圖書資訊學研究 1：1 (December 2006)：1-22

Web Relevant Term Suggestion Using Log-based and Text-based Approaches

Hsiao-Tieh Pu
Graduate Institute of Library and Information Studies,
National Taiwan Normal University

Hsin-Chen Chiao
Institute of Information Science, Academia Sinica

【Abstract】

The study attempts to integrate log-based and text-based methods to obtain appropriate relevant terms for web search. The merits and limitations of using these two types of resources for term suggestion are also discussed. Further, the study proposes to cluster the relevant terms extracted into certain concept clusters hierarchically, which allows users to browse the terms in a more intuitive and meaningful way. The integrated model will enhance effectiveness of current web information retrieval systems and benefit further research.

[Keyword]
Web Information Retrieval, Relevant Term Suggestion, Log Analysis

Introduction

With the increase of web resources, users rely on search engines to retrieve information. However, the overwhelmingly large number of search results by any given query often causes information overload. Further, users are usually in a state of knowing the problem, but not what would resolve the problem when seeking information (Belkin, Oddy, & Brooks, 1982). It is hard for a web user to express information needs to a search engine using only just a few keywords. In addition, most users are not familiar either with the system functions or the indexing structures, and let along to utilize them well. According to several web user studies (Silverstein, et al., 1998; Jansen, Spink, & Saracevic, 2000; Pu, Chuang, & Yang, 2002), web users often input short queries and use less retrieval functions. Besides, it would not be possible for web users to predict what words might be appearing on the pages they need. In sum, it is not easy to formulate an appropriate query to retrieve complete and accurate search results in web searching.

To remedy the above problems of information overload, cognitive insufficiency and lack of retrieval skills, there are increasing research interests in applying interactive information retrieval techniques (Harman, 1992; Belkin, 2001). Major concern of these studies is to develop Information Retrieval (IR) systems containing iterative and interactive functions to enhance retrieval effectiveness, such as relevance feedback, query expansion, and term suggestion, etc. Relevance feedback provides a form of query-free retrieval where documents are retrieved according to a measure of equivalence to a given document indicated by a user (Salton & Buckley, 1990). On the other hand, query expansion is the process of adding related search terms to a user's search. The additional terms may be taken from a thesaurus or from documents that a user has specified as being relevant (Xu & Croft, 1996). Term suggestion has the same process like query expansion, i.e., systems match a user's query with the terms either from a manual-made controlled vocabulary or an automatic co-occurrence word list; difference is that in term suggestion, systems would allow users to select terms instead of automatically adding terms to the query. Through the above techniques, a web user may interact more with a system. However, facing large amounts of data on the

web, more query formulation tools are needed. Various search engines now provide similar features, such as Google's Similar Pages and Suggest functions (http://google.com/), Vivisimo's clustering of search results (http://vivisimo.com/), Grokker's visual display of search results (http://grokker.com/), and Openfinds News Term Suggestion (http://www.openfind.com.tw/), etc.

The above studies or systems were mostly based on text-based approaches, i.e., using documents such as pages retrieved as a source for relevant term extraction. From a user's perspective, these index terms may not be consistent or similar to the queries, and hence, vocabulary mismatch or search failure may occur frequently (Fidel, 1991; Spink, 1995). Such difficulties have led researchers to investigate another alternative, i.e., so-called log-based approaches in this paper (Wen, Nie, & Zhang, 2001; Beeferman & Berger, 2000; Chuang & Chien, 2002; Huang, Chien, & Oyang, 2003). Basically, such approaches mainly use query logs as the base for term extraction. Query logs from web IR systems like search engines are considered to be the foremost source of unobtrusive data about users' requests. The log-based approaches normally employ users' query terms as sources for the key terms selection, but there are different methods to determine term associations and corpus sources. Detail discussion will be included in the next section.

Following previous researches of the authors' co-research group in Web Knowledge Discovery (WKD) lab in Academia Sinica, Taiwan (http://wkd.iis.sinica.edu.tw/), several efficient techniques have been developed, including a query classification technique to find users' search interests (Pu, Chuang, & Yang, 2002), a query-session-based approach to extracting relevant terms for queries (Huang, Chien, & Oyang, 2003), and an effective algorithm for clustering relevant queries into clusters (Chuang & Chien, 2002). With the above basis, this paper further presents an integrated approach to combine these techniques to investigate its potential for web term suggestion applications. Different from the above text-based or log-based approaches, the study proposed an approach combining both the text-based method with the query-session-based method to extracting relevant terms for each search term. Using this combined method, the relevant terms of each search term can be extracted via two different manners. One is to extract the terms that co-occur

in similar query sessions from search engine logs, and the other from that in the same retrieved web pages returned from search engines. Many terms that are conceptually related to the search request but do not frequently co-occur in documents, therefore, can be extracted. Since the extracted relevant terms might be many and contain different meanings, to make them more understandable and well organized, the proposed approach also used a query clustering method to further organize the relevant terms into classes. The clustering method mainly exploits the highly ranked documents retrieved from the web by a search term as the additional feature source. It significantly reduces the lack of sufficient features in characterizing query terms. Some initial experiments have been conducted to test the feasibility of the proposed approach, and implications for the applications are also discussed.

Related Studies

Web Term Extraction and Thesaurus Construction

It has been well recognized that a thesaurus is crucial for representing vocabulary knowledge and helping users to reformulate queries in IR systems (Jing & Croft, 1994). One of the important functions of a thesaurus is to provide term associations information for IR systems. Conventional approaches for estimating term associations are mostly text-based and collection-dependent (Smadja, 1993; Schutze & Pedersen, 1997), i.e., the thesauri were constructed based on the document collections used. These approaches rely on a term extraction technique to extract key terms from a set of domain-specific documents, and an association estimation technique that determines term associations mainly based on their co-occurrence in the same documents. Such approaches need to deal with term segmentation difficulties and interference of irrelevant documents. Moreover, it is not easy to extract terms that are conceptually related to the intended purpose of a search term but not frequently co-occurred with it in documents (Wan, et al., 1997; Lin, 1998). Concerning the dynamic and diverse nature of the web, it is not feasible to apply the text-based approaches to extract relevant terms for web term suggestion. Since the number of terms in web documents is huge and their meanings are dynamically changed, determining a set of representative and

up-to-date key terms is the first challenge. Further, it is necessary to design scalable techniques for estimating key terms associations, which allow adaptability to changes of both web documents and users' usages of the vocabularies.

As mentioned, the above difficulties have led researchers to investigate another alternative, i.e., so-called log-based approaches in this paper. Such approaches normally employ users' query terms as a source for the term selection, and there are different methods to determine term associations and corpus sources. For example, Beeferman & Berger (2000) proposed a relevant query clustering method based on "click-through data" to discover correlations between queries and clicked URLs. Query terms with more common clicked URLs were taken as similar and grouped. Without merely using the clicked URLs, Wen, Nie, & Zhang (2001) developed a similar method to combine the indexed terms from the clicked pages to estimate the similarity between queries, which achieved better performance. However, the number of distinct URLs is huge and may cause many similar queries not being able to be grouped together due to the lack of common clicked URLs. With fewer clicked URLs as the feature set, it is therefore difficult to find similar terms, especially for those new query terms or queries with less usage. As discussed in the previous section, the study proposed an approach combining both the text-based method with the query-session-based method to extract relevant terms for each search term.

Web Interactive Information Retrieval

Reflecting from the development of search engines for decades, web IR of early periods focused on subject directory service like Yahoo!, then full-text search like AltaVista, and the linking and ranking technique like Google. Though search functions of the above systems are simple and powerful, very few interactive functions were developed. Once AltaVista (*http://altavista.com/*) provided term suggestion function, but now such function was mostly limited to specific domains (such as news services) or applications (such as keyword advertising). Recently, there are systems that offer more interactive functions, such as Vivisimo and Grokker, as briefly discussed below.

Openfind's News Relevant Term Suggestion

Openfind provides relevant term suggestion for any given query in its News search services. It estimated the similarity between a user's query and terms extracted from the news collected from the web, then list candidate terms for a user to select. Terms extracted would be dynamically changed when contents of the news changed accordingly. Such feature allows a web user to associate related terms and concepts, and also prevent from missing topics of important news. For example, at the time of August 2006, when a user input a query like "荷蘭" (Netherland), the system would list terms like "天皇" (Emperor), "渡假" (vacation), "足球賽" (World Cup), "法國" (France), "西班牙" (Spain), etc. Investigating concepts behind these terms may include foreign affairs (Japan Emperor visiting Netherland) and FIFA World Cup, etc. The user may then further narrow down or expand the search using these suggested terms. In addition, there are keyword suggestion tools for web design or search engine position optimization. These tools provide similar function regardless of their initial purposes may not be for end-user searching. Examples are like Google AdWords Keyword Tool or Overture View Bid Tool (Agarwal, 2006). Basically, these tools were developed in similar process, only the sources for extraction may be different, i.e., systems allow a user to input keywords, and search different vocabulary sources, such as Google Directory or goto.com search engine for relevant keywords used by recent searchers.

Vivisimo's Search Results Clustering

There is an analogy that searching on the web is more like finding the needle in a haystack. Though many search engines provide advanced search functions, most users lack the search skills or tactics to fully utilize them. Better tools are needed to assist users' query formulation. Valdes-Perez, co-founder of Vivisimo Inc. describes his search engine as "If the Internet is a giant bookstore in which all the books are piled randomly on the floor, then Vivisimo is like a superfast librarian who can instantly arrange the titles on shelves in a way that makes sense." Also "consider it a 21st century Dewey Decimal System designed to fight information overload. But unlike libraries, Vivisimo doesn't use predefined categories. Its software determines them on

the fly, depending on the search results" (Library Stuff, 2004). Basically, Vivisimo analyzes search results and automatically sorts them into categories that may present far more information than the typical textual list like Google.

Grokker's Visual Display of Search Results

Grokker has similar process like Vivisimo, and adds more visual display feature. In other words, it not only sorts search results into categories but also maps the results in a more holistic way, showing each category as a colorful circle. Within each circle, subcategories appear as more circles that can be clicked on and zoomed in on. The power of visual display is that a user may be more easily filter out un-wanted foci of search results. For example, a user was looking for accommodation in France, and entered a search like "Paris Hilton." Google recognizes this as a search in the category of "Regional-Europe-Travel and Tourism-Lodging-Hotels" but still produces page after page with links about celebrity socialite Paris Hilton and her exploits (CNN, 2004). However, the resulting circle shows all the possible categories of information the web, including reviews, maps and online booking sites for the Hilton Hotel in Paris, which are all but might be buried in the Google rankings.

Overview of the Proposed Approach

One of the important features of the proposed approach is utilizing query-session logs to extract relevant terms and organizing them into concept clusters. It is based on the assumption that some related terms, such as synonyms and alternative terms, may appear in a set of query sessions with similar requests. Table 1 lists an example showing the idea of the proposed approach. The example is extracted from a query session log containing 615,634 query sessions for web search (Huang, Chien, & Oyang, 2002). It contains a number of real Chinese query sessions with the appearance of the term "圖片" (equivalent term of "picture" in English). In these sessions, the term seems related to a number of relevant terms, and can be categorized into several search subjects, including cartoon photos, screen savers, online picture banks, greeting cards, etc. It would be very useful for web term suggestion application, in which these subject classes are automatically clustered and important relevant terms being categorized into them.

Table 1. Example query sessions containing relevant terms with the search term "圖片" ("picture" in English)

Chinese Query Session	English Translation
卡通圖片__kitty	cartoon picture __ kitty
可愛圖片__卡通	lovable picture __ cartoon
母親節圖片__母親	mother's day picture __ mother
母親節圖片__母親節	mother's day picture __ mother's day
母親節圖片__母親節卡片	mother's day picture __ mother's day greeting card
母親節圖片__趴趴熊	mother's day picture __ bear
母親節圖片__康乃馨	mother's day picture __ carnation
圖片__kitty	Picture __ kitty
圖片__卡通	Picture __ cartoon
圖片__卡通圖片	Picture __ cartoon picture
圖片__布丁狗	Picture __ pudding dog
圖片__母親節	Picture __ mother's day
圖片__母親節卡片	Picture __ mother's day greeting card
圖片__母親節圖片	Picture __ mother's day picture
圖片__皮卡丘	Picture __ picachu
圖片__有趣	Picture __ funny
圖片__狗	Picture __ dog
圖片__趴趴熊	Picture __ bear
圖片__桌面	Picture __ theme
圖片__桌面王	picture __ themeking
圖片__神奇寶貝	picture __ pokemon
圖片__動物	picture __ animal
圖片__動畫	picture __ animation
圖片__康乃馨	picture __ carnation
圖片__遊戲	picture __ game
圖片__遊戲下載	picture __ game download
圖片__圖	picture __ graph
圖片__圖片下載	picture __ picture download
圖片__圖庫	picture __ picture bank
圖片__圖檔	picture __ picture file
圖片__漫畫	picture __ comic

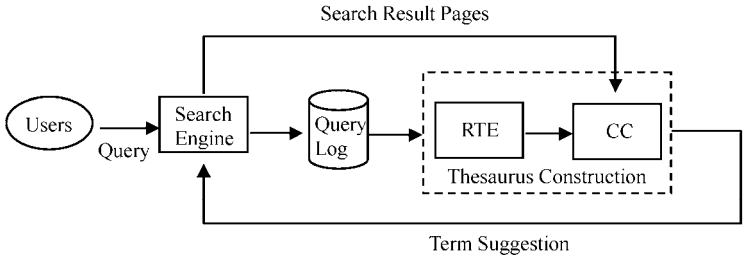

Figure 1. An abstract diagram of the proposed approach.

To exploit web users' queries and web documents as the corpus source, the proposed approach is designed to integrate with a search engine, as shown in Figure 1. It consists of two kernel modules, i.e., Relevant Term Extraction (RTE) and Concept Clustering (CC) modules. Since the web is dynamic, terms suggested with the proposed approach is also dynamic and changed with the query usages and web pages currency. To make the approach more adaptable, each search term is considered a potentail key term in this study. Its relevant terms and corresponding concept classes are generated depending on a dynamic request, i.e., once a search term is submitted, the RTE module would extract its corresponding relevant terms and the CC module would organize the relevant terms into concept classes. The obtained clustered relevant terms and concept classes, therefore, can be exploited by search engines for various applications, such as term suggestion or search results browsing. Details of RTE and CC are described in the following sections.

RTE (Relevant Term Extraction)

Suppose that t is a search term to extract relevant terms, D a set of highly ranked documents retrieved by t, and L a collected log of users' query sessions. The relevant term extraction module is to perform relevance analysis between t and all candidate terms obtained from D and L, and extract relevant terms for t. The proposed approach combines a log-based and a text-based method to extract relevant terms as described below.

The Log-based Method

The log-based method finds the relevant terms of a given search term from similar query sessions containing the search term of concern. The query

log that contains the query transactions submitted to search engine(s) is first partitioned into a number of query sessions. The log contains timestamps at which the query transaction was submitted to and the IP addresses of the machine at which the query transactions were submitted. The user query analysis conducted by Silverstein et al. (1998) revealed that queries for a single information need could be clustered in time. Based on this observation, we therefore use a time threshold as delimiter to segment query sessions from the query log used. With the segmented query sessions, next is to perform relevance analysis among query terms and then extract relevant terms for each given subject term. The relevance analysis in this paper is based on a term co-occurrence matrix defined as follows:

Definition: *The co-occurrence matrix* ***C*** *of the distinct query terms in a query log, denoted by* $\boldsymbol{q_1, q_2, ..., q_n}$*, is an* ***n*** *by* ***n*** *matrix with* $\boldsymbol{C_{i,j}}$ *= the number of query sessions containing both query terms* $\boldsymbol{q_i}$ *and* $\boldsymbol{q_j}$*.* ($\boldsymbol{f_i = C_{i,i}}$)

Three notable similarity estimation functions are applied in our relevant analysis - the Jaccard measure, dependence measure and cosine measure, which are defined below:

$$\text{Jaccard}(q_u, q_v) = \frac{C_{u,v}}{f_u + f_v - C_{u,v}}, \tag{1}$$

$$\text{Dependence}(q_u, q_v) = \frac{C_{u,v}}{\min(f_u, f_v)}, \tag{2}$$

$$\cos(q_u, q_v) = \frac{\sum_{\forall q_j \in Q} (C_{u,j} \cdot C_{v,j})}{\sqrt{\sum_{\forall q_j \in Q} C_{u,j}^2} \cdot \sqrt{\sum_{\forall q_j \in Q} C_{v,j}^2}}, \tag{3}$$

According to our experiments, the Jaccard measure is useful when q_u and q_v contain high co-occurrence frequencies. The dependence measure is a degenerated form of the Jaccard measure. When the difference between the frequencies of q_u and q_v is large, the larger item from the original formula will be eliminated in order to deal with low-frequency query terms.

Figure 2 shows the log-based RTE algorithm implemented. Note that the algorithm does not apply a global formula to all cases due to our empirical tests. When the co-occurrence value $C_{u,v}$ of q_u and q_v is between $\sqrt[4]{f_u}$ and $\sqrt{f_u}$, either the dependence measure or the Jaccard measure will be applied. If the difference between the frequencies of q_u and q_v is large, we apply the

```
function compute_relevant_term_set(q_u, Q, C, R)
{
  Input:
    q_u: the query term of concern
    Q: the set of all query terms in the log
    C: the co-occurrence matrix
  Output:
    R: relevant term set
  R =∅
  For every q_v in Q {
    if( C_{u,v} ≥ √f_u )
      R = R ∪ {q_v}
    else if( ∜f_u ≤ C_{c,v} ≤ √f_u )
      if( f_u ≫ f_v or f_u ≪ f_v )
        if( Dependence( q_u, q_v ) > threshold_1 )
        then R = R ∪ {q_v}
      else
        if( Jaccard( q_u, q_v ) > threshold_2 )
        then R = R ∪ {q_v}
    else if( C_{u,v} ≤ ∜f_u ) {
  If( cos( q_u, q_v ) > threshold_3 )
      then R = R ∪ {q_v}
  }
  return R;
}
```

Figure 2. The algorithm for computing the relevant term set.

dependence measure; otherwise, we apply the Jaccard measure. Queries that frequently co-occur with query q_u, i.e., $C_{u,v} \leq \sqrt[4]{f_u}$, are taken as being relevant to query q_u. For queries which seldom co-occur with query q_u, i.e., $C_{u,v} \leq \sqrt[4]{f_u}$, we apply the cosine measure, in order to obtain more less-relevant query terms. The required parameter values, including $\sqrt[4]{f_u}$, $\sqrt{f_u}$ and *threshold1~3*, have all been obtained through experiments.

The Text-based Method

An integrated approach that combines both log-based and text-based methods is found more efficient and useful, since either method has some problems as mentioned. Therefore, in the RTE module, a basic text-based

method is also implemented for relevant term extraction. Our basic idea in using the text-based method is to extract co-occurring key terms for each test search term *t* from a set of Web documents *D*, which are the high-ranked documents retrieved using *t* as query for a real Web search engine. Also, to reduce the term extraction difficulties, the RTE employs the terms appearing in the query log as the basic key terms *W*.

Term relevance is measured based on a conventional mutual-information-based association measure between each subject term *t* and its candidate term *w*. *w* is assumed to appear in *D* and belong to *W*. Also, this satisfies the condition that $N(w,t)/(N(w)+N(t)) >$ a threshold value, in which $N(w)$, $N(t)$, and $N(w,t)$ are the numbers of Web documents in *D* containing term *w*, containing term *t*, or containing both term *w* and *t*, respectively. In our experiences, it is promising to collect up to 100 search result entries for each search term and extracted each entry's title and description as the representation of the corresponding document containing the search term. Since the text-based method is used as a complementary one in this study, term relevance measurement is relatively simple to reduce its computational complexity.

CC (Concept Clustering)

Based on the results obtained from the previous RTE module, the proposed approach attempts to organize related concept classes of a search term from its relevant term set. Clustering short terms into concept classes is a difficult problem given that, unlike long documents, short terms typically don't contain enough information to extract reliable features. A lack of domain-specific corpora to describe short terms is usually the case in practice. In the proposed approach, we incorporate the search result pages returned from search engines into the process of acquiring features for short terms.

Feature Source Collection and Feature Selection

Since a relevant term is often short in length and simple in structure, we take their highly-ranked search-result pages retrieved from the search engine as their feature source to judge the relevance of the relevant terms. Each relevant term is used as a query to search and up to 100 result entries are collected. The title and description of each entry are extracted as the representation of the corresponding document, i.e., as the feature source for the

key terms of concern. After collecting the feature source, we then extract a set of feature terms from it. We adopt the n-gram method and choose bi- and tri-grams extracted from the feature source to represent the feature space of the key terms.

The Concept Clustering Algorithm

There exist many different clustering algorithms. The proposed approach adopts the Hierarchical Agglomerative Clustering (HAC) algorithm. An HAC algorithm groups a set of objects by using their inter-object distance matrix and constructs a binary tree from leaves to root. Each non-leaf node is a cluster merged from two clusters, i.e., each cluster is a set of two clusters. First, each object forms a singleton cluster, and all the singleton clusters C_1, C_2, ..., C_n become the leaves of the binary tree. Second, the closest pair of clusters C_i and C_j are chosen to merge and form a new cluster $C_{n+1} = \{C_i, C_j\}$. The process of the two steps is iterated on the remaining unmerged clusters (including the new formed cluster) until only one unmerged cluster is left.

To execute the HAC algorithm, we first define the inter-object distance. We adopt vector-space model to represent each relevant term as a vector of weights of feature terms, therefore the dimension of the term vector is the number of feature terms. Given n relevant terms, we have n vectors v_1, v_2, ..., v_n. Let T be the set of feature terms, t_j be the j-th feature term in T, and $v_{i,j}$ be the j-th element of the term vector v_i. By adopting the conventional *tf-idf* term weighting scheme, $v_{i,j}$, the weight of a feature term, is defined as:

$$v_{i,j} = (0.5 + 0.5\frac{tf_{i,j}}{max_{t_k \in T} tf_{i,k}}) \log \frac{n}{n_j}$$

where $tf_{i,j}$, the term frequency, is the number occurrences of feature term t_j in the v_i's corresponding feature source, and n_j is the number of relevant terms whose feature source contain feature term t_j. The similarity of a pair of relevant terms is then defined as the cosine measure of the two vectors of relevant terms:

$$sim(v_a, v_b) = \frac{\Sigma_{t_j \in T} v_{a,j} v_{b,j}}{\sqrt{\Sigma_{t_j \in T} v^2_{a,j}} \sqrt{\Sigma_{t_j \in T} v^2_{b,j}}}$$

The distance between a pair of relevant terms is then defined as one minus

the similarity of the pair of key terms:

$$dist(v_a, v_b) = 1 - sim(v_a, v_b)$$

The core of an HAC algorithm is to choose a specific inter-cluster distance function. We used the complete-linkage function as the inter-cluster distance functions:

$$\max_{v_a \in C_i, v_b \in C_j} dist(v_a, v_b)$$

Experiments and Observations

Relevant Term Extraction

To estimate the performance of the two different RTE methods, an experiment has been performed using the query session log shown in Table 2. The query session log used was collected using a proxy server from National Taiwan University, which served 52 organizations including 20 universities and colleges in Northern Taiwan. In order to select a reasonable threshold value, we conducted some analysis on the query log. With extensive experiments, we therefore selected 5 minutes as the time threshold, which was similar to that of Silverstein, et al (1998).

Table 2. Some statistics of the test session log for relevant term extraction.

	All query sessions	Sessions with more than one query
1. Number of obtained query sessions	615,634	160,180
2. Total query terms in the sessions	2,369,282	1,213,226
3. Number of unique query terms	218,362	177,324
4. Average unique query terms per session	1.45	2.75

The log-based method was further performed to extract relevant query terms for 5,366 distinct query terms. It successfully extracted relevant terms for 3,330 of them. On average, 9.28 high-frequency terms, 4.82 medium-frequency terms, and 2.77 low-frequency terms were extracted. The achieved relevance between the query terms and extracted relevant terms exceeded our

expectations.

In order to realize the performance between the log-based and text-based method, some further experiments were performed. The test query set contained 100 queries randomly selected from the 3,330 queries whose relevant query terms have been extracted with the proposed log-based method. For each of the test queries, its text-based relevant terms were also extracted with the text-based method for comparison. The relevant terms extracted with the two different methods were merged and inspected manually to judge the relevance between the test query terms. For each test query term, its relevant term set is defined as all of the relevant terms extracted with the two different methods. In addition to precision and recall rates, we also propose a new measure called average relevancy for those extracted relevant terms. When judging the relevance between the test query terms and the extracted really relevant terms, the volunteers were asked to assign numeric values ranged from 1 to 3 to indicate the degree of the relevance. Value 1 means less relevant and value 3 highly relevant. The average relevancy is then the average relevancy value of the obtained really relevant terms. Table 3 lists the obtained performance based on the average relevance values.

Table 3. Performance comparison between the log-based and the text-based method for relevant term extraction.

	The Text-Based Method	**The Log-Based Method**
Precision	0.26	0.90
Recall	0.76	0.35
Average Relevance	1.73	2.40
Avg. number of Extracted Relevant Terms	22.8	10.6
Avg. number of Extracted Terms	89.5	11.8
Comments	- Low precision - High Recall - Useful for low frequency queries	- High precision - Low recall - More comprehensive - Needs sufficient log

It is clear that the log-based method outperformed the text-based method in terms of the precision rate. The low precision rate of the text-based method

might have resulted from the low relevancy of the retrieved documents and the diverse nature of the Web documents. Also, the log-based method achieved very high average relevancy. Although relevancy could not be evaluated without some subjectivity, we still have strong confidence in the results obtained. This is because the log-based method achieved better average precision rates for almost every test query, and because among the extracted relevant terms, 7.2 terms could not be found using the text-based method. However, the log-based method was not superior in every respect. The number of extracted relevant terms was less than that for the log-based method. An integrated approach that combines both methods is therefore proposed.

The log-based method was found easier to extract highly-associated relevant terms, and explore term relationships such as synonyms and alternative terms. Table 4 illustrates the test query "Sina" (name of a Chinese Web portal http://www.sina.com/) to show differences between two methods. The relevant terms extracted using the text-based method were mostly popular services provided by Sina. No portal competitors in Taiwan such as "Yahoo-Kimo", "Pchome", "Yam", etc were extracted. This result was probably due to the fact that the top relevant pages retrieved from Google were limited to Sina's sites.

Table 4. An example showing the different characteristics of the relevant terms extracted using the log-based and document-based methods, respectively, in which the test query was Sina in Chinese.

The Log-Based Method	The Text-Based Method
Pchome	copy right (in Chinese)
Yahoo	IP right (in Chinese)
Kimo	Chat (in Chinese)
sex	News (in Chinese)
Yam	Nasdaq
Yahoo (in Chinese)	chat room (in Chinese)
tomail	personal finance (in Chinese)
free email (in Chinese)	banking (in Chinese)
Yahoo Taiwan (in Chinese)	company name (in Chinese)
Kimo (in Chinese)	mail box (in Chinese)
Search engine (in Chinese)	translation (in Chinese)
	fortune telling (in Chinese)

This confirms that the text-based method may not extract relevant terms that do not frequently co-occur in the same documents. Yahoo is a typical example; it might occur frequently with Sina in news stories but seldom co-occur in the same Web pages. To extract conceptually related terms, such as the names of competitors is an important merit of the proposed log-based method. In addition, RTE takes less than 1 second for each query using either log-based or text-based methods.

Concept Clustering

A number of experiments have been conducted to test the feasibility and the performance of the proposed approach on different domains of search terms. In this paper we show one of the experiments for illustration. To have a standard basis for performance evaluation, we collected the following experimental data: The category names in the top three levels of the Yahoo! Computer Science (CS) directory were collected. There were 36,177 and 278 category names in the first, second, and third levels, respectively. These category names were short in length and specifically expressed some key concepts in the CS domain; therefore, they could play the roles of typical relevant terms, the terms relevant to "Computer Science", and were used as the major experimental data in this study.

In the experiment, we classified the 177 CS terms at the second level into clusters. We adopted F-measure, a standard evaluation measurement used in document clustering, for the generated clusters. Based on the 36 classes at the first level as the target classes, the F-measure computes the performance of the machine-generated clusters. F measure of cluster j with respect to class i is defined as:

$$F_{i,j} = \frac{2R_{i,j}P_{i,j}}{R_{i,j} + P_{i,j}}$$

where R and P are recall and precision which are defined as $n_{i,j}/n_i$ and $n_{i,j}/n_j$ respectively, in which $n_{i,j}$ is the number of members of class i in cluster j, n_j is the number of members in cluster j, and n_i is the number of members of class i. For an entire hierarchical clustering, the F-measure of any class is the maximum value it attains at any node in the tree, and an overall F-measure is computed by taking the weighted average of all values for the F-measure as

given by the following:

$$F = \sum_i \frac{n_i}{n} \max \{ F_{i,j} \}$$

where the maximum is taken over all clusters at all levels, n is the total number of relevant terms, and n_i is the number of relevant terms in class i.

It was found that, except some terms isolated in singleton clusters, the terms being grouped together were normally relevant in some degree. The F-measure value obtained in the experiment can reach to 0.8324. It means most of the test CS terms were successfully grouped with the terms owning close concepts. The high performance was really out of our expectation and encouraging. To further illustrate the performance of the concept clustering method, the example relevant terms in Table 4 was clustered and the obtained term clusters shown in Figure 3. It would be easy to see that the related concepts of the search term "Sina" are easier understood.

However, clustering is a similarity-driven approach to grouping instances, i.e., it groups instances based on their similarities. In our further work, it is good in finding related or similar terms, but has a limitation in determining the broader or narrower relationship between terms. Thus, determining the different levels of granularity or the subsumption relationship among concepts could not be performed by our current approach. Even so, with the concept clustering method the relevant terms of a search terms are

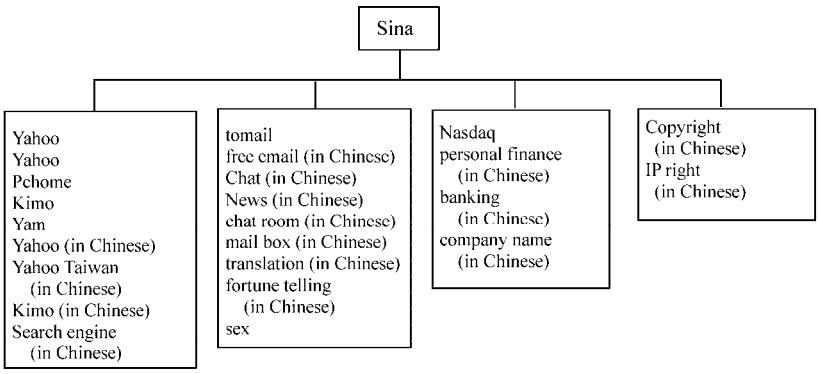

Figure 3. An example showing concept clustering results using the proposed approach.

likely to be further clustered with their concept similarity. The proposed approach, therefore, can benefit the construction of a thesaurus for Web applications.

Conclusion

In this paper, we have presented an approach combining several useful techniques to benefit relevant term suggestion for web IR. The obtained preliminary results have shown the potential of the approach. With the help of terms extracted combining both logs and texts, a web user may utilize it to clarify topics for search, explore unexpected yet highly relevant concepts, and therefore, obtain more satisfactory results. Further through concept clustering, users may browse these terms in a more intuitive and convenient way. With the increase of web content, the experimental results may benefit further related research, such as personalized web search and web thesaurus construction, etc. Based on the experimental results, we are developing a large-scale thesaurus for Chinese web IR. However, a few challenges need to be further explored, for example, to investigate whether the generated concept classes are really helpful for users' search. Meanwhile, the computational complexity of the proposed approach needs to be improved to make the thesaurus more scalable too.

Acknowledgement

This research is sponsored in part by grant NSC 94-2413-H-003 -023 to the first author. Also thanks to the Web Knowledge Discovery Lab in the Institute of Information Science, Academia Sinica, Taiwan for technical support.

References

Agarwal, A. (2006). Online keyword suggestion tools. Retrieved July 10, 2006. from http://labnol.blogspot.com/2006/01/online-keyword-suggestion-tools.htm

Beeferman, D. & Berger, A. (2000). Agglomerative clustering of a search

engine query log. *Proceedings of ACM SIGKDD Int'l Conference*, 407-415.

Belkin, N.J., et.al. (2001). Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval. *Information Processing & Management*, 37(3), 403-434.

Belkin, N.J., Oddy, R.N. & Brooks, H.M. (1982). ASK for information retrieval: part 1: background and theory. *Journal of Documentation*, 38(2), 61-71.

Chuang, Shui-Lung & Chien, Lee-Feng. (2002). Towards automatic generation of query taxonomy: A hierarchical query clustering approach. *IEEE Conference on Data Mining*, 75-82.

CNN (2004). Better search results than Google? Next-generation sites help narrow Internet searches. January 5, 2004. Retrieved July 10, 2006. from http://www.cnn.com/2004/TECH/internet/01/05/seeing.search1.ap/

Fidel, R. (1991). Searcher's selection of search keys. *Journal of the American Society for Information Science*, 42(7), 490-500 & 515-27.

Harman, D. (1992). Relevance feedback and other query modification techniques. In Frakes, W.B. & R. Baeza-Yates (eds). Englewood Cliffs, NJ: Prentice Hall. *Information Retrieval: Data Structures and Algorithms*, 241-63.

Huang, C.K., Chien, Lee-Feng & Oyang, Y.J. (2003). Relevant term suggestion in interactive web search based on contextual information in query session logs. *Journal of the American Society for Information Science & Technology*, 54(7), 638-649.

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), 207-27.

Jing, Y. & Croft, W.B. (1994). An association thesaurus for information retrieval. *UMass Technical Report,* 94-17.

Library Stuff (2004). *Vivisimo like a "superfast librarian"*. January 2, 2004. Retrieved July 10, 2006. from *http://www.librarystuff.net/2004/01/vivisimo-like-superfast-librarian.html*

Lin, Dekang (1998). Automatic retrieval and clustering of similar words. *COLING-ACL'98*, 768-774.

Pu, Hsiao-Tieh, Chuang, Shui-Lung & Yang, C. (2002). Subject categorization of query terms for exploring Web users' search interests. *Journal of the*

American Society for Information Science & Technology, 53 (8), 617-630.

Salton, G. & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science & Technology*, 41(4), 288-297.

Schutze, H. & Pedersen, J.O. (1997). A coocurrence-based thesaurus and two applications to information retrieval. *Information Processing & Management*, 33, 307-318.

Silverstein, C., et al. (1998). Analysis of a very large AltaVista query log. *SRC Technical Report* 1998-014, COMPAQ Systems Research Center.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, 19, 143-177.

Spink, A. (1995). Term relevance feedback and mediated database searching: implications for information retrieval practice and systems design. *Information Processing & Management*, 31(2), 161-171.

Wan, T.L., et al. (1997). Experiments with automatic indexing and a relational thesaurus in a Chinese information retrieval system. *Journal of the American Society for Information Science*, 48(12), 1068-1096.

Wen, J.-R., Nie, J.-Y. & Zhang, H.J. (2001). Clustering user queries of a search engine. *Proceedings of the 10th International World Wide Web Conference*, 162-168.

Xu, J. & Croft, B. (1996). Query expansion using local and global document analysis. *Proceedings of the 19th annual international ACM SIGIR,* 4-11.