Material Acquisitions Using Discovery Informatics Approach

Chien-Hsing Wu
National University of Kaohsiung, Taiwan, ROC

Tzai-Zang Lee
National Cheng Kung University, Taiwan, ROC

INTRODUCTION

Material acquisition is a time-consuming but important task for a library, because the quality of a library is not in the number of materials that are available but in the number of materials that are actually utilized. Its goal is to predict the users' needs for information with respect to the materials that will most likely be used (Bloss, 1995). Discovery informatics using the technology of knowledge discovery in databases can be used to investigate in-depth how the acquired materials are being used and, in consequence, can be a predictive sign of information needs.

BACKGROUND

Material searchers regularly spend large amounts of time to acquire resources for enormous numbers of library users. Therefore, something significant should be relied on to produce the acquisition recommendation list for the limited budget (Whitmire, 2002). Major resources for material acquisitions are, in general, the personal collections of the librarians and recommendations by users, departments, and vendors (Stevens, 1999). The collections provided by these collectors are usually determined by their individual preferences, rather than by a global view, and thus may not be adequate for the material acquisitions to rely on. Information in the usage data may show something different from the collectors' recommendations (Hamaker, 1995). For example, knowing which materials were most utilized by the patrons would be highly useful for material acquisitions.

First, circulation statistics is one of the most significant references for library material acquisition decisions (Budd & Adams, 1989; Tuten & Lones, 1995; Pu, Lin, Chien, & Juan, 1999). It is a reliable factor by which to evaluate the success of material utilization (Wise & Perushek, 2000). Second, the data-mining technique with a capability of description and prediction can explore patterns in databases that are meaningful, interpretable, and decision supportable (Wang, 2003). The discovery informatics in circulation databases using induction mechanism are used in the decision of library acquisition budget allocation (Kao, Chang, & Lin, 2003; Wu, 2003). The circulation database is one of the important knowledge assets for library managerial decisions. For example, information such as "75.4% of patrons who made use of Organizations also made use of Financial Economics" via association relationship discovery is supportive for the material acquisition operation. Consequently, the data-mining technique with an association mechanism can be utilized to explore informatics that are useful.

MAIN THRUST

Utilization discovery as a base of material acquisitions, comprising a combination of association utilization and statistics utilization, is discussed in this article. The association utilization is derived by a data-mining technique. Systemically, when data mining is applied in the field of material acquisitions, it follows five stages: collecting datasets, preprocessing collected datasets, mining preprocessed datasets, gathering discovery informatics, interpreting and implementing discovered informatics, and evaluating discovered informatics. The statistics utilization is simply the sum of numeric values of strength for all different types of categories in preprocessed circulation data tables (Kao et al., 2003). These need both domain experts and data miners to accomplish the tasks successfully.

Collecting Datasets

Most libraries have employed computer information systems to collect circulation data that mainly includes users identifier, name, address, and department for a user; identifier, material category code, name, author, publisher, and publication date for a material; and users

identifier, material identifier, material category code, date borrowed, and date returned for a transaction. In order to consider the importance of a material category, a data table must be created to define the degree of importance that a material presents to a department (or a group of users). For example, five scales of degree can be "absolutely matching," "highly matching," "matching," "likely matching," and "absolutely not matching," and their importance strength can be defined as 0.4, 0.3, 0.2, 0.1, and 0.0, respectively (Kao et al., 2003).

Preprocessing Data

Preprocessing data may have operations of refinement and reconstruction of data tables, consistency of multityped data tables, elimination of redundant (or unnecessary) attributes, combination of highly correlated attributes, and discretization of continuous attributes. Two operations in this stage for material acquisitions are the elimination of unnecessary attributes and the reconstruction of data tables. For the elimination of unnecessary attributes, four data tables are preprocessed to derive the material utilization. They are users tables (two attributes: department identifier and user identifier), category tables (two attributes: material identifiers and material category code), circulation table (three attributes: user identifier, material category code, and date borrowed), and importance table (three attributes: department identifier, material category identifier, and importance). For the reconstruction of data tables, a new table can be generated that contains attributes of department identifier, user identifier, material category code, strength, and date borrowed.

Mining Data

Mining mechanisms can perform knowledge discovery with the form of association, classification, regression, clustering, and summarization/generalization (Hirota & Pedrycz, 1999). The *association* with a form of "If Condition Then Conclusion" captures relationships between variables. The *classification* is to categorize a set of data based on their values of the defined attributes. The *regression* is to derive a prediction model by altering the independent variables for dependent one(s) in a defined database. *Clustering* is to put together the physical or abstract objects into a class based on similar characteristics. The *summarization/generalization* is to abridge the general characteristics over a set of defined attributes in a database.

The association informatics can be employed in material acquisitions. Like a rule, it takes the form of P=>Q (α, β), where P and Q are material categories, and α andβ are support and confidence, respectively (Meo, Pseila, & Ceri, 1998). The P is regarded as the condition, and Q as the conclusion, meaning that P can produce Q implicitly. For example, an association rule "Systems =>Organizations & Management (0.25, 0.33)" means, "If materials in the category of Systems were borrowed in a transaction, materials in Organization & Management were also borrowed in the same transaction with a support of 0.25 and a confidence of 0.33." *Support* is defined as the ratio of the number of transactions observed to the total number of transactions, whereas *confidence* is the ratio of the number of transactions to the number of conditions. Although association rules having the form of P=>Q (α, β) can be generated in a transaction, the inverse association rules and single material category in a transaction also need to be considered.

When two categories (C1 and C2) are utilized in a transaction, it is difficult to determine the association among C1=>C2, C2=>C1, and both. A suggestion from librarians is to take the third one (both) as the decision of this problem (Wu, Lee, & Kao, 2004). This is also supported by the study of Meo et al. (1998), which deals with association rule generation in customer purchasing transactions. The number of support and confidence of C1=>C2 may be different from those of C2=>C1. As a result, the inverse rules are considered as an extension for the transactions that contain more than two categories to determine the number of association rules. The number of rules can be determined via 2*[n*(n-1)/2], where n is the number of categories in a transaction. For example, {C1, C2, C3} are the categories of a transaction, and 6 association rules are then produced to be {C1=>C2, C1=>C3, C2=>C3, C2=>C1, C3=>C1, C3=>C2}. Unreliable association rules may occur because their supports and confidences are too small. Normally, there is a predefined threshold that defines the value of support and confidence to filter the unreliable association rules. Only when the support and confidence of a rule satisfy the defined threshold is the rule regarded as a reliable rule. However, no evidence exists so far is reliable determining the threshold. It mostly depends on how reliable the management would like the discovered rules to be. For a single category in a transaction, only the condition part without support and confidence is considered, because of the computation of support and confidence for other transactions.

Another problem is the redundant rules in a transaction. It is realized that an association rule is to reveal the company of a certain kind of material category, independent of the number of its occurrences. Therefore, all redundant rules are eliminated. In other words, there is only one rule for a particular condition and only one conclusion in a transaction. Also, the importance of a material to a