revill, D. (1995) "Towards the active collection: the use of circulation analyses in collection evaluation", Journal of Librarianship and Information Science 27(3), pp. 149-157.

Eldredge, J.D. (1998) "The vital few meet the trivial many: unexpected use patterns in a monographs collection", Bulletin of the Medical Library Association 86(4), pp. 496-503.

Han, J., and Fu, Y. (1995) "Discovery of multiple-level association rules from large databases", Proceedings of the 21st International Conference on Very Large Data Bases, pp. 420-431.

Losee, R.M. (1992) "A Gray code based ordering for documents on shelves: classification for browsing and retrieval", Journal of the American Society for Information Science 43, pp. 312-333.

Losee, R.M. (1993) "The relative shelf location of circulated books: a study of classification, users, and browsing", Library Resources & Technical Services 37, pp. 197-209.

Naylor, M., and Walsh, K. (1994) "A time-series model for academic library data using intervention analysis", Library and Information Science Research 16, pp. 299-314.

Srikant, R., and Agrawal, R. (1995) "Mining generalized association rules", Proceedings of the 21st International Conference on Very Large Data Bases, pp. 407-419.

Zboray, R.J. (1991) "Reading patterns in Antebellum America: evidence in the charge records of the New York Society Library", Libraries & Culture 26, pp. 301-333.