Subject Categorization of Query Terms for Exploring Web Users' Search Interests

Hsiao-Tieh Pu
Institute of Information Management, National Chiao Tung University, Hsinchu, Taiwan 300.
E-mail: htpu@cc.shu.edu.tw

Shui-Lung Chuang
Institute of Information Science, Academia Sinica, Taipei, Taiwan 115. E-mail: slchuang@iis.sinica.edu.tw

Chyan Yang
Institute of Information Management, National Chiao Tung University, Hsinchu, Taiwan 300.
E-mail: cyang@cc.nctu.edu.tw

Subject content analysis of Web query terms is essential to understand Web searching interests. Such analysis includes exploring search topics and observing changes in their frequency distributions with time. To provide a basis for in-depth analysis of users' search interests on a larger scale, this article presents a query categorization approach to automatically classifying Web query terms into broad subject categories. Because a query is short in length and simple in structure, its intended subject(s) of search is difficult to judge. Our approach, therefore, combines the search processes of real-world search engines to obtain highly ranked Web documents based on each unknown query term. These documents are used to extract cooccurring terms and to create a feature set. An effective ranking function has also been developed to find the most appropriate categories. Three search engine logs in Taiwan were collected and tested. They contained over 5 million queries from different periods of time. The achieved performance is quite encouraging compared with that of human categorization. The experimental results demonstrate that the approach is efficient in dealing with large numbers of queries and adaptable to the dynamic Web environment. Through good integration of human and machine efforts, the frequency distributions of subject categories in response to changes in users' search interests can be systematically observed in real time. The approach has also shown potential for use in various information retrieval applications, and provides a basis for further Web searching studies.

Introduction

As Web searching has grown, research interests in using search engine logs to understand Web users' searching behaviors has increased. Many recent studies have been performed with focuses on the statistical or linguistic characteristics of Web searching, such as query length, query structure, and query lexical patterns (Jansen & Pooch, 2001). Subject content analysis of queries, such as exploring the search topics and observing changes in their frequency distributions with time, has been less investigated. This is largely due to the difficulties involved in manually processing large numbers of dynamic queries, including problems related to scalability, currency, subjectivity, and lack of domain knowledge. It is believed that good integration of human and machine efforts is necessary to explore Web searching more cost effectively. This article, therefore, presents a query categorization approach to automatically classifying query terms from search engine logs into broad subject areas. It can help Information Retrieval (IR) systems obtain query terms in various subject domains and assist them in observing users' search interests on a larger scale.

The current research continues our previous study (Pu, Chuang, & Yang, 2001), and the purpose is twofold: (1) to develop a feasible approach to categorizing large numbers of dynamic queries into predefined taxonomy, and (2) to explore Web searching interests through integration of the proposed autocategorization approach and human analysis using real-world search engine logs. Based on the need for broad coverage of user groups and query transactions from long periods of time, three search engine logs in Taiwan were collected and tested. They contained over 5 million queries from different periods of time.

In this study, query terms submitted to a search engine were the target of analysis. Note that a query term is not limited to the Chinese language; it may contain one or multiple words in English or a sequence of characters in

Chinese. For queries containing more than one query term, for example, queries concatenated with Boolean operators, each of the composed query terms will be treated as a single term in this article.

Because a Web query is short in length and simple in structure, its intended subject(s) of search is difficult to judge, especially in the diverse Web environment. To obtain more current information for identifying the subject domain(s) of an unknown query term (unknown term), our idea is to employ highly ranked documents retrieved by the unknown term as a source for feature extraction. Therefore, our approach combines the search processes of real-world search engines. These engines index huge numbers of Web pages, and only a small number of users' queries have no matching documents. Each unknown term is sent to more than one search engine to obtain highly ranked Web documents. Only documents that can be retrieved by all of the engines are used to extract the unknown term's cooccurring precategorized terms (feature terms) to create a feature set. An effective ranking function has also been designed for assigning appropriate weighted value to each feature term, and to help find the most probable categories for the unknown term. If the feature set contains discriminated feature terms, the most possible subject domain(s) of the unknown term can be determined even though these documents may not exactly serve the search purposes of the user.

Besides designing a high-performance categorization algorithm, the proposed approach also needs to (1) employ an adequate subject taxonomy covering popular search interests, (2) extract a sufficient number of seed terms for each subject category, and (3) precategorize seed terms as feature terms for the autocategorization process. Our subject taxonomy was built in a bottom-up manner through qualitative analysis of a number of popular queries, which consisted of some broad subject categories in the initial stage. An adequate number of popular queries, which lasted for a long period in time, were extracted from the collected logs and treated as seed terms. These terms are assumed to be easier for users to use to express search topics and also appeared frequently in Web documents with similar subjects. Human analysts then manually categorized these seed terms into the predefined taxonomy as feature terms for the later autocategorization process.

The accuracy of the proposed autocategorization approach and the effects of different seed term sets were extensively tested. The achieved performance was quite encouraging compared with that of human categorization. For systematic analysis of users' search interests, an observation system that integrated both the autocategorization approach and human analysis was constructed to deal with all of the collected logs. Changes in the frequency distributions of subject categories in response to users' search interests can, therefore, be observed in real time. This approach has been shown to be efficient for analyzing users' subject interests on a larger scale than could be achieved in previous studies. With this approach as a basis, various Web IR applications can be developed, including approaches designed to enhance Web retrieval effectiveness, to aid Web content organization, and to facilitate Web user studies.

The remainder of the article is organized as follows: the next section briefly reviews some of the major related studies. The Problem Considered section describes the problem considered and gives an overview of the proposed approach. The Environmental Environment section provides background information about the experimental environment, including data collection and analysis, and preparation of the subject taxonomy. The Proposed Categorization Approach section introduces the proposed categorization approach in detail, including extraction of seed terms, manual categorization of seed terms, and autocategorization of unknown terms. The experimental results then discussed and a performance evaluation is given. The Observations of Users' Search Interest section presents a systematic way to observe users' search interests. Various possible Web IR applications and future research are then discussed, and then the paper concludes.

Related Work

Recently, there has been growing interest in analysis of Web query logs from Internet search engines. Jansen and Pooch (2001) provided a good review of the state of research in the field. Three representative search engines have been studied so far, namely, AltaVista (Silverstein, Henzinger, Marais, & Moricz, 1999), Excite (Jansen, Spink, & Saracevic, 2000; Xu, 1999), and Fireball (Hoelscher, 1998). These studies have provided basic information about Web users' searching characteristics, such as the fact that users mostly input short queries, rarely use advanced search features, browse few search result pages, and usually input only one query in each search session. Besides general statistical analysis, a variety of research topics have drawn interest. Ross and Wolfram (2000) conducted query term-pair topic analysis, and showed that there is some commonality among the most popular Web term cooccurrences. Jansen, Spink, and Pfaff (2000) examined the lexical patterns of queries divided into five categories, and found that Web queries are mostly noun phrases. Similar to the above search engine studies, basic statistical analysis of queries from Chinese users in Taiwan was conducted in our previous work, as will be discussed in the next section (Pu, Chuang, & Yang, 2000).

Studies on what users search for have been conducted recently by Ross and Wolfram (2000) and by Spink, Wolfram, Jansen, and Saracevic (2001). Ross and Wolfram used cluster analysis to group the top 1,000 term pairs extracted from Excite's log containing 363,282 unique queries collected within 1 day. Interestingly, the results were similar to those we obtained using our proposed approach, such as the fact that there were very distinct domains of adult entertainment and of computing/computer-mediated communication and play, and that there were less distinct domains of media interest and information needs. However, the authors stated that a limited amount of data is unable to reflect the Internet

in general, and that the clustering method is rather an exploratory technique that cannot provide irrefutable proof. Spink et al. (2001) took a different approach. They used term cooccurrence analysis and a human classification method. Their study proved that query-level analysis is more productive than term-level analysis in answering subject content of a query term.

On the other hand, transaction logs have also been used to study subject searching in traditional IR systems (Drabenstott & Vizine-Goetz, 1994; Hildreth, 1985). For example, many generalizations have been obtained about the subject terms users enter into online catalogs (Carlyle, 1989). An important advantage of transaction logs is the unobtrusiveness of this kind of data collection approach (Kaske, 1993). However, this method has disadvantages. Drabenstott and Vizine-Goetz pointed out that it is difficult to determine exactly what users are looking for using computer analysis only. Manual analysis is more accurate than computer analysis because a human intermediary can demarcate individual searches by employing both time stamps and the meanings of user queries. In addition, transaction log analyses of users' subject terms can also aid the development of subject control vocabularies and facilitate subject searching in on-line catalogs. For example, users' searches and subject heading lists as a supplemental vocabulary can be matched to enrich a classification scheme (Vizine-Goetz & Godby, 1996), and on-line catalogs that respond to a wide variety of user queries for subjects can be designed (Drabenstott & Weller, 1996). Based on the above studies, it is believed that good integration of human and machine efforts is necessary to explore Web searching more cost effectively.

Some early research on term clustering was related to our work, including works on latent semantics, SVD, and term relationship analysis (Baeza-Yates, & Ribeiro-Neto, 1999; Salton, 1989). Most of these studies dealt with clustering of relevant terms based on cooccurrence analysis; i.e., terms were clustered if they cooccurred in similar documents. Different from these works, in our approach, each unknown term is categorized into a predefined taxonomy based on cooccurring feature terms extracted from retrieved documents. Besides, the approach to the estimation of similarity between each unknown term and candidate subject category is similar to those used in some research on document classification. Van Rijsbergen (1979) provided a good discussion of these automatic methods. Some related studies on the classification of bibliographic information can be found in (Larson, 1992; Shafer, Subramanian, & Fausey, 1999; Schwartz, 1981).

The Problem Considered and an Overview of the Proposed Approach

A query usually represents a compromised information need (Taylor, 1962), and is the primary means of translating a user's request into a form that an IR system can understand. Query logs from Web IR systems like search engines are, thus, thought to be the foremost source of unobtrusive data related to users' requests for information. To observe users' search interests, it is necessary to analyze the subject contents of their queries. The problem considered in this article is, then, defined as follows.

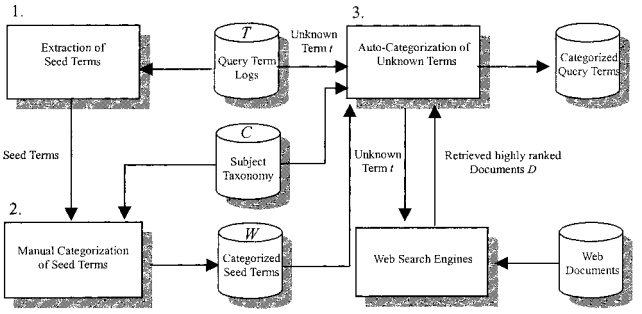


FIG. 1. An abstract diagram showing the design of the proposed autocategorization approach.

Suppose that T is a set of unknown terms, each of which can be a single word or multiple words in English, or a sequence of characters in Chinese, and assume that C is a predefined subject taxonomy for organizing these query terms. The problem is to develop an automatic categorization method that is effective in classifying each term t in T into one or more appropriate categories in C that indicate the subject domain(s) of ts search interests. More precisely, for each $t \in T$, the categorization method determines a set $C(t) \in C$, where each $c \in C(t)$ represents a category that term t may be related to. The above problem is certainly challenging, because queries are usually short, and a large number of new requests are continuously appearing.

The proposed categorization approach consists of three analysis steps shown in Figure 1. In the first step, a modest number of high-frequency terms called seed terms are extracted from query log T. Using them as the features for categorization is believed to be more effective than just using common terms, as will be discussed in a later section. In the second step, human analysts categorize these seed terms into a set W using the predefined subject taxonomy C, which is restricted to a moderate number of popular subject areas in the initial stage. With the help of C and W, the third step involves autocategorizing each unknown term t in log T into appropriate categories as the output. In addition, the autocategorization process is combined with the use of Web search engines to obtain necessary parameters; i.e., the unknown term t is sent to obtain the categories of the cooccurring precategorized seed terms from the retrieved documents D. Each step will be discussed in detail in the Proposed Categorization Approach section after the experimental environment is introduced.

Experimental Environment

Data Collection and Analysis

In our study, we collected three query logs in different periods of time from Dreamer, GAIS, and Openfind, which

TABLE 1. Experimental data sets from three search engines.

Data sets	Source (period of data collection)	No. of distinct query terms	Total frequencies
D-1998	Dreamer (3 months in early 1998)	228,566	2,184,256
G-1999	GAIS (2 weeks in mid-1999)	114,182	475,564
O-2000	Openfind (top queries of 12 months in 2000)	3,011	2,493,211

are representative Web search engines in Taiwan. The effects of time locality and potential bias in choosing only one set of test data are, therefore, alleviated. The three engines serve general Web users by providing Web directory and page search and also provide other network services similar to those of Yahoo! and Google. The scale of their coverage of Chinese Web pages is large; for example, Openfind at the time of our research contained over 21 million Chinese Web pages, making it number one among the major Chinese search engines in Chinese communities (Openfind, 2001). As to their users, varied types of people presumably exist. For example, Dreamer is a media-oriented portal; GAIS was initiated by a university lab and has the longest history of development among search engines in Taiwan; and Openfind is a Chinese portal comparable to other internationally known search engines (He, 1999; Lighthouse, 2001). However, generally there is no clear distinction among their users according to our analysis of the query logs and other related user surveys. Although the three engines may only represent a small portion of Internet uses, they contain consecutive time frames with the years of 1998, 1999, and 2000 useful for longitudinal studies.

As listed in Table 1, the Dreamer's log was collected from a period of over 3 months in early 1998 (D-1998); GAIS's log was collected within a period of 2 weeks in mid-1999 (G-1999); and Openfind's log contained top 1,000 query terms of each month for 12 months in 2000 (O-2000). D-1998 and O-2000 contained only distinct query terms and the corresponding aggregated frequency, and G-1999 contained search requests in which each included a query and a corresponding submitted time stamp. Among them, D-1998 was more comprehensive in coverage and used as the basis for observing the frequency distributions of users' search interests and extracting seed terms. G-1999 and O-2000, on the other hand, were used to observe changes of search interests with time.

In our analysis of the factual characteristics of the above query logs, we examined the query length, query structure, and language used. The statistical results showed that the average length of a query in Chinese was 3.18 characters. According to general statistics for the Chinese language (Chien, & Pu, 1996), the average word length is 1.5–1.6 characters; hence, 3.18 characters can be treated as a word bi-gram or word pair. However, the languages used in the queries are not limited to Chinese; thus, the proposed approach has to deal with both Chinese and English queries. The average length of an English query was 1.10 words, which was much shorter, and the queries mostly consist of proper nouns like "IBM," "Microsoft," etc. As for the structure of the queries, although the three search engines provide some advanced search functions like Boolean operators, these functions appeared in less than 1% of the queries. This coincides with results of the previous related studies, which found that the query structure was not complex. Because query terms are mostly short in length and simple in structure, it is assumed in this article that the problem of categorizing an unknown query can be simplified to that of categorizing a query term. A query term then is treated as a single term, which can be one word or multiple words in English or a sequence of characters in Chinese.

Meanwhile, it was noted that a small number of queries were repeated many times. Taking D-1998 as an example, 4.33% of the query terms covered 74.89% of the total frequencies. Although Web users do not limit themselves to a small number of popular queries, overall Pareto's 20/80 Law is observed. This reveals that an adequate number of high-frequency query terms may represent most of the popular requests. Thus, we used the top nearly 20K query terms from D-1998 as the seed terms, which represented 81% of the total number of query terms in the log. Although the coverage of the logs used may not be comparable to those of search engines like AltaVista reported previously, they were basically sufficient for observing users' popular search interests. In addition, the results are also valuable for observation of cultural differences between Western and Chinese users, which to our knowledge have not been studied before.

Structuring the Subject Taxonomy

It is necessary to construct adequate subject taxonomy beforehand for the subject categorization. Such taxonomy can be intuitively derived from some well-established schemes like library classification schemes or Internet directories. As Weinberg (1996) pointed out, such discipline-based indexing systems may suffer from the problems of currency and specificity. It was not suitable to directly apply these schemes; instead, they were used only as references. Our subject taxonomy was a two-level scheme consisted of 15 major categories and 85 subcategories. Although its structure was similar to that of a discipline-based system, the categories were determined by analyzing topics of a number of popular queries.

The development of our subject taxonomy can be said to be based on the grounded theory approach (Strauss & Corbin, 1990), which suggests that the taxonomy built from the bottom up, using raw data. The raw data included popular query terms and directory trees obtained from several commercial search services. In our research, the process of building an adequate scheme for categorizing query terms was derived from those used to develop an empirical taxonomy of a value in use of library and information services as reported by Saracevic and Kantor's study

(1997). The tasks to build the subject taxonomy and the later categorization were mainly done by a team of five Library & Information Science (LIS) students and a reference librarian. These professionals have substantial experience with net surfing and adequate training in subject indexing techniques.

First, one analyst quickly perused the top 5K query terms from D-1998 and identified representative subjects based on their importance in terms of frequencies. Next, the analyst grouped these query terms into categories associated with specific directory names collected from commercial directory search services and merged, and gave each category a label. Then, these categories were grouped into a working classification scheme to be used in testing. Next, the analyst went back to the query logs and classified each query term according to that working scheme. Then, the analyst wrote a set of instructions for encoding to test the reliability of the scheme when the query logs were encoded by other analysts. Finally, two analysts not previously engaged tested the scheme for intercoder agreement until the level of consistency was acceptable. The steps were reiterative and involved considerable feedback. In the end, the scheme contained 15 major categories and 85 subcategories, including Adult, Arts & Humanities, Business & Finance, Computers & Networks, Education, Entertainment, Games, Healthcare, News & Media, Politics & Society, Recreation & Chat, Science & Technology, Shopping, Travel, and Unknown, listed alphabetically. Each major category consisted of several subcategories as well; for example, the Computers & Networks category was divided into eight subcategories, namely, BBS, Company, Graph & Picture, Hardware, Network Services, Other, Search Engines, and Software.

Although the taxonomy constructed was based on a small sample of data, it represented much of the popular search interests concerning with their usage in frequencies as described before. With the aid of the autocategorization process discussed in a later section, new query terms can be categorized and the taxonomy can be examined and updated accordingly. With such a basis, it will also be helpful for further analyzing the topics in these categories toward building a topic-based indexing system in the future.

The Proposed Categorization Approach

The proposed categorization approach as shown in Figure 1 consists of three analysis steps. Each step is discussed in detail in the following subsections.

Extraction of Seed Terms

The first step involves analyzing and extracting a sufficient number of seed terms from the test logs for each subject category. The set of seed terms extracted should be both as broad in coverage and as modest in size as possible. Meanwhile, because many queries may be related to ephemeral interests, such as a new movie or some recent events, seed terms must also be sustainable.

TABLE 2. Coverage comparison between the D-1998 and G-1999 search engine query logs based on the numbers and percentages of common distinct query terms.

G-1999 D-1998	Top 1K terms	Top 20K terms	All
Top 1K terms	583/58.30%	977/97.70%	992/99.20%
Top 20K terms	914/91.40%	9,709/50.71%	14,721/76.89%

Before seed terms were extracted from the logs, query terms from D-1998 were matched and filtered using the G-1999 log to compare their coverage. Table 2 shows that 76.89% (see the third row of the fourth column in Table 2) of D-1998's top 20K query terms still existed in G-1999's 2 week randomly selected log. This indicates that many important search interests were not much affected by time. Note that only eight popular terms in D-1998's top 1K did not appear in G-1999's top 1K (see the second row of the fourth column in Table 2). Also interesting is that 417 and 86 popular terms in G-1999's top 1K did not appear in D-1998's top 1K and 20K, respectively (see the second and third rows of the second column in Table 2).

It was also found that query terms affected by time were mostly proper nouns. On the other hand, terms not affected by time, except for some proper nouns like the names of famous Web sites and people, were mostly subject terms like "movie," "baseball," or "flight ticket." These terms are considered to be core terms in this article because they are long-lasting, modest in size, and rich in content. From our observations, core terms are more comprehensive in meaning and are often used by Web users to express popular search interests. Using them as features in the categorization process is believed to be more effective than just using high-frequency terms and random-selected terms. Comparisons will be made in a later section. In this study, 9,709 terms were used as test core terms as they appeared both in the top 20K query terms of D-1998 and the top 20K of G-1999.

Manual Categorization of Seed Terms

In the second step, human analysts categorize seed terms as feature terms for the later autocategorization process. During the categorization process, many principles can be derived through subject analysis as bibliographic information is organized (Chan, 1994). However, four steps are important when categorizing a query term: (1) determine useful principles for categorization, (2) deal with short query terms containing little information for analysis, (3) alleviate the lack of subject domain knowledge that human analysts have, and (4) reduce the level of inconsistency that occurs due to human indexing.

Human analysts are recommended for judging the subject domain(s) of a query based on possible search purposes. According to our observations, most query terms are single nouns, and phrase terms are rare. It is reasonable to assume that a query term usually represents one topic. As a consequence, in most cases, each query term was assigned to one major category and one subcategory; for example, the term "mutual fund" was categorized into the Personal Finance under the major category Business & Finance. Yet it is very likely that a single query term represents multiple information requests from different users. In such cases, up to three categories could be assigned to a given query term. For example, the term "ICQ" could represent a request for either chat sites or software downloads; therefore, it was classified into both the subcategory Chat under the major category Recreation & Chat and the subcategory Software under Computers & Networks.

To facilitate the classification task, human analysts were recommended to utilize various resources to increase their domain knowledge, not limited to Internet directories or Web page searches. Although some terms could not be categorized appropriately into the predefined 85 subcategories, relatively few of these were popular queries. Meanwhile, for categorization to be as accurate as possible, each selected seed term was categorized by at least two analysts from our team. Although they were trained to assign subjects as consistently as possible, it was found that about 10 and 20% of the seed terms were assigned into different major categories and subcategories, respectively. In addition, as inconsistency in human indexing was inescapable (Leonard, 1977; Soergel, 1994), another analyst not previously engaged in the categorization process made the final decision. On average, it took about 80 seconds for an analyst to categorize a query term. In total, it took over 400 man-hours to finish categorizing the 9,709 seed terms.

While the human analysts may have suffered from subjectivity, inconsistency, and lack of domain knowledge, in the initial stage, they were involved mainly in classifying the seed terms into appropriate subject categories. If there were an adequate number of feature terms in each category, the distortion caused by a few invalid human categorizations, in fact, would not seriously affect the accuracy of the autocategorization approach.

Autocategorization of Unknown Terms

The third step involves instant categorizing of each unknown term. As mentioned above, autocategorization has to do with feature extraction from retrieved documents. However, such documents are diverse in terms of their contents, and subject determination is not very straightforward. A number of issues need to be carefully considered beforehand to achieve better performance in classification, including the corpus for each subject domain and appropriate extraction of discriminated features in the documents (Goller, Loning, Will, & Wolff, 2000). In our approach, we simulate the work of a human analyst in determining the corresponding subjects of a term that is outside of his/her domain knowledge. When the subjects of a document are difficult to determine, a human analyst usually refers to the categories of some known terms contained in that document. The assignment of subject categories for an unknown term can then be based on the categories of its cooccurring seed terms.

Because we use cooccurring seed terms in Web documents containing an unknown term as the feature set for subject analysis, our approach is very similar to that of classifying a document based on the composed key terms in a conventional document classification process, or to that of tagging a part of speech for a word with its cooccurring neighboring words in linguistic analysis. Assuming that there is a set of highly ranked Web documents D related to the unknown term t, we obtain cooccurrence information by utilizing a seed term set W, which has been precategorized using the subject taxonomy C. For each possible category c for the term t, a ranking function is defined to estimate the categorization confidence of t belonging to c as follows:

$$R(t,c) = \sum_{w \in W_{t,c}} N(D_{w,t}) * f_w / N(D_w),$$

where $W_{t,c}$ is the cooccurring term set for category $c \in C$; D_w and $D_{w,t}$ are the sets of documents in D, which contain the term w and both terms w and t, respectively. $N(D_w)$ and $N(D_{w,t})$ are the numbers corresponding to D_w and $D_{w,t}$, respectively, and f_w indicates the total frequency occurrence of the term w in D_w. Extraction of $W_{t,c}$ will be further described in a later paragraph.

The function $R(t,c)$ is employed in a TFIDF-like approach, a ranking method commonly used in the well-known vector space model of IR research (Salton & McGill, 1983), to assign a weighted value for each cooccurring seed term. Whether t can be categorized into c depends on the number of cooccurring terms in $W_{t,c}$. The developed categorization process ranks all candidate categories in C using the ranking function $R(t,c)$ to find the most appropriate categories for term t, and to also judge whether the accumulated weighted values of these categories are large enough. From the above discussion, it is clear that cooccurring seed terms play a crucial role in implementing the ranking function. If an unknown term contains a sufficient number of cooccurring seed terms with correct subject categories in D, then its categories can be more easily and correctly assigned. On the other hand, it may not be correctly assigned if such information is lacking.

Considering that document set D may not be large enough for cooccurrance analysis, the estimate of the number of cooccurring terms in set $W_{t,c}$ is obtained using global information. In other words, the estimate is based on mutual information-based association estimation between t and each term $w \in W$, where w is in D and is classified into category c. $w \in W_{t,c}$ should also satisfy the condition that $N(w,t)/(N(w)+N(t)) >$ a threshold value.

For more queries to have a sufficient number of features, real-world search engines are used to retrieve highly ranked Web documents. Each unknown term t is sent to the search engines to retrieve the overlapping top n documents as t's most relevant documents. These engines index huge numbers of Web pages, and only a small number of the queries in the test log have no matching documents. Furthermore, it is assumed that the search engines consider the relevancy and popularity of retrieved documents, although this may not be perfectly implemented. In fact, the ranking of the retrieved documents was not considered in the extraction of feature terms. If there are sufficient feature terms in the retrieved set, the most probable subject domain(s) of the unknown term can be determined. Therefore, we retrieved at least the top 100 documents appearing in all of the engines used to obtain more reliable feature terms.

The above mutual information-based association estimation process can also be performed by querying the search engines. For each candidate seed term w, its corresponding $N(w)$, $N(t)$, and $N(w,t)$ can be estimated using the numbers of retrieved documents queried by w, t, or w and t, respectively. In addition, our ranking process can be performed along with an interactive computation process for each cooccurring seed term w in $W_{t,c}$, where $N(D_{w,t})$, f_w, and $N(D_t)$ are obtained by analyzing the retrieved documents in D. We present the major concepts behind the autocategorization algorithm in the Appendix for reference. These include two procedures: the input of the subject-categorization procedure consists of the unknown term, the predefined subject taxonomy, feature terms, and retrieved highly ranked Web documents, and the output is the suggested categories for the unknown term. The ranking procedure helps rank the categories of cooccurring seed terms extracted from retrieved documents, and the output is the confidence value for each candidate category for use in the subject-categorization procedure.

Note that the above process is dynamic and adaptable to the changing environment of the Web, i.e., changes in Web documents or modifications of search engines. However, the document set D has a great impact on the performance of the proposed approach. Many documents in D may not be relevant to term t. Furthermore, the proposed ranking function does not yet consider the authority of Web documents, and detailed comparisons of various ways of extracting useful document features for the categorization process are also needed. Despite the above issues, the current ranking function is basically a useful initial approach to observing the feasibility of the proposed approach. This function will be refined in our future research.

Experimental Results and Performance Evaluation

Two experiments using D-1998 and O-2000 were conducted. D-1998 was used to test the accuracy of the autocategorization approach compared with that of human categorization. O-2000 was used to test the sustainability of different term sets as feature sets for categorization, such as core terms and high-frequency terms.

The first experiment was conducted to evaluate the performance in categorizing the test query terms into the predefined 15 major and 85 subcategories. The 9,709 core terms that had been manually precategorized were taken as the seed term set. Another 1K noncore terms, randomly selected from the top 20K query terms in D-1998, were treated as the test term set and also manually categorized as a basis for comparing the accuracy achieved in subject categorization.

Next, the test terms were sent as queries to three real-world search engines, consisting of Google Chinese, Openfind, and Yahoo-Kimo, to obtain the required Web documents. The number of existing indexed Chinese Web pages exceeded 21 million at the time of our test. Up to 100 highly ranked Web documents were collected for each test term, and the titles and some descriptions of the documents were extracted as surrogates of the full documents. Cooccurring seed terms could, therefore, be extracted from the surrogates.

Table 3 gives some of the autocategorization results obtained. The three columns on the left side list the query terms (English translations are provided for Chinese queries), categories assigned by humans, and the top five categories suggested by the machine, respectively. The right column lists the name and symbol of each major category and subcategory listed in the left columns. Taking "real player" as an example, it was classified into both the "cd" and "cn" categories by humans, and these were also the first and second categories suggested by the machine. The categories suggested by the machine that matched the categories assigned by humans are in italics. As shown in the right column, the "cd" represents the subcategory Software under the major category Computers & Networks, and "cn" indicates the subcategory Network Services under Computers & Networks.

Note that although in many cases, the top five categories were not exactly the same as the categories assigned by humans, some of them were, nevertheless, related to the manually assigned categories and should not be just considered examples of incorrect categorization. For example, "chinatrust" (one of the largest commercial banks in Taiwan) was categorized into the Banks (bb) by humans, and this category is listed fourth among the top five categories suggested by the machine. However, the other four categories are also related to various services provided by that Bank. Personal Finance (bm) is possibly related to personal investment opportunities, Travel Local (tl) and Travel Abroad (tf) are related to travel package promotions, and Network Services (cn) is related to the finance portal services. In addition, the suggested categories can also be used by human analysts to reexamine correctness of manual categorization.

In addition, to understand the effects of various sizes of core term sets, we ran experiments with the top 100, 300, . . . , 9,709 terms, respectively. Some of the obtained

TABLE 3. Sample results of autocategorization compared with those of human categorization.

Query term (translation in English)	Categories assigned by humans	Top 5 categories suggested by the machine					Major category	/Subcategory	:Symbols
chinatrust (a commercial bank)	bb	bm	tl	tf	*bb*	cn	Arts & Humanities	/Arts	:aa
							Business &	/Banks	:bb
wap	bn	lp	cn	bb	*bn*	kb	Finance	/Electronics Industry	:be
								/Business Information	:bf
								/Personal Finance	:bm
real player	cd, cn	*cd*	*cn*	mn	ch	pc	Computers &	/Telecom Industry	:bn
							Networks	/Software	:cd
								/Hardware	:ch
michael jackson	ei	em	cd	tl	en	cn		/Network Services	:cn
								/Search Engines	:cs
shu-ma-bao-bei (Digimon, a popular animated film)	fa	mn	cs	*fa*	cn	cn	Entertainment	/Stars	:ei
								/Popular Music	:em
								/Entertainment News	:en
liao-tian-shi (chat rooms)	fc	*fc*	cn	cs	mn	be	Recreation & Chat	/Animation	:fa
								/Chat	:fc
								/Sports	:fs
shi-ji-di-guo (a popular computer game)	gg	*gg*	cd	cn	gk	cd	Games	/Computer Games	:gg
								/Game Codes	:gk
sheng-wu-jing-pian (bio-chips)	ks	*ks*	mn	bf	bm	cs	Sci. & Tech.	/Bibliographic Info.	:kb
								/Science	:ks
café	le	*le*	cn	mn	cd	em	Shopping	/Food & Restaurants	:le
								/Festivals	:lo
								/Mobile Products	:lp
can-ting (restuarant)	le	*le*	cn	tf	be	bf	Media & News	/News	:mn
							Politics & Society	/Local Culture	:pc
sheng-dan-ka (christmas card)	lo	*lo*	aa	be	fs	cs	Adult	/Sex Photos	:sp
								/Sex Info.	:ss
xie-zhen-ji (sex-related photos)	sp	*sp*	cn	cs	ss	em	Travel	/Travel Abroad	:tf
								/Travel Local	:tl

inclusion rates are shown in Table 4, and the corresponding curves are also depicted in Figure 2, where the top *n* inclusion rates are the average rates of the highly ranked *n* candidate categories suggested by the machine that matched with the categories assigned by humans. If only the top one categorization result was considered, the average inclusion rate was 51.05%. If the top five categories were considered, the average inclusion rate was 81.37%. In addition, the obtained top one inclusion rate could reach almost 80% if only 15 major categories instead of 85 subcategories were considered. Although it is assumed that the level of accuracy that can be achieved will be even higher if more core terms are used, the accuracy achieved using these 9,709 core terms was quite stable. The performance achieved in this study is quite encouraging compared with the results obtained using human indexing.

Besides using core terms as the seed term set (*coreterm* in Fig. 3), there are various ways to select seed terms, such as high-frequency terms or random-selected terms. The high-frequency terms are selected based on their frequencies (*freqterm* in Fig. 3). The *freqterm* set used in this study contained the most popular terms in the log and could contain some noncore terms. In addition, the random-selected terms can be selected from different query term sets (*freqrand* and *baseline* in Fig. 3). The *freqrand* set and the *baseline* set contained the terms randomly selected from the 9,709 core terms set and the top 20K terms in D-1998, respectively. The top one inclusion rates obtained using the

TABLE 4. Average top 1–5 inclusion rates with various core term sets used as the seed term set for categorization based on 85 subcategories. (The horizontal rows indicate the sizes of the core term set used for categorization, and the vertical columns list the obtained inclusion rates for the top 1–5 candidate categories.)

	100	300	500	1,000	3,000	5,000	7,000	9,000	9,709
Top-1	25.36	36.66	39.63	44.31	46.45	47.74	49.58	50.57	51.05
Top-2	31.35	46.91	50.98	57.84	61.49	63.32	65.16	66.38	66.77
Top-3	34.11	51.67	56.01	63.70	68.01	70.58	72.39	73.83	74.23
Top-4	35.71	54.97	59.60	67.81	72.11	74.61	76.60	78.08	78.44
Top-5	37.02	57.46	62.30	70.56	75.56	77.65	79.70	81.13	81.37

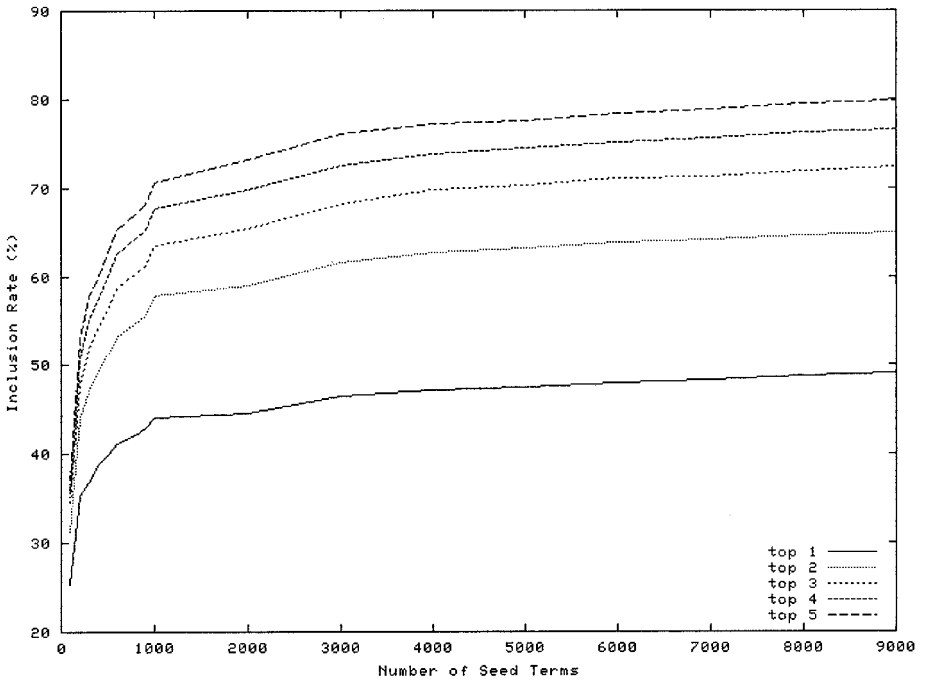


FIG. 2. Curves showing the average top one to five inclusion rates with various core term sets used as the seed term set for categorization based on 85 subcategories.

four different term sets to categorize 9,709 seed terms are shown in Figure 3. The *baseline* set obviously had the poorest performance, and the *coreterm* set and *freqterm* set achieved comparable results. Although the *freqterm* set slightly outperformed the *coreterm* set, the results were likely influenced by the test data derived from the same log.

Meanwhile, the sustainability of core terms and high-frequency terms as the seed term sets for categorization is also tested. The second test was further conducted using the 3,011 distinct query terms from O-2000 as the test term set. The test set has a 2-year lag different from the seed term sets obtained using D-1998 in the first test. The obtained exper-

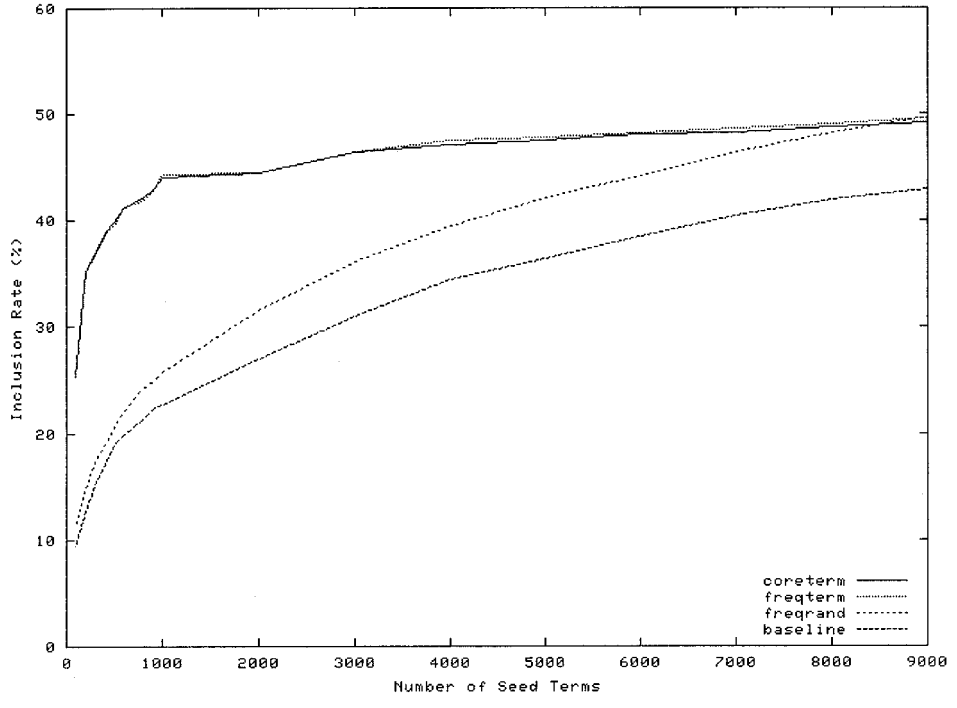


FIG. 3. Curves showing the top one inclusion rates obtained using four different term sets as the seed term set for categorization based on 85 subcategories.

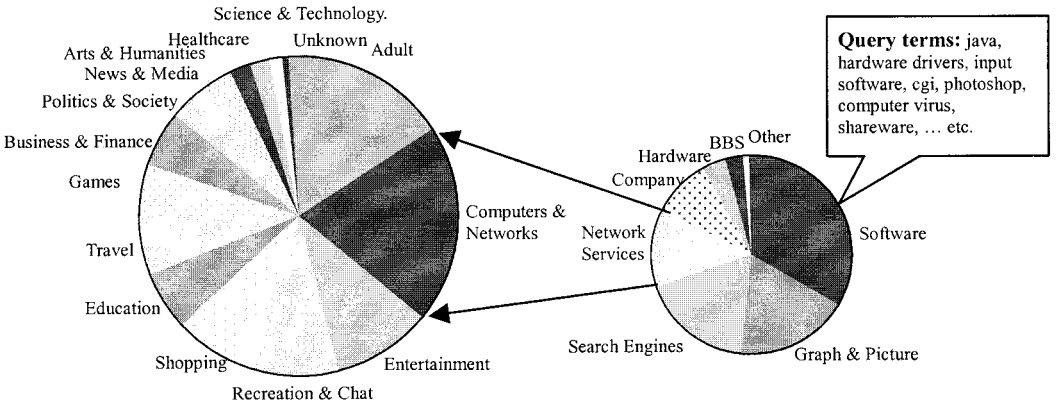


FIG. 4. Sample distributions of subject categories of Web users' search interests.

imental results support our previous assumption that core terms perform better than high-frequency terms in the categorization process. However, it is assumed better performance would be achieved if longer time lags are considered.

Observations of Users' Search Interests

An observation system integrating both the autocategorization approach and human analysis was constructed to allow a systematic way to analyze users' search interests. Two types of observations were conducted in the initial stage; i.e., the distributions of popular search interests in terms of various subject categories, and changes in search interests within a certain period of time.

Subject Distributions of Users' Search Interests

Users' search interests in various subject categories were observed by the subject distributions of the top 20K query terms from D-1998. We used the 9,709 core terms as the feature set and autocategorized the other 10,291 top noncore terms from D-1998. Each test term was automatically assigned to at most five categories depending on its confidence values, as described before. Compared with over 400 man-hours needed to categorize the 9,709 core terms, it took only 1 hour to categorize the 10,291 noncore terms using the proposed approach. The time spent could be reduced to a few minutes if the required computations were performed tightly coupled with the search engine itself. Meanwhile, to ensure accuracy in categorization, human analysts could check the results provided by the autocategorization process. A human analyst was estimated to be able to inspect 500 terms/hour instead of 45 terms/hour without the aid of the categorization process. In total, it took about 21 hours to process the whole test term set instead of hundreds of hours needed by human analysts.

After the above process was finished, the subject distributions of the top 20K query terms were obtained, which represented fully 81% of the total number of query terms in D-1998 as mentioned. The observation system provided the distributions of each subject category and the total frequencies of its query terms along with changes in query term usage. However, when a query term is classified twice, it may have a certain effect on the frequency distribution of each subject category. In the test log, there was less than 10% of the queries were multicategorized. For each query term with multiple categories, its frequency contributed to each of these categories was then estimated as the total occurrence frequency divided by the number of the categories. However, refinement of the method is necessary for further research.

Figure 4 shows a sample illustrating users' search interests in terms of major categories and subcategories. Among the 15 major categories, the Computers & Networks category had the highest percentage, 19.9%, followed by Adult, 16.9%, Entertainment, 9.8%, Recreation & Chat, 8.8%, Shopping, 8.1%, Education, 6.1%, Travel, 5.8%, Games, 5.7%, Business & Finance, 5.4%, Politics & Society, 4.2%, News & Media, 3.1%, Arts & Humanities, 2.3%, Healthcare, 1.8%, and Science & Technology, 1.4%. Each major category was also decomposed into related subcategories. The Computers & Networks category was subdivided into eight subcategories as showed in the figure. The distributions for the major categories were found to be similar to those observed by Ross and Wolfram's study (2000) based on the analysis of Excite's logs. Although different approaches and test logs were used, it is interesting that Web users shared many similar search interests despite of many cultural differences. Furthermore, observation was extended to the query terms in each subject category and their particular topics of interests. For example, the Software subcategory was the most popular subcategory in the Computers & Networks category. The composed query terms may indicate related topics of interests such as "input software," "java," "hardware drivers," etc. Observing the distributions of subject categories is obviously valuable for determining users' popular search interests, and for collecting a large number of up-to-date related query terms for more in-depth subject analysis of the topics for search.

Changes in Users' Search Interests

The observation process is found also efficient in observing trends of the search interests. Possible causes for changes can be then by analyzing the composed related query terms in each category. For the analysis, we used the O-2000 as the test set. It was then categorized using the proposed autocategorization approach and inspected by the human analysts.

The experimental results are illustrated in Figure 5a and Figure 5b, which show the trends of 14 major subject categories (excluding the Unknown category) over a 12-month time frame. The total frequencies of each subject category in different months provide information about their shifts in popularity. In all, although the rank of each category was the same as that in D-1998, each category has its own trend of distribution with time. To illustrate, the Computers & Networks category steadily increased in popularity, which indicated that the need for computer-related information remained strong. The Adult category obviously decreased in popularity with time, and the causes for changes were many, such as the search engine had implemented sensitive query filters starting from February and sources for accessing such information had changed. The Education category increased sharply in popularity in July, which is very reasonable because it was the season for taking various examinations in Taiwan. The Science & Technology category increased in popularity, and the possible reasons came from the searches for report writing materials by students after our examining its composed query terms. The Shopping category did not change much in popularity, but the composed query terms showed different contents for search, such as the shifts of shopping interests for certain products.

Similar observations can be made regarding the 85 subcategories. Taking the frequency distribution of the eight subcategories in the Computers & Networks category as an example, many categories remained popular, like Software; some categories grew in popularity, like Company; some diminished gradually in popularity, like BBS. Meanwhile, we may further analyze particular subject categories to learn more about what users searched in that category. For example, in the Entertainment category, users searched for new movie stars; in the Business & Finance category they searched for new IPO companies; and in the Science & Technology, they searched for in-depth academic information, etc.

Additionally, changes in search interests can be observed by analyzing the distributions of query terms in each subject category. For example, the query term "mp3" remained high in frequency, indicating its continuous popularity. On the other hand, "y2k" decreased sharply in frequency because losing of interests. The above two types of search interests are easily detected and helpful in understanding some hot or nonpopular search interests. However, from our observations, query terms like "electronic books" and "theses/dissertation" deserve more attention. The query term "electronic books" increased gradually in popularity which revealed its potential for becoming an important search interest, and the query term "theses/dissertation" remained stable, indicating that it is an important information request regardless of its relatively low frequency. To sum up, the above longitudinal analysis of search interests enable us to understand more about users' information requests and behaviors, and are helpful for various Web IR applications.

Discussion

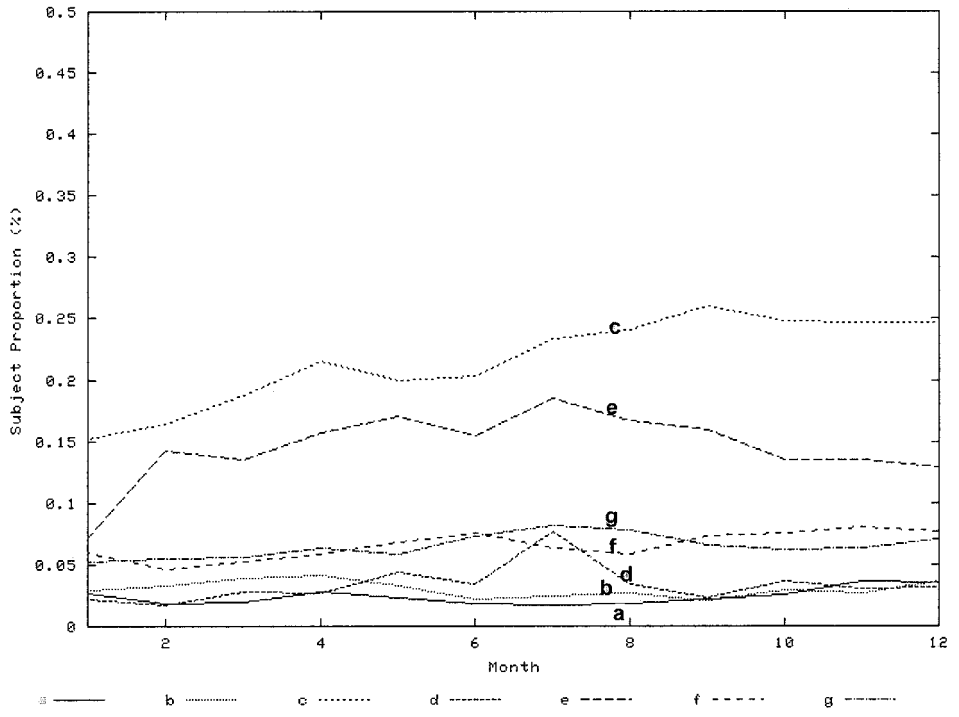
Implications for Various Web IR Applications

The proposed approach can be applied in three areas of Web IR applications: (1) it is valuable for use in the design of Web IR systems, such as implementing query filters; (2) it is useful for Web content organization, such as constructing user-oriented subject controlled vocabularies; and (3) it provides an alternative way to understand users' searching behaviors, such as facilitating Web user studies.

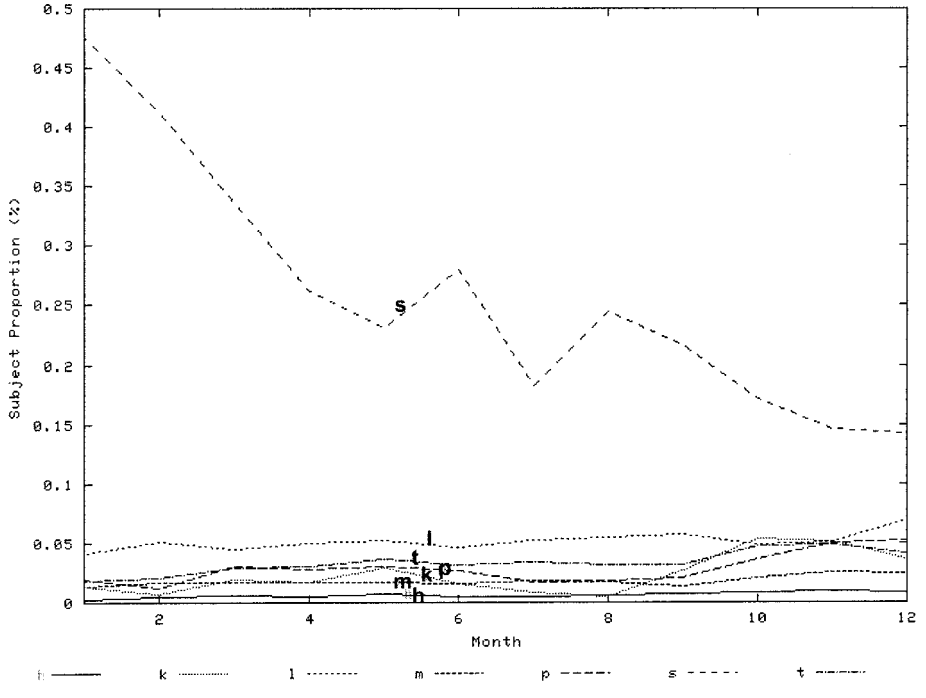
For the design of Web IR systems, the proposed approach can be used to improve Web retrieval in many ways. For example, it is likely to collect a large number of related terms valuable for query expansion or term suggestion in IR systems. Furthermore, it may be used as a basis for developing a thesaurus especially for Web searching (Chuang, Pu, Lu, & Chien, 2000), and also useful for summarizing and ranking search results obtained from search engines. Another immediate application is query filtering. A real-world query filter for filtering out pornography-related terms from Web image searches has been successfully developed in our research. In the filter, more than 20K sensitive query terms have been collected based on the concept behind the proposed approach. Because requests for adult materials can change dynamically, it is more efficient to obtain related terms through an automatic categorization process than through manual analysis.

As for the organization of Web contents, knowledge of users' search interests surely aids the organization of Web directory services. According to a search satisfaction and behavior survey (I-Search, 2000), users often complain that subject directories do not contain the keywords they input, making searching and navigation inconvenient. With the help of the proposed approach, it is possible for human analysts to perform in-depth subject analysis, like synonym term grouping and term relationship analysis, on users' instant queries. Meanwhile, human analysts usually lack domain knowledge, especially within the varied and dynamic Web environment. Our approach at least provides the possible subject domains of queries for further investigation. Therefore, subject-based access tools like Web directory services can benefit from the proposed approach to improve classification structures, to make use of query terms as descriptors for directory names, and to maintain vocabulary consistency between users and human analysts.

Finally, the proposed approach can also be helpful in Web user studies. For example, surveys like questionnaires



a



b

FIG. 5. (a) Sample trends of distributions of subject categories within 1 year Pt. 1. (Legend for the subject categories: a. Arts & Humanities, b. Business & Finance, c. Computers & Networks, d. Education, e. Entertainment, f. Recreation & Chat, g. Games.) (b) Sample trends of distributions of subject categories within 1 year Pt. 2. (Legend for the subject categories: h. Healthcare, k. Science & Technology, l. Shopping, m. News & Media, p. Politics & Society, s. Adult, t. Travel.)

or interviews are usually conducted to understand users' searching behaviors. However, such methods suffer from inefficiency in collecting sufficient amounts of data for analysis, and bias of small groups of users who are surveyed. With the proposed approach, it may provide other aspects of information helpful for the surveys. For example,

according to several user surveys conducted using questionnaires or interviews (GVU, 1998; Yam, 2000), requests for adult materials were rather low, which was different from the results obtained in our research and other related log-based analysis. Both user surveys and log-based analysis are needed to analyze this discrepancy and to achieve a more comprehensive understanding of users' searching behaviors. Furthermore, other observations can be easily made using the proposed approach. For example, it may well serve as a tool to compare common and local search interests revealing cultural differences. For example, according to several log-based-analysis studies, requests for computer-related information are many regardless of the users' geographic locations. However, each log analysis will contain local popular search interests due to different cultural backgrounds; for example, requests for examination materials are high in number in the test logs from Taiwan. The proposed approach provides a new research tool for Web user studies.

Limitations and Future Research

Although many difficulties were encountered when dealing with short queries, the proposed approach has been shown to be efficient and useful for categorizing query terms and observing users' search interests. The obtained results are similar to those of previous related research, and have been even more thoroughly analyzed. With such basis, various research topics for Web IR applications have been presented in the previous section. However, there are some limitations that need to be considered. First, 85 subcategories are rather broad in describing all aspects of users' search interests, more thorough analysis about the topics existing in these broad subject areas are necessary for more effective Web IR applications. Second, accuracy in autocategorization depends on the quality of the documents retrieved from search engines, and should be continually improved. Finally, a more comprehensive understanding of users' search interests can be obtained if some additional data sets can be collected and utilized because query logs are more sparse data for knowing users' information requests. Resources like users' query sessions and click streams can provide much contextual information and reveal more about how users search, browse, and access information on the Web. For example, query sessions provide contextual information, which may help us discover more meaningful relationships among query terms, and click streams may help us disambiguate the subject contents of highly relevant documents.

Conclusions

Subject categorization of Web queries is essential for understanding Web searching interests. Categorizing a large number of highly diverse queries presents theoretical and methodological challenges. In this article, we have presented a feasible and efficient approach incorporating real-world search engines to explore Web users' search interests. The experimental results demonstrate that the approach is scalable and adaptable when used to categorize query terms into predefined subject taxonomy within the dynamic Web environment. The approach has also shown its potential for use in various Web IR applications, and provides a basis for future Web searching studies.

References

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval. New York: ACM.

Carlyle, A. (1989). Matching LCSH and user vocabulary in the library catalog. Cataloging & Classification Quarterly, 10(1/2), 37–63.

Chan, L.M. (1994). Cataloging and classification: An Introduction (2nd ed.). New York: McGraw-Hill.

Chien, L.-F., & Pu, H.-T. (1996). Important issues on Chinese information retrieval. Computational Linguistics and Chinese Language Processing, 1(1), 205–221.

Chuang, S.-l., Pu, H.-T., Lu, W.-H., & Chien, L.-F. (2000). Auto-construction of a live thesaurus from search term logs for interactive Web search. (poster). ACM SIGIR 2000. Athens: Greece.

Drabenstott, K.M., & Vizine-Goetz, D. (1994). Using subject headings for online retrieval. San Diego, CA: Academic Press.

Drabenstott, K.M., & Weller, M.S. (1996). Failure analysis of subject searches in a test of a new design for subject access to online catalogs. Journal of the American Society for Information Science, 47(7), 519–537.

Goller, C., Loning, J., Will, T., & Wolff, W. (2000). Automatic document classification: A thorough evaluation of various methods. IEEE Intelligent Systems, 14(1), 75–77.

GVU Center, College of Computing, Georgia Institute of Technology. (1998). GVU WWW user survey [On-Line]. Available: http://www.cc.gatech.edu/gvu/user_surveys.

He, S. (1999). Chinese search engines for retrieving Chinese information on the Internet: Search capabilities, retrieval performances and evaluation criteria. In C.-c. Chen (Ed.), IT and global digital library development (pp. 171–182). West Newton, MA: MicroUse Information.

Hildreth, C.R. (1985). Monitoring and analyzing online catalog user activity. LS/2000 Communique (pp. 3–6).

Hoelscher, C. (1998). How Internet experts search for information on the Web. WebNet98—World Conference of the WWW, Internet & Intranet, Orlando, FL.

I-Search (2000). Search satisfaction survey [On-line]. Available: http://www.searchenginewatch.com/sereport/00/12-isearch.html.

Jansen, B.J., & Pooch, U. (2001). A review of web searching studies and a framework for future research. Journal of the American Society for Information Science, 52(3), 235–246.

Jansen, B.J., Spink, A., & Pfaff, A. (2000). Linguistic aspects of Web queries. Proceedings of the American Society for Information Science 2000, Chicago, IL.

Jansen, B.J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the web. Information Processing and Management, 36(2), 207–227.

Kaske, S. (1993). Research methodologies and transaction log analysis: Issues, questions, and a proposed model. Library Hi Tech, 11(2), 79–86.

Larson, R.R. (1992). Experiments in automatic Library of Congress classification. Journal of the American Society for Information Science, 43(2), 130–148.

Leonard, L.E. (1977). Inter-indexer consistency studies, 1954–1975: A review of the literature and summary of study results. Technical Report, University of Illinois, Graduate School of Library Science, Champaign, IL.

Lighthouse. (2001). Evaluation of Chinese search engines [On-line]. Available: http://www.haiyan.com/steelk/navigator/b5index.htm (in Chinese).

Openfind. (2001). News announcement on new features of Openfind services [On-line]. Available: http://www.openfind.com.tw/ (in Chinese).

Pu, H.-T., Chuang, S.-l., & Yang, C. (2000). Auto-categorization of search terms toward understanding Web users' information needs. ICADL 2000—International Conference on Asian Digital Libraries. Soeul, Korea.
Pu, H.-T., Chuang, S.-l., & Yang, C. (2001). Exploration of Web users' search interests through automatic subject categorization of query terms. ASIST 2001 Annual Meeting.
Ross, N., & Wolfram, D. (2000). End user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine. Journal of the American Society for Information Science, 51(10), 949–958.
Salton, G. (1989). Automatic text processing: The transformation, analysis, and retrieval of information by computer. Reading, MA. Addison-Wesley.
Salton, G., & McGill, M. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.
Saracevic, T., & Kantor, P.B. (1997). Studying the value of library and information services. Part II: Methodology and taxonomy. Journal of the American Society for Information Science, 48(6), 543–563.
Schwartz, C. (1981). Automatic classification of retrieved sets in online database searching. ASIS 44th Annual Meeting.
Shafer, K., Subramanian, S., & Fausey, J. (1999). Measures for evaluating automatic subject assignment of electronic resources. Dublin, OH: OCLC. [On-line]. Available: http://orc.rsch.oclc.org:6109/measures.html.
Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large Web search engine query log. SIGIR Forum, 33(1), 6–12.
Soergel, D. (1994). Indexing and retrieval performance: The logical evidence. Journal of the American Society for Information Science, 45(8), 589–599.
Spink, A., Wolfram, D., Jansen, B.J., & Saracevic, T. (2001). Searching the Web: The public and their queries. Journal of the American Society for Information Science, 52(3), 260–273.
Strauss, A., & Corbin, J. (1990). Basics of qualitative research: Grounded theory procedures and techniques. Newbury Park, CA: Sage Publications.
Taylor, R.S. (1962). The process of asking questions. American Documentation, 13(4), 391–396.
van Rijsbergen, C.J. (1979). Information retrieval (2nd ed.). London: Butterworths.
Vizine-Goetz, D., & Godby, J. (1996). Library classification schemes and access to electronic collections: Enhancement of the Dewey Decimal Classification with supplemental vocabulary. ASIS 1996 Classification Workshop.
Weinberg, B.H. (1996). Complexity in indexing systems—Abandonment and failure: Implications for organizing the Internet. ASIS 1996 Annual Meeting.
Xu, J.L. (1999). Internet search engines: Real world IR issues and challenges. Conference on Information and Knowledge Management. Kansas City, MO.
Yam.com. (2000). Surveys on uses of the Internet in Taiwan. [On-line]. Available: http://survey.yam.com/ (in Chinese).

Appendix

Procedure Subject-Categorization (t, C, W, D)
{
Input:
  t: the unknown query term
  C: the predefined subject taxonomy based on a number of popular queries
  W: the seed term set which has been manually categorized
  D: the highly ranked Web document set retrieved by term t
Output:
  $C(t)$ the set of subject categories that term t is related to
$C(t) = \varnothing$
obtain W' through scanning D
where W' is a subset of W and each $w \in W'$ should appear in D for every $c \in C$ {
  /* obtain the cooccurring seed term set for category c */
  $W_{t,c} = \varnothing$
  for every $w \in W'$ and w are categorized into category c
  {
    obtain $N(w)$, $N(t)$, and $N(w,t)$ through Web search
      where each $N(x)$ is obtained by sending x as a query to the search engine and taking the number of retrieved Web documents as the value
    if $(N(w,t)/(N(w) + N(t)) >$ threshold1)
    $W_{t,c} = W_{t,c} \cup \{w\}$ /* w is taken as a cooccurring term */
  }
  /* obtain a confidence value by calling rank function R */
  if ($R(W_{t,c}, D) >$ threshold2)
    /* ranking function R is implemented with the procedure call */
    $C(t) = C(t) \cup \{c\}$ /* c is a candidate category */
}
return $C(t)$
}
Procedure R ($W_{t,c}$, D)
{
Input:
  $W_{t,c}$: the cooccurring term set for category c
  D: the highly ranked Web document set retrieved by term t
Output:
  R: the categorization confidence value
$R = 0$
for every $w \in W_{t,c}$ {
  obtain D_w, $D_{w,t}$ by scanning the surrogate of each document in D
  count f_w in D_w
  count $N(D_w)$, $N(D_{w,t})$
  $R = R + N(D_{w,t}) * f_w/N(D_w)$
}
return R
}