Enriching user-oriented class associations for library classification schemes

Hsiao-Tieh Pu and
Chyan Yang

The authors

Hsiao-Tieh Pu is an Associate Professor at the Shih Hsin University, Taipei, Taiwan.
Chyan Yang is a Professor at the National Chiao Tung University, Hsinchu, Taiwan.

Keywords

Hierarchy, Data, Library users, User studies, Web site classification

Abstract

This paper explores the possibility of adding user-oriented class associations to hierarchical library classification schemes. Some highly associated classes not grouped in the same subject hierarchies, yet relevant to users' knowledge, are automatically obtained by analyzing a two-year log of book circulation records from a university library in Taiwan. The library uses the Chinese Decimal Classification scheme, which has similar structure and notation to the Dewey Decimal Classification. Methods, from both collaborative filtering and information retrieval research, were employed and their performance compared based on similarity estimation of classes. The results show that classification schemes can, therefore, be made more adaptable to changes of users and the uses of different library collections by analyzing the circulation patterns of similar users. Limitations of the methods and implications for applications are also discussed.

Electronic access

The Emerald Research Register for this journal is available at
http://www.emeraldinsight.com/researchregister

The current issue and full text archive of this journal is available at
http://www.emeraldinsight.com/0264-0473.htm

The Electronic Library
Volume 21 · Number 2 · 2003 · pp. 130-141
© MCB UP Limited · ISSN 0264-0473
DOI 10.1108/02640470310470507

Introduction

Hierarchical library classification schemes have been used for decades to organize bibliographic resources (Chan, 1995). The major schemes include Dewey Decimal Classification (DDC) in the USA and the Chinese Decimal Classification (CDC) in the Chinese communities. These schemes are organized based on disciplines, followed by a hierarchical arrangement of topics. Although library items have mostly been classified using these schemes, the purpose has been to help people locate items on shelves. In fact, with the growth of electronic environments, schemes are now being exploited in new ways for information organization and retrieval. Recently, classification schemes have been adapted to organize Web resources (Hickey, 2000) and also to enhance the design of a view-based Online Public Access Catalog (OPAC) interface for the Web (Tinker *et al.*, 1999). In these Web applications, a high degree of class integrity of schemes like DDC has been demonstrated (Thompson *et al.*, 1997). The subject classes in these schemes are well defined and distinct. However, they do not take user understanding into consideration.

From the user point of view, some highly related yet non-hierarchical classes, i.e. classes not grouped in the same subject hierarchies, may not be easy to perceive in these schemes. The discovery of hidden associations between classes is important in many applications, such as constructing or updating relative indexes of schemes and enhancing OPAC subject browsing. Associations between classes 170 (psychology) and 527 (school management), and between classes 170 and 548 (social pathology) in the CDC scheme are examples shown in Figure 1. CDC has similar structure and notation to DDC, but the disciplines included, and their arrangements, are different in some classes (Lai, 1989). Table I lists ten main classes in CDC and DDC respectively for comparison. For example, CDC divides history and geography into two main classes, i.e. China and the World, and merges literature and language into one main class. When someone searches for psychology books in CDC, class 170 and its neighboring

The work is partially sponsored by NSC90-2413-H-128-001 for the first author, and by NSC90-2416-H-009-012 for the second author.

Figure 1 An example of subject distribution of class 170 (psychology) in the Chinese Decimal Classification

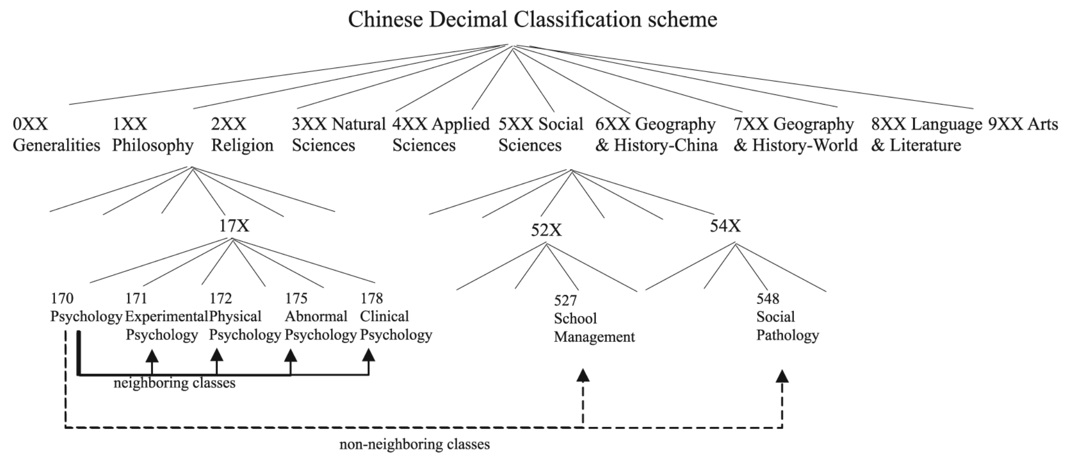


Table I Ten main classes of Dewey Decimal Classification and Chinese Decimal Classification schemes

Class	DDC	CDC
000	Computers, information and general reference	Generalities
100	Philosophy and psychology	Philosophy
200	Religion	Religion
300	Social sciences	Natural sciences
400	Language	Applied sciences
500	Science	Social sciences
600	Technology	History and geography – China
700	Arts and recreation	History and geography – World
800	Literature	Language and literature
900	History and geography	Arts

classes (the classes grouped in the same subject hierarchy) like 171 (experimental psychology) are assumed to be the most appropriate places to find related items (see the solid arrow lines in Figure 1). However, by analyzing users' uses of library collections, additional related classes, which are outside the neighborhood, are found scattered in the 52X (education) and 54X (sociology) hierarchies, such as 527 and 548 (see the dashed arrow lines in Figure 1). As Solomon (1992) pointed out:

> . . . classification schemes fail too often because they are not grounded in the language and knowledge of users or in the task or situation of use. Yet methods for adding user dimension to classification schemes are neither well established nor well tested.

The purpose of this paper is, therefore, to explore the possibility of adding user-oriented associations to hierarchical classification schemes, and to provide a new basis for information organization and retrieval applications.

This paper presents initial experiments on obtaining highly related classes by analyzing users' usage of library collections. A book circulation log was used as the basis for analysis. The experimental data contained a two-year log of approximately 166,000 book circulation records from nearly 8,000 patrons of a university library. The library uses the CDC scheme to organize books in Chinese. Classes in the third-level, i.e. classes 000-999, were the target classes for association analysis. The approach to discovering class associations is based on the similarity estimation methods commonly used in the collaborative filtering (Sarwar *et al.*, 2001) and information retrieval research (Salton, 1989), including conditional-probability-based and cosine-based methods. The major hypothesis behind the approach is that the similarity between two classes can be estimated by the resemblance of users who have borrowed the books in the same classes.

The results obtained show that many neighboring classes are strongly associated in terms of the local users' circulation patterns. More importantly, it is also found that many associated classes scattered across different subject hierarchies could be discovered from the circulation patterns of similar users. Such subject distributions were investigated and organized into various types of associations. Further, the obtained association norms between classes were found to be useful in understanding users' subject preferences for a given class. For example, class 435 (horticulture) was found to be associated with 43 classes. Among them, there were 40 non-hierarchical classes (classes not grouped in the 43X agriculture subject hierarchy). To name a few, classes 375 (botany), 374 (economic botany), and 929 (landscape architecture) were all associated with 435. Through further analysis of their subject contents using the titles and subject headings of the corresponding bibliographic records, some topics of interests could be observed, such as ornamental plants, perfume flowers, and urban beautification, respectively. Also, the subject preferences could be ordered according to the rank of each of their association norms.

With the proposed approach, it is possible to enrich user-oriented class associations to the classification scheme, and the scheme can, therefore, be made more adaptable to changes of users and the uses of different library collections. There are implications for applications in information organization and retrieval. For example, catalogers could refer to the ranked associated classes when they perform multi-classification, and users could also browse the associated classes for related subjects. However, there are some limitations of this study. For example, the small size and sparseness of the circulation log used may have prevented the discovery of more in-depth associations. In future research, more empirical studies will be necessary to validate the findings, and methods for obtaining user-oriented associations can still be improved.

Related work

Recently, much research effort has focused on enhancing library classification schemes to meet the needs of new information environments. For instance, OCLC has been applying DDC to the Internet environment of its NetFirst database and has also enhanced the traditional DDC summaries to create a knowledge base using various vocabulary sources (Mitchell and Vizine-Goetz, 2000). On the other hand, many studies have concentrated on investigating the value of classificatory structures for searching and browsing. In one project, Visual Dewey was applied to the DDC at the interface of a view-based searching OPAC (Pollitt, 1998). The IFLA guidelines for OPAC displays (Yee, 1998) also recommend that a user be able to view the hierarchical and cross-disciplinary context of any classification number for a specified search term.

In order to understand how users view classification schemes, circulation logs can be used to unobtrusively obtain information about users' circulation patterns. The bulk of previous works on the analysis of library circulation logs have focused on collection management with various purposes in mind, for example, to learn more about how the reading patterns of user groups can help libraries focus their services (Bertland, 1991), to assess how particular resources can be used to make decisions about collection development (Eldredge, 1998), and to measure the usage of collections to create predictive models of shelf arrangement (Barr and Sichel, 1991). Other studies have investigated how subject analysis can be used to increase usage of collections (Wilson and Spillane, 2000).

Association discovery may employ various techniques to perform data mining, collaborative filtering, and information retrieval research. Among the available data mining techniques, association analysis has been widely used to analyze consumers' purchasing patterns, specifically, to detect products that are frequently purchased together (Chen *et al.*, 1996). However, very few studies have focused on library applications (Banerjee, 1998). In Cunningham and Frank's (1999) study, they used market basket analysis to detect subject classification categories that co-occurred in a circulation log. The authors noted that the library circulation log used was small and sparse; consequently, potential support for more complex association rules was lacking in their study. Collaborative filtering or recommendation systems apply similar discovery techniques to the problem of making personalized recommendations for information and products (Sarwar *et al.*, 2001). Meanwhile, a variety of similarity estimation techniques have been developed in information retrieval research, such as the term frequency theta inverse document frequency (TFIDF) measure in the vector space model (Salton, 1989).

Research design

Data collection

The data used were from Shih Hsin University (SHU), which is a medium-size, urban multiversity in Taiwan. For the experiment, a raw circulation log was collected that included 166,065 book items (books) borrowed from the SHU library during the period from September 22, 1997 to September 23, 1999 (excluding transactions of English books). At the time of our research, 130,898 books were available for borrowing, and the number of patrons was 10,876. Other necessary information like the user's department code and book's classification number were extracted from the patron and bibliographic files of the SHU library automation system, and then merged in to the experimental data set. Meanwhile, in order to protect individual privacy, each user id was made anonymous as a unique serial number. The experimental set then included a user anonymous id, user department code, and three-digit classification number.

The user department code included at most four levels, i.e. unit, department, section, and year. For example "A0114" indicated undergraduate (A), department of journalism (01), news editing (1), and the fourth year (4). The classification number was based on the CDC scheme as mentioned above. It is divided into ten main classes (hundreds), each main class is further divided into ten divisions (tens), and each division into ten sections (ones), etc. Each class has a minimum length of three digits to indicate its position in the classification hierarchy, and those longer than three have a decimal point after the first three digits. In this study, classes in the third level, 000-999, were used for association analysis.

Data analysis

(1) *Circulation statistics*. The use of similarity estimation techniques usually requires a large group of users who have overlapping circulation patterns and who have interacted with the system for a certain period of time. However, statistics on library use indicate that most books are utilized by very few patrons (Kent and Williams, 1979). The data set shows that only 33.1 percent of the total available books had been borrowed, among them, 32.1 percent had been borrowed only once, and that on average, each book that had been borrowed had only been borrowed 3.83 times. Of the users, 71.5 percent had borrowed books, but each of them had borrowed only 5.57 books on average. Books from 74.58 percent of the total available 893 classes had been borrowed, and on average, books from each class were borrowed by 110.34 users. The standard deviation 274.02 was rather high. Figure 2 shows the frequency distribution of each class borrowed. Table II lists the top 20 classes with over 1,000 circulation counts, which were considered popular classes. Obviously these popular classes accounted for large portions of usage; for example, classes 857 (fiction) and 312 (computer) accounted for nearly 18 percent of the total usage. This explains the above high standard deviation. On the other hand, each user borrowed books from 9.48 classes on average, and the standard deviation of 8.48 was not significantly different.

Figure 2 Frequency distribution of classes of borrowed books

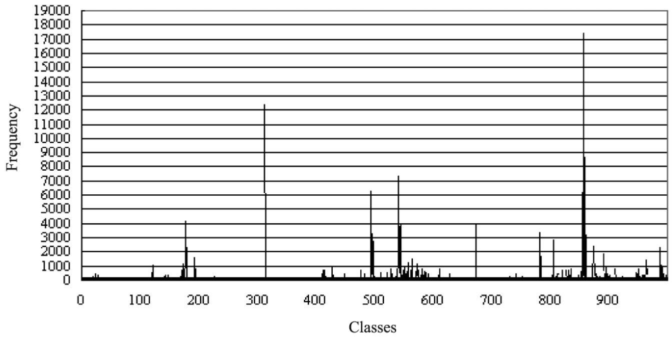


Table II Top 20 classes with over 1,000 circulation counts

CDC number (caption)	Count	Percent	Accumulated
857 (Fiction)	17,368	10.46	10.46
312 (Computer)	12,392	7.46	17.92
541 (Sociology)	7,310	4.40	22.32
494 (Business management)	6,287	3.79	26.11
855 (Short literature essays)	6,212	3.74	29.85
177 (Applied psychology)	4,185	2.52	32.37
544 (Family and kinship)	4,015	2.42	34.79
673 (Taiwan)	3,944	2.37	37.16
782 (Chinese biography)	3,325	2.00	39.16
861 (Japanese literature)	3,261	1.96	41.13
496 (Marketing)	2,969	1.79	42.92
805 (Linguistics – English)	2,875	1.73	44.65
497 (Advertisement)	2,671	1.61	46.25
873 (English literature)	2,420	1.46	47.71
987 (Movies)	2,290	1.38	49.09
874 (American literature)	2,155	1.30	50.39
890 (Journalism)	1,828	1.10	51.49
192 (Personal ethics)	1,609	0.97	52.46
563 (Finance)	1,533	0.92	53.38
540 (Sociology)	1,493	0.90	54.28

(2) *Data sparseness analysis.* In order to perform similarity computation of classes, the data set was converted into a user-class frequency matrix *MI*, which is an $m \times n$ matrix as shown in Figure 3. Assume that there is a list of m users $U = \{u_1, u_2, \ldots, u_m\}$ and a list of n classes $C = \{\vec{c}_1, \vec{c}_2, \ldots, \vec{c}_n\}$, and that each user u_a has a list of classes C_{u_a}, which are the corresponding classes of the borrowed books. Each entry $k_{a,x}$ in *MI* then represents the borrowing frequency of the ath user from the xth class of borrowed books, and it can be 0, indicating that the user has not borrowed any books from that class. That is, each c_i in *MI* is taken as a vector such that $\vec{c}_i = <k_{1i}, k_{2i}, \ldots, k_{mi}>$. Another matrix *MD* was also created to measure similarities by user department instead of at the individual user level.

In the experiments, the sparseness level of the data set was also considered. In Table III, the columns show the total numbers of active users (users who had borrowed books), active classes (classes from which books had been borrowed), non-zero entries, and the sparseness level, respectively. The sparseness level was defined as:

$$1 - \frac{no\_non\_zero\_entries}{no\_active\_users \times no\_active\_classes}.$$

Figure 3 An abstract diagram showing the co-borrowed classes and similarity computation

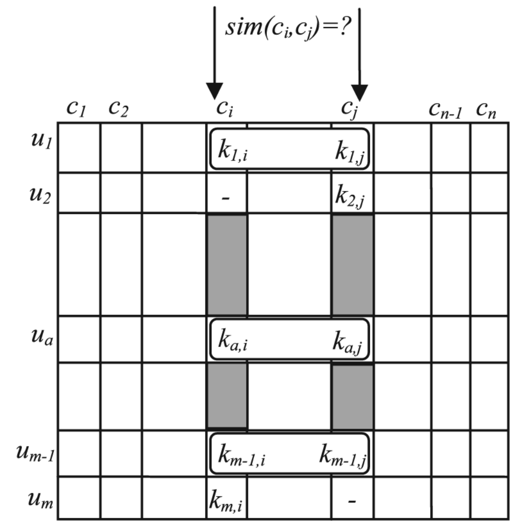


Table III Data sets with different sparseness levels

Data set	No. of active users	No. of active classes	Non-zero entries	Sparseness level
***MI* (individual user)**	7,779	666	73,697	0.98578
***MD* (user department)**	58	666	8,009	0.79266

Association discovery methods

The premise behind the methods used for discovering associations is similar to the concept derived from collaborative filtering, which uses community opinion and behavior to determine the value of information and identify important trends (Paepcke *et al.*, 2000). If there is a large enough sampling, the associations among books or classes can be discovered according to the circulation patterns of similar users. Various similarity-based methods can be used in the initial stage of discovering associations between classes. Here are two types of methods, i.e. conditional-probability-based and cosine-based methods.

(1) *Conditional-probability-based similarity.* In this case, the similarity between two classes is based on the conditional probability of borrowing one of the classes given that the other class has already been borrowed. In particular, the conditional probability of borrowing c_j given that c_i has already been borrowed $P(c_j|c_i)$, is the number of users that have borrowed both classes c_j and c_i divided by the total number of users that have borrowed c_i as follows:

$$P(c_j|c_i) = \frac{Freq(c_jc_i)}{Freq)c_i)},$$

where *Freq*(*X*) is the number of users that have borrowed books from the classes in set *X*.

This method has a drawback in that, quite often, $P(c_j|c_i)$ is high as a result of the fact that c_j occurs very frequently and not because c_j and c_i tend to occur together. Also, it provides no mechanism by means of which to discriminate between users who borrow many books and users who borrow few books. Users that borrow fewer books may be more reliable indicators when determining the similarity between books. For the above two reasons, the normalized conditional probability method was used with the following adapted equation (Karypis, 2001):

$$sim(c_i, c_j) = \frac{\sum_{\forall a:k_{c_i}>0} k_{a,c_i}}{Freq(c_i) \times Freq((c_j))^{a}},$$

where $k_{a,x}$ is an entry in MI and α is a parameter that takes a value between 0 and 1. Furthermore, instead of using the co-occurrence frequency, the equation uses the sum of the corresponding non-zero entries of the xth column in the user-class matrix. Since the rows are normalized to be unit length, users that have borrowed more books will tend to contribute less to the overall similarity, thus lending emphasis to the borrowing decisions of the users that have borrowed fewer books.

(2) *Cosine-based similarity.* A variety of similarity estimation techniques, including basic cosine and normalized cosine methods, have been developed in information retrieval research (Salton, 1989). These techniques use the vector form for both documents and queries. In the basic cosine method, two classes are thought of as two independent vectors in the m dimensional user space. The similarity between them is measured by computing the cosine of the angle between these two vectors. Formally, with the $m \times n$ user-class frequency matrix shown in Figure 3, the similarity between classes c_i and c_j, denoted by $sim(c_i, c_j)$, is given by:

$$sim(c_i, c_j) = \cos(\vec{c_i}, \vec{c_j}) = \frac{\vec{c_i} \cdot \vec{c_j}}{|\vec{c_i}| \times |\vec{c_j}|},$$

where $|\vec{c_i}|$ and $|\vec{c_j}|$ are the norms of the two class vectors, and "." denotes their inner product.

From the above equation, it is noted the similarity between two classes will be high if two users that borrowed books from the same number of classes also borrowed books from the other class as well. Besides, the cosine-based method considers the borrowing frequency of every class, and, therefore, offsets the drawback resulted in the class whose books were frequently borrowed. However, the basic cosine measure has one important drawback, i.e. the differences in the scale of borrowing between different users are not taken into account. Therefore, besides testing the basic cosine method, the normalized cosine method was also used (Salton and Buckley, 1988). The normalized cosine method solves the above problem by using the TFIDF weight. Each entry $k_{a,x}$ in MI is replaced by a weight $w_{a,x}$, which is associated with each user u_a of a class c_x, and is calculated as follows:

$$w_{a,x} = \left(0.5 + \frac{0.5\,freq_{a,x}}{\max_l freq_{l,x}}\right) \times \log\frac{N}{n_a}.$$

The value calculated using the formula in brackets represents the normalized frequency of user u_a borrowing books from class c_x. $freq_{a,x}$ is the raw frequency of user u_a borrowing books from class c_x and the maximum l is computed for all users who have borrowed books from class c_x. On the other hand, the log value represents the normalized frequency of the user u_a borrowing books from all classes C, where N is the total number of classes in the library collection and n_a the number of classes whose books user u_a borrowed. Using the weighting scheme, each class pair in C can be ranked based on its cosine value.

Experimental results

Evaluation of the similarity estimation methods

One of the goals in the experiments was to evaluate the quality of the obtained associations using the normalized conditional probability and normalized cosine methods. The test set included randomly selected ten classes with medium frequencies of circulation. Each class in the test set obtained its correct associated classes from three sources, i.e. the neighboring classes in the corresponding hierarchies of the scheme, the non-neighboring classes in the cross-references of the scheme, and the associated classes obtained using the different similarity estimation methods which were manually validated and extracted. The evaluation metric adopted the recall and precision measure. Figures 4 and 5 show the change of the recall and precision rates, respectively, based on the number of top n associated classes obtained using the two different similarity estimation methods.

From Figures 4 and 5, it is easy to see that the normalized cosine method outperformed the normalized conditional probability

Figure 4 The change of the recall rates obtained using the two different methods for class similarity estimation

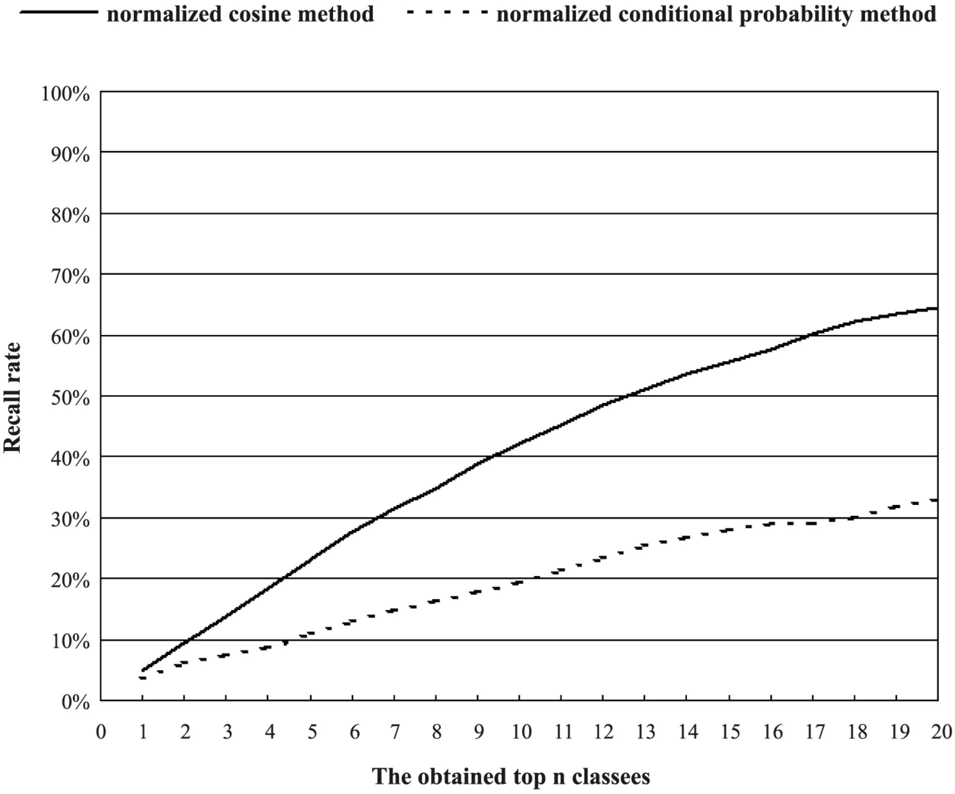
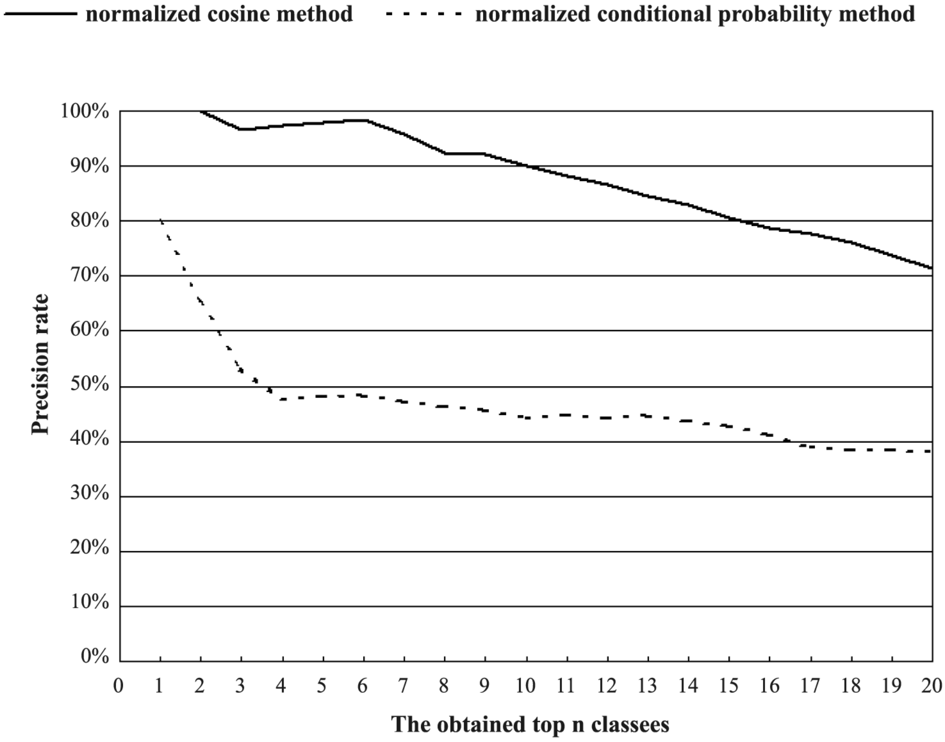


Figure 5 The change of the precision rates obtained using the two different methods for class similarity estimation



method. With the conditional-probability-based method, many noisy high associations arose from some infrequent classes, which easily obtained high associations with some popular classes. These infrequent classes were those classes whose books were borrowed a few times by a few users. On the other hand, the cosine-based method favored classes with large numbers of similar users and was less sensitive to infrequent classes.

However, there were many infrequent classes in the test circulation log, e.g. 70 percent of the classes whose books had been borrowed had circulation counts of less than 100 and 76 percent had fewer than 100 users. This case is especially true in a university library setting and is less frequently found in a commercial environment. In the former situation, there is a need to balance the collection based on factors besides user interests. In the latter case, products rarely purchased by customers will be withdrawn within a short time; hence, most products will accumulate a certain number of transactions more easily.

On the other hand, some experiments were also conducted to observe the effect of different sparseness levels. Large numbers of ambiguous associations were found when using the user's department as the basis for discovering class associations, i.e. 58 departments rather than 7,779 individual users in Table III.

Although the sparseness level is lower if using department information, it helps to find more associations and the noises produced become even more serious. Therefore, considering the above problems, the normalized cosine method was used with individual users as the basis for extracting class associations. In addition, to make the results more accurate, some of the infrequent and popular classes were removed.

Analysis of association types

As mentioned above, the normalized cosine method achieved better performance and is not as sensitive to the problem of data sparseness, so it was chosen to obtain class associations in this study. For the purpose of in-depth observation, a 1,000 ×1,000 class matrix was used to investigate these associations as depicted in Figure 6. It shows how circulation patterns affected the classification hierarchy. The obtained, highly associated, class pairs are plotted to indicate the associations of the corresponding classes (class pairs) with cosine values larger than 0.3. The X-axis and Y-axis both correspond to 1,000 classes. Each class pair may have another reverse pair running parallel to itself. It is clear that many class pairs follow the 45° angle line; i.e. the associated classes based on the users' circulation patterns are somehow identical to the hierarchy of the scheme. However, there are associations scattered across the scheme as well (see the scattered points that are not near the 45° angle line). Bibliographic records of books belonging to these highly associated classes were verified and found that their subject contents were quite related to each other. The quality of the associations discovered using the proposed approach was, therefore, high.

These class pairs were, then, investigated and organized into four different types of associations discussed as follows:

(1) *Hierarchical class associations.* The first type includes classes near the 45° angle line. They are subject classes hierarchically related in CDC; for example, class 023 (library personnel administration) is associated with class 025 (special libraries), and both are arranged under 02X (library science) in the subject hierarchy. These associations provide a basis for observing users' preferences in the classification hierarchy. For example, the experimental results show that class 020 (library and information sciences) is associated with class 023 (library personnel administration) more frequently than with class 028 (reading and use of other information media). In addition, many hierarchically related classes do not appear in the association matrix, including classes 021 (library public relationships), 022 (administration of the physical plant), and others.

(2) *Non-hierarchical class associations.* The second type includes associated classes scattered across the scheme. Such scattering may have resulted in the unavoidable inter-disciplinary associations in the scheme, or may have been due to various needs or interests of the local users. Table IV lists ten examples of interesting class pairs obtained from the association matrix based on cosine values > 0.3.

(3) *Popular class associations.* The third type includes popular classes like 857 (fiction), which are associated with many classes without strong subject relationships. These classes are like stop words in an indexing system, which contribute little discrimination value for identifying differences among documents. Such "stop" classes may have noise effects on the weight of each class and may make

Figure 6 Association matrix of classes at the individual user level based on cosine > 0.3

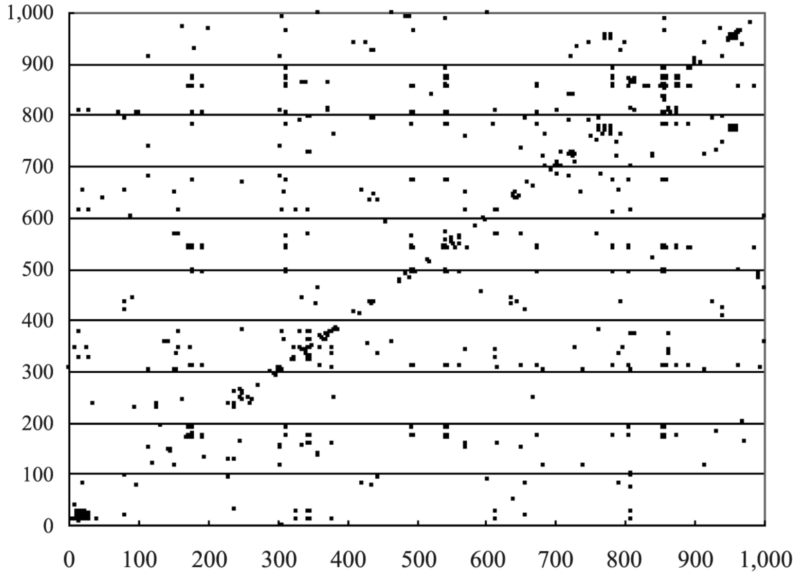


Table IV Examples of non-hierarchical highly associated class pairs

Class number (caption)	Class number (caption)
019 (Reading)	811 (Literary writing)
143 (Modern western philosophy)	549 (Social reform movement)
284 (Western mythology)	870 (Western literature)
367 (Ecology)	445 (City planning and hygiene)
435 (Horticulture)	929 (Landscape architecture)
521 (Educational psychology and teaching)	176 (Psychology theories)
610 (General history of China)	820 (Chinese literature)
731 (Japan)	803 (Oriental languages)
876 (French literature)	147 (French philosophy)
992 (Travel)	427 (Food, cookery)

observation of interesting associations more difficult. However, they reflect the general circulation interests of users.

(4) *Erroneous class associations.* Finally, there were some erroneous associations resulting from human indexing errors. For example, class 425 is an unassigned number in CDC, yet an association between it and 805 (English languages) was obtained. It is likely that the correct class would be found to be 424 (beauty culture) after checking the subject contents of the books included. Human analysts sometimes mix up nearby numbers. Another example is class 310 (mathematics) associated with class 501 (sociology), which seems rather unusual. After checking the books borrowed under class 310, they were found to be more suited to class 312 (computer).

Discussion and applications

This paper presents the results obtained by exploring associations between classes of hierarchical classification schemes by computing the similarity values of books borrowed. However, there are some limitations to such an approach. First, the sparseness of the circulation log presents a major problem for discovering useful associations. Comparing the library situation with other business contexts, books are generally more diverse than business products. Furthermore, once a book is purchased and added to a collection, it is seldom moved or withdrawn. In addition, users of a university library may leave the school and no longer borrow books from the library. Owing to these factors, only a small portion of the collection is circulated, and the average frequency of circulation is low. In business environments on the other hand, the number of different products is normally small and any unsold products are withdrawn within a short time as discussed in a previous section.

Furthermore, the circulation behaviors of local users in a university library present some problems as well. Most users borrow books from only a few classes; i.e. circulation interests are usually focused on particular classes. Therefore, it is rather difficult to obtain associated classes for all classes. In addition, many other factors can affect the class associations obtained; for example, users in a university library often borrow books for courses they take instead of borrowing books based on their actual subject interests. This obviously makes it difficult to obtain more valid class associations. In addition, associations derived from old transactions may be not so useful in predicting current usage. In any of the above situations, the discovery process could not be greatly improved even if the size of the circulation log is increased. In other words, results of the similarity analysis could be limited if the data set is rather sparse, regardless of the size of a library or the length of a classification number extracted. However, it is still worthwhile to understand special subject associations through analyzing local users' circulation patterns.

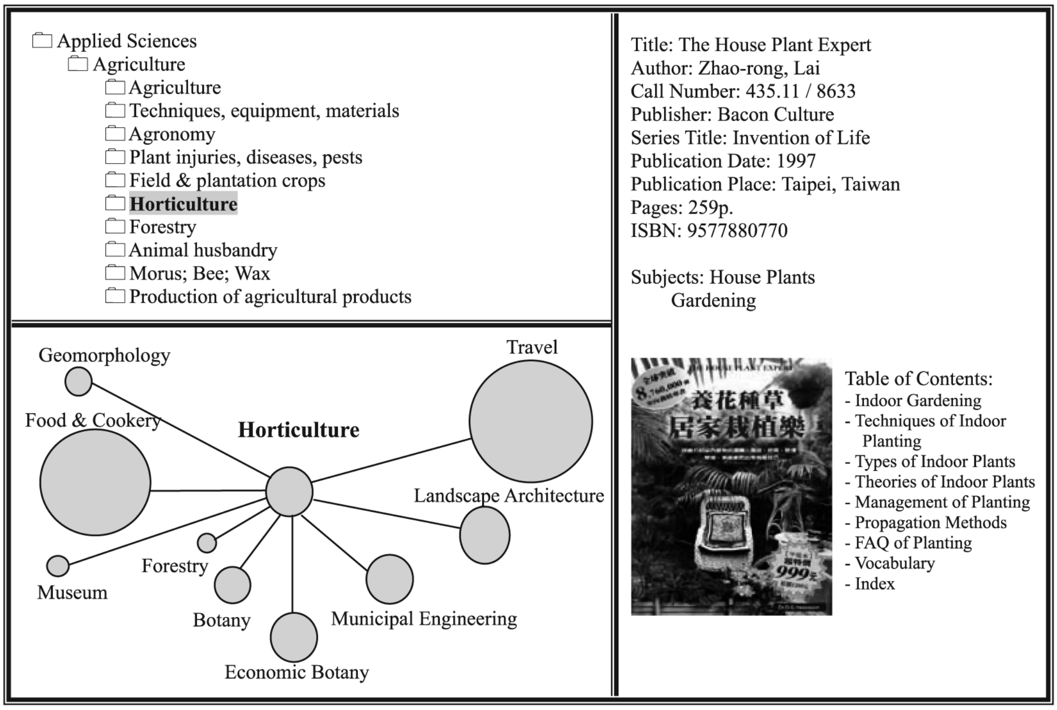
The results of this study are useful for both information organization and retrieval applications. As described previously, class associations can provide a basis for multi-classification of bibliographic records. Since, so far, there is no relative index in the CDC scheme, the results obtained here can be used as a reference for constructing such an index. As for DDC, the proposed approach can also help in updating or enriching the relative index through the addition of user-oriented associated classes or the re-ranking of existing associated classes for each class. For example, Table V lists examples of classes associated with class 435 (horticulture), which are ranked according to their similarity cosine values. Sample book titles with subject headings are listed for reference. Though the cosine value of each class is rather low, it is useful for further analyzing various topics of interest in these associated classes, which are associated with 435. Such topics may include forest esthetics in 436 (forestry), ornamental plants in 375 (botany), perfume flowers in 374 (economic botany), flora and fruit teas in 427 (food), and urban beautification in 929 (landscape architecture). The ranking of associated class may reveal the importance of each topic associated with 435. Such analysis can be helpful for obtaining newly associated topics for a given class in response to changes of users' circulation interests.

For information retrieval applications, an initiative prototype of enhanced view-based OPAC interface has been designed. The prototype exploits the idea of enabling classificatory browsing of a library collection. Figure 7 shows a snapshot of the design. The right window displays the final result of the retrieved bibliographic record selected by an OPAC user. The two left windows display the associated classes for the target class of the bibliographic record listed in the right window. The upper part displays its neighboring classes and their hierarchical relationships in CDC, and the lower part uses a graphic display to show all the associated classes, including non-neighboring classes. The distance from a node (an associated class) to the target class in the graph indicates the proportional strength of their association, and the size of the node indicates the circulation frequency. With such an environment, users may browse the hierarchical structure of the scheme (upper left window) and also view some highly related classes dispersed in the scheme (lower left window).

Table V Examples of ranked associated classes for a given class 435 (horticulture)

Associative classes (caption)	Sample Chinese book title (English translation)	Chinese subject headings (English translation)	Cosine
436 (Forestry)	Lin Ye Zhe Xue Yu Sen Lin Mei Xue (*the esthetics of forests and the philosophy of forestry*)	Lin Ye – Zhe Xue, Yuan Li (*forestry – principles*)	0.267261
375 (Botany)	Lyu Mei Hua Jing Guan Zhi Wu: Cao Hua Pian (*ornamental plants: herbaceous flowers*)	Zhi Wu-Tai Wan (*plants – Taiwan*)	0.194225
374 (Economic botany)	Xiang Hua Zhi Wu (*perfume flowers*)	Hua Hui (*flowers and plants*); Yuan Yi (*horticulture*)	0.174816
427 (Food)	Ou Shi Hua Guo Cha Pu (*recipes for European style flora and fruit teas*)	Cha (*teas*); Yin Liao (*drinks*); Shi Pu (*recipes*)	0.16533
929 (Landscape architecture)	Zhi Zai, Lyu Hua Yu Jing Guan (*planting, virescence and landscaping*)	Zao Yuan (*landscape gardening*); Du Shi Mei Hua (*urban beautification*)	0.149533

Figure 7 A sample snapshot of the prototype of enhanced OPAC interface for subject browsing



□ Applied Sciences
- □ Agriculture
  - □ Agriculture
  - □ Techniques, equipment, materials
  - □ Agronomy
  - □ Plant injuries, diseases, pests
  - □ Field & plantation crops
  - □ **Horticulture**
  - □ Forestry
  - □ Animal husbandry
  - □ Morus; Bee; Wax
  - □ Production of agricultural products

Title: The House Plant Expert
Author: Zhao-rong, Lai
Call Number: 435.11 / 8633
Publisher: Bacon Culture
Series Title: Invention of Life
Publication Date: 1997
Publication Place: Taipei, Taiwan
Pages: 259p.
ISBN: 9577880770

Subjects: House Plants
Gardening

Table of Contents:
- Indoor Gardening
- Techniques of Indoor Planting
- Types of Indoor Plants
- Theories of Indoor Plants
- Management of Planting
- Propagation Methods
- FAQ of Planting
- Vocabulary
- Index

Conclusion

The paper has presented an approach to adding user aspects to hierarchical classification schemes and shown the potential for various applications. With the proposed approach, many associated classes scattered across different subject hierarchies can be discovered, based on the circulation patterns of similar users. The schemes, therefore, can be made more adaptable to changes of users and the uses of different library collections. Various applications in information organization and retrieval have also been discussed in this paper, including enrichment of the relative indexes of schemes and enhancement of the OPAC interface for subject browsing. However, there are some limitations on applying the results, especially since the size and sparseness of the circulation log may prevent the discovery of more in-depth associations. Although such an approach is useful for discovering class associations, other techniques are needed to deal with infrequent classes, such as document-based approaches to extracting sufficient features for similarity computation. In future research, more empirical studies will be needed to validate the findings obtained here, and the methods for obtaining user-oriented associations can be still improved.

References

Banerjee, K. (1998), "Is data mining right for your library?", *Computers in Libraries*, Vol. 18 No. 10, pp. 28-31.

Barr, A. and Sichel, H.S. (1991), "A bivariate model to predict library circulation", *Journal of the American Society for Information Science*, Vol. 42 No. 8, pp. 546-53.

Bertland, L.H. (1991), "Circulation analysis as a tool for collection development", *School Library Media Quarterly*, Vol. 19 No. 2, pp. 90-7.

Chan, L.M. (1995), "Classification, present and future", *Cataloging and Classification Quarterly*, Vol. 21 No. 2, pp. 5-17.

Chen, M.-S., Han, J. and Yu, P.S. (1996), "Data mining: an overview from a database perspective", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8 No. 6, pp. 866-83.

Cunningham, S.J. and Frank, E. (1999), "Market basket analysis of library circulation data", *Proceedings of the 6th International Conference on Neural Information Processing*, Perth, pp. 825-30.

Eldredge, J.D. (1998), "The vital few meet the trivial many: unexpected use patterns in a monographs collection", *Bulletin of the Medical Library Association*, Vol. 86 No. 4, pp. 496-503.

Hickey, T.B. (2000), "CORC: a system for gateway creation", *Online Information Review*, Vol. 24 No. 1, pp. 49-53 (for current information, see www.oclc.org/corc/).

Karypis, G. (2001), "Evaluation of item-based top-n recommendation algorithms", *Proceedings of the 2001 ACM CIKM International Conference on Information and Knowledge Management*, Atlanta, GA, pp. 247-54.

Kent, A. and Williams, J.G. (1979), *Use of Library Materials: The University of Pittsburgh Study*, Marcel Dekker, New York, NY.

Lai, Y.-H. (1989), *New Classification Scheme for Chinese Libraries: Table*, 7th ed., Sun-Ming, Taipei, Taiwan (in Chinese).

Mitchell, J.S. and Vizine-Goetz, D. (2000), "A research agenda for classification", available at: www.oclc.org/dewey/research/research_agenda.html

Paepcke, A., Garcia-Molina, H., Rodriguez-Mula, G. and Cho, J. (2000), "Beyond document similarity: understanding value-based search and browsing technologies", *SIGMOD Records*, Vol. 29 No. 1, pp. 80-92.

Pollitt, A.S. (1998), "The application of Dewey Classification in a view-based searching OPAC", *Proceedings of the 5th ISKO Conference*, Ergon Verlag, Würzburg, pp. 176-83.

Salton, G. (1989), *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, Reading, MA.

Salton, G. and Buckley, C. (1988), "Term weighting approaches in automatic text retrieval", *Information Processing and Management*, Vol. 24 No. 5, pp. 513-23.

Sarwar, B., Karypis, G., Konstan, J. and Riedl, J. (2001), "Item-based collaborative filtering recommendation algorithms", *WWW10 Conference*, Hong Kong.

Solomon, P. (1992), "User-based methods for classification development", *Advances in Classification Research*, No. 2, pp. 163-70.

Thompson, R., Shafer, K. and Vizine-Goetz, D. (1997), "Evaluating Dewey concepts as a knowledge base for automatic subject assignment", OCLC, Dublin, OH, available at: http://orc.rsch.oclc.org:6109/eval_dc.html

Tinker, A.J., Pollitt, A.S., O'Brien, A. and Braekevelt, P.A. (1999), "The Dewey Decimal Classification and the transition from physical to electronic knowledge organization", *Knowledge Organization*, Vol. 26 No. 2, pp. 80-96.

Wilson, M.D. and Spillane, J.L. (2000), "The relationship between subject headings for works of fiction and circulation in an academic library", *Library Collections, Acquisitions, and Technical Services*, Vol. 24 No. 4, pp. 459-65.

Yee, M. (1998), "Guidelines for OPAC displays prepared for the IFLA task force on guidelines for OPAC displays", available at: www.ifla.org/VII/s13/guide/opac.htm