Live Thesaurus Construction for Interactive Voice-based Web Search

Shui-Lung Chuang\textsuperscript{1}, Hsiao-Tieh Pu\textsuperscript{2,3}, Wen-Hsiang Lu\textsuperscript{1,4}, Lee-Feng Chien\textsuperscript{1}

1. Institute of Information Science, Academia Sinica
2. Institute of Information Management, Chiao Tung University
3. Department of Library & Information Studies, Shih Hsin University
4. Department of Computer Science and Information Engineering, Chiao Tung University
Taiwan, China

(slchuang, whlu, lfchien) @iis.sinica.edu.tw, htpu@cc.shu.edu.tw

ABSTRACT

Since Web users' queries are often too short, an accurate and interactive speech interface is believed very helpful, especially for WAP-based Web search engines. To provide high accurate speech recognition and effective interactive search, a rigid and live web thesaurus that contains users' search terms plus a set of relations between their associated terms is highly in demand. The purpose of this paper is intended to present a log-based approach for live thesaurus construction. Based on the live thesaurus, certain kinds of users' information behaviors could be characterized and a more effective voice-based search engine could be developed.

1. INTRODUCTION

Since Web users' queries are often too short to contain sufficient keyterms to discriminate ambiguous documents in the process of information retrieval [5], a high-performance Web search engine needs to provide efficient interactive search and/or term suggestion techniques [3]. An accurate and interactive speech interface is believed very helpful, especially for WAP-based search engines. To provide high accurate speech recognition and effective interactive search [2], a rigid web thesaurus that contains users' search terms plus a set of relations between their associated terms is highly in demand [1]. The purpose of this paper is intended to introduce an innovative work called live thesaurus construction. Different from previous work on automatic thesaurus construction, the research attempts to construct a live thesaurus from search term logs of real-world search engines rather than through term extraction from online documents. In addition, the information to be collected in such a thesaurus is mainly the representative search terms with their associated terms and corresponding subject categories.

The proposed approach is being developing and testing with two primarily Chinese search engine logs in Taiwan: the Dreamer's log and GAIS's log. The Dreamer's log contained 228,566 distinct search terms with a total frequency of 2,184,256 within a period of over 3 months in 1998 (D-1998), and the GAIS's contained 114,182 distinct queries with a total frequency of 475,564 within a period of 2 weeks in 1999 (G-1999). The query logs contain a series of requests, and each request includes a search term, and the corresponding timestamp (when the query was submitted) etc.

The proposed approach is a well integration of human and machine's efforts. It is a three-step process: search term log analysis, new term categorization and similar term clustering for the construction of a live thesaurus. The human efforts mainly involve in the analysis of search term logs, such as extracting of "Core Terms", structuring subject taxonomies for these terms and classifying each of them appropriate subject categories. There are twenty thousands terms which form 81% of the total number of search terms in practical usage were taken as core terms to be manually categorized into 100 predefined categories. In addition, the machine efforts are to make it is possible to automatically categorize each new important search term appropriate subject categories, and cluster similar search terms for each corresponding category.

With the above approach, a prototype system we are constructing is to allow user using voice input to select a subject domain and speak out keywords for search. Based on the auto-constructed thesaurus, the system can recommend related terms for refined search if necessary. The preliminary results obtained so far have showed its possibility that a thesaurus suited for interactive voice-based web search can be automatically constructed and updated with the change of search term logs. A more effective and robust voice-based search engine could be developed.

2. OVERALL DESIGN

As mentioned before, the main focus of the ongoing research is to make a live thesaurus can be automatically constructed from search term logs of real-world search engines. On the basis of the live thesaurus, the research will further attempt to build a front-end search engine with capability of automatic term suggestion and voice input. Such a search engine is designed to be properly integrated with existing search engines as back-end engines in our
research. As depicted in Fig. 1 for each voice or typed query term, the developing front-end engine will suggest related terms through the access of the live thesaurus and retrieve web results via meta search. On the other hand, for each unknown but highly occurred typed query term, it will categorize appropriate subject categories and extract similar search terms. All of users’ query inputs and visiting web pages through the queries and suggested relevant terms will be recorded in the search term log, and update the corresponding information of the thesaurus. The thesaurus turns out will record up-to-date information including search term frequency, and even the frequency distribution of each subject category. The following introduction will describe more about the construction of live thesaurus.

3. ANALYSIS OF SEARCH TERM LOG

As shown in Fig. 1, our research consists of three major steps of analysis. The first-step analysis involves analyzing characteristics of search terms and extracting core terms. During the process of analyzing search terms, there is a valuable finding that some "core" search terms exist without the effects of time. Many searches would be affected by ephemeral trends in querying, such as searches related to a new movie being released or some specific occasions appearing during the year. To find out the effects of time locality, search terms from D-1998 were matched and filtered using the G-1999 log to have a comparison of coverage. It shows that 14,721 (near 77%) of D-1998’s top 20,000 search terms still exist in G-1999’s two weeks randomly-selected log, which indicates many core or important information needs are not much affected by time and worthy of further study. It is also noticed that only 8 popular terms in D-1998’s top 1,000 disappear from those of G-1999’s. These terms are taken as core terms in our research.

The first-step analysis also involves structuring subject taxonomies for building a classification scheme, and categorizing high-frequency search terms into certain subject categories in terms of users' possible intentions. To build such a proper scheme for categorizing search terms, we used a bottom-up methodology rather than top-down design commonly found in Internet search services like Yahoo or library communities like Dewey Decimal Classification. After quick reviewing the 14,721 core terms from D-1998’s log, we constructed a classification scheme mainly based on the analysis of human observations. Fourteen major categories with one hundred subcategories were developed. The major categories include Adult, Business, Chat, Computer, Education, Entertainment, Game, Health, Humanities, Media, Science, Shopping, Society, and Travel etc. Each category consists of several subcategories as well. The categorization of the 14,721 core terms was done manually by our team of five Library & Information Science students and a reference librarian for three months. Based on the careful judgement of users' possible intentions, i.e. for what purpose would be the corresponding search results used, each term was analyzed and classified into an appropriate subject category according to its major
intention. Since it is very likely that a search term can have multiple intentions from different users, besides one major category, each term was also assigned with a secondary category if necessary for cross-reference.

The second-step analysis involves instant categorization of each new search term. For efficiency and usability, the above manual efforts on core term categorization could be done only once. With such bases, a search term either from the non-core term set or from newly collected (logged) set after a period of time; it will be automatically categorized into appropriate subject categories by machine. As will be described in the Section 3, the categorization process has to be combined with the search process of the search engine. Whenever all search terms in the log have been assigned appropriate categories, the distribution of each subject category and the total frequency of its composed search terms can be calculated and monitored concurrently with the change of search term usage. Finally, the third-step analysis performs the extraction process of relevant terms for each subject category.

4. NEW TERM CATEGORIZATION

To make the categorization process incremental and adaptive, we need a mechanism to automatically categorize each new given term. First the above manually categorized core terms were taken as the seed vocabulary set $V$, and were used as the basis for such an automatic process. Let $C$ be a category set that can cover all requested topics by Web users, our research problem of new term categorization can be defined as: given a new search term $t$ and the seed vocabulary set $V$ with corresponding categories for each composed core term, the goal is to determine the most proper category or categories in $C$ for $t$ responding to users’ information needs; for example, to categorize a new search term "Microsoft" as an instance of Company Subcategory included in the major Computer Category.

A basic idea of our approach is to take term $t$'s co-occurrence search terms, the search terms that appear together with term $t$ in some Web documents, as $t$'s feature set. The approach is much similar to classifying a document based on the composed search terms within a conventional document classification process. Usually the first step of classifying a document is to determine how the terms in the vocabulary set contribute to each candidate category, which can be performed by a training corpus. Then the confidence that the given document belongs to each category is determined by all terms appearing in that document.

To continue the discussion about new term categorization problem, suppose there is a Web document collection $D$ in hand, and the categorization process developed is to rank all candidate categories in $C$ to find out the most probable category or categories that $w$ belongs to $W_w$. The proposed method of estimating the rank for candidate category $c$ is defined as follows:

$$ R_c = \sum_{w \in W_w} |D_w \cap D_t| $$

where $t$ is the term to be categorized, $W_w$ is the set of $t$'s co-occurrence terms in $V$ with category $c$, $w$ belongs to $W_t$, $D_w$ and $D_t$ are the sets of documents in $D$ which contain term $t$ and $w$ respectively, and $|D_w \cap D_t|$ is the number of documents containing both $t$ and $w$.

In order to obtain the necessary parameters, the above ranking process needs to be combined with the search process of search engine. Basically, each search term to be categorized is taken as a query for the search engine in retrieving the documents containing the search term. All co-occurred search terms would then be extracted from the documents as the feature set for the search term. The new term categorization could be performed.

A preliminary experiment has been done for testing the performance of the above method. The experimental data are the top 20,000 distinct query terms from D-1998's log, which cover about 81% query requests and mostly belong to core terms. We randomly select 1,000 terms from the 20,000 categorized term set as the testing set, and the rest are treated as the vocabulary $V$. The selected terms were then been sent as queries to a well-known search engine to collect the required Web documents. The returned documents are assumed highly correlated to $t$ and the co-occurrence term set $T$ can be therefore extracted accordingly. The obtained correct rates with the proposed method is shown in Table 1, where top $n$ means the highly ranked $n$ candidate categories contain the correct category. The experiment showed promising results that various new terms given by users could be categorized automatically.

<table>
<thead>
<tr>
<th>Correct Rate of Categorization</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Top1</td>
<td>48.30%</td>
</tr>
<tr>
<td>Top2</td>
<td>63.50%</td>
</tr>
<tr>
<td>Top3</td>
<td>68.90%</td>
</tr>
<tr>
<td>Top4</td>
<td>74.50%</td>
</tr>
<tr>
<td>Top5</td>
<td>77.10%</td>
</tr>
</tbody>
</table>

Table 1. The obtained correct rates with the proposed method compared with manual results for new term classification experiment.

With the above automatic mechanism, all of the search terms in the D-1998 log were then categorized, i.e. 20,000 terms were categorized by human and the other near 210,000 by machine. The machine categorization process will assign a search term at most five categories depending on the confidence in process.

5. RELEVANT TERM EXTRACTION
Users' queries for Web search are usually short. To deal with the short query problem, term suggestion techniques that attempt to identify users' intentions behind the short queries and recommend more precise search terms for users are, therefore, developing in many Web search engines. Towards the development of more effective term suggestion techniques, auto-extraction of relevant terms for huge amount of Web search terms is a crucial and challenging issue.

To determine relevant terms for each given search term, a conventional term suggestion process often relies on the information obtained from co-occurrence analysis of key terms in the indexed documents [4,6]. Applying such kinds of approaches into Web search, it is hard to determine searchable key term set. Extraction of relevant terms can be performed not limited in document space. The key term set could come from search engine's logs.

For each new term categorized with the above method, the purpose of relevant term extraction is to extract similar terms in the term sets with the same categories. It also assumes that the search terms, which could retrieve more shared search results, have a stronger relevance. According to this assumption, the relevance degree between different search terms can be calculated. The similarity estimation between two terms is similar to mutual-information-like metrics. A pair of terms with high co-occurrences in some Web pages will obtain higher similarity value. The selection of key term set could be easier and the suggested relevant terms better then document-based approaches.

6. VOICE-BASED INTERACTIVE SEARCH

Relevance feedback techniques have been widely used in IR for several years. The idea behind relevance feedback is to add the important terms in the retrieved relevant documents to the original query to an attempt to expand the subjects requested in the query in order to retrieve more relevant documents from the database. In order to find really important topic terms and to filter out non-related terms, automatic term suggestion techniques are sometimes needed. In the research, both relevance feedback and term suggestion techniques are implemented based on the above live thesaurus construction.

A prototype system we are constructing is to allow user using voice input to select a subject domain and speak out key terms for search. Based on the auto-constructed thesaurus, the system can recommend related terms for refined search if necessary. Many topic terms which are conceptually relevant but do not appear in the query might be recommended. The extracted topic terms can help the user formulate more precise queries and perform more searches. The user can choose to reformulate queries via voice after looking up the extracted topic terms. Using these methods, most of the desired documents having concepts relevant to the query can be finally retrieved.

The preliminary results obtained so far have showed its possibility that a thesaurus suited for interactive voice-based web search can be automatically constructed and updated with the change of search term logs. Besides, based on the live thesaurus, certain kinds of users' information behaviors could be characterized and a more effective voice-based search engine could be developed.

5. REFERENCE