A Specialized Search Engine for City Classification Information Retrieval

Zhijiang Wang **Jiangning Wu**

Institute of Systems Engineering, Dalian University of Technology, Dalian, 116024, China
zjwang@student.dlut.edu.cn, jnwu@dlut.edu.cn

Abstract

This paper describes the process of constructing a specialized search engine for city classification information retrieval. Compared with the general search engine, it can provide more precise information and domain-specific knowledge for users by means of some effective approaches. The forward maximum matching algorithm is used for Chinese segmentation. The query expansion method is conducted to the given queries based up on the synonymy lexicon. And the query recommendation approach is proposed in terms of user logs. The experimental results show where the proposed approaches are tested on corpora of significant scale, showing clear improvements with respect to conventional keyword-based search.

Keywords: Search Engine, Information Retrieval, Query Expansion, Query Recommendation

1 Introduction

A search engine is a system that collects and organizes web documents, and presents a way to select documents based on certain words, phrases or patterns within documents. In the last couple of years, search engine technology had to scale up dramatically in order to keep up with the growing amount of information available on the web. Current large-scale search engines such as Google have the ability to handle billions of pages. But the recall of even the largest commercial search engines is rather low, covering 50-70% of the web today [1]. As the amount of web sites is growing rapidly, the number and size of stored documents is growing even faster and site contents are getting updated more and more often. Centralized systems cannot grow that fast and therefore they cover an ever-decreasing segment of the web. The large size and general focus of their indices entail a rather low precision. Naturally, specialized search engines and domain-specific web portals have seen an increasing popularity in recent years. They are called vertical search engines and vertical web portals, respectively [2]. Compared with general search engines, obviously, vertical search's processing scope is greatly narrowed down, and with a specialized index it has consequently a more structured content and offers higher precision than a generalized search engine, as it has been intelligently extracted from the web[3].They are more appropriate to classified information retrieval for specialized markets and target groups.

In this paper, we construct a vertical-like search engine based on the city catering service information. Compared with the general search engine, it can provide more precise information and domain-specific knowledge for users in which the forward maximum matching algorithm is used for Chinese segmentation, the user queries are expanded based on the synonymy lexicon and the query recommendation is gotten based on user logs. An experiment is carried out to vertify the effectiveness of the proposed approaches and satisfying results come back from the developed searching system.

The rest of this paper is organized as follows. In Section 2, an architecture of Web search engine is proposed and three main process modules, including Indexer, Chinese segmentation and query expansion are addressed. In Section 3, query recommendation based on query logs is introduced. An experiment is conducted to test the performance of the developed search engine and the experimental results are discussed in Section 4. Finally, conclusions and future works are summarized.

2 Architecture and Algorithms for the Search Engine

2.1 Architecture of the search engine

Most practical and commercially operated Internet search engines are based on a centralized architecture that relies on a set of key components: crawler, indexer, page and index storage and searcher as shown in Figure 1.

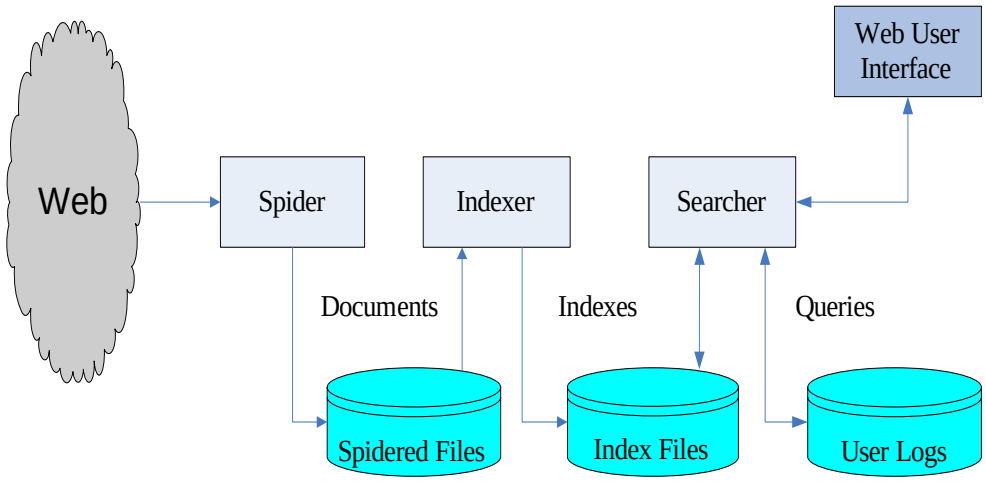

Figure 1. The architecture of search engine

The crawler is a module aggregating documents from World Wide Web in order to make them searchable and passing retrieved documents into page storage. Several heuristics and algorithms exist for crawling, and most of them are based upon following links. The indexer is a module which is designed to handle documents and build a searchable index from them. Index is a data structure that allows fast random access to words stored inside it. Terms are extracted from the documents and the index is created and saved as files. Common practices are inverted files, vector spaces, suffix structures and hybrids of these. The core of all today's web search engines is inverted indexes. This data structure makes efficient use of disk space while allowing quick keyword lookups. What makes this structure inverted is that it uses tokens extracted from input documents as lookup keys instead of treating documents as the central entities. In other words, instead of trying to answer the question "What words are contained in this document?" this structure is optimized for providing quick answers to "Which documents contain word X?" The searcher is working on the output files from the indexer. The searcher module therefore is responsible for receiving search requests from users and analyzing user queries, including Chinese segmentation, query expansion and query recommendation. This module relies heavily on the index, and always includes the ranking module which has the task of sorting the results according to their relevance to the given query so that results near the top are the ones most likely to be what the user is looking for. Systems usually run the crawler, indexer, and searcher sequentially in cycles. First the crawler retrieves the documents, then the indexer generates the searchable index, and finally, the searcher provides functionality for searching the indexed data [4].

2.2 Algorithms of the search engine

2.2.1 Forward maximum matching algorithm for Chinese segmentation

Segmentation is applied for both documents and user queries. Documents are first segmented to yield the index terms that are stored for the subsequent query matching process. User queries are also segmented prior to matching. And the results of segmentation will influence the precision of search engine.

The Chinese word segmentation is a challenge

due to the fact that Chinese text has no delimiters to mark word boundaries other than the punctuations, while English text uses space as word delimiter. Many methods for Chinese word segmentation have been proposed [5]. These methods can be roughly classified into two groups, namely, character-based approaches and word-based approaches. The former ones can be defined as purely mechanical processes that extract certain number of characters from texts. According to the number of characters extracted, character-based approaches can be further divided into single character-based approach and multi-character-based approaches. Single character-based approach divides Chinese texts into single characters and is the simplest method to segment Chinese text. Multi-character-based (or N-gram) approaches segment texts into strings containing two (bi-gram), three or more characters. The latter ones, as the name implies, attempt to extract complete words from sentences. They can be further categorized as statistics-based, dictionary-based and hybrid approaches. Dictionary-based approach is commonly used in most current systems utilizing the word-based approach for text segmentation [6].

Forward maximum matching (FMM) algorithm is one of the most popular segmentation methods[7][8]. We take the dictionary-based forward maximum matching algorithm to segment user queries. The dictionary contains all possible words and phrases used in Chinese texts. Before segmentation, some stop words will be filtered from the user queries based on the stop-word list. Some Chinese words such as 的, 地, 得 without practical meanings are defined as stop words and stored in stop-word list. Then the segmentation starts at the beginning of the query and proceeds left-to-right finding one segment boundary at a time. At each segment boundary, it searches for the longest word in a lexicon that matches the next few characters from the segment boundary. If no word is found to match the next few characters, the next single character is considered as a word and the word boundary moves forward by one character. Otherwise, the word boundary moves forward by the number of characters in the matched word. Take the query in Chinese 搜索引擎的新体验 as an example where the lexicon contains the words 搜索, 引擎, 搜索引擎 and 体验. The stop word 的 will be filtered before segmentation. Then starting at the first segment, FMM finds two matched words 搜索 and 搜索引擎 from the lexicon. It then chooses the longest word 搜索引擎 and moves on to segment the remaining character sequence 新体验. As there is no word in lexicon that matches the character 新, this character is treated as a single-character word. The procedure continues until the end of the sentence, resulting in the segmented sentence 搜索引擎 新 体验. FMM algorithm is an efficient algorithm and can be implemented easily.

2.2.2 Query expansion based on synonymy lexicon

Query expansion is processed by searcher module, which is a technique, widely used for obtaining additional terms relevant to a given query (search keywords). Query expansion is needed due to the ambiguity of natural language and also the difficulty in using a single term to represent an information concept. It is usually used to help searchers express their intentions more accurately and increase the precision of search results [9]. Krovetz and Croft (1992) observed that the most benefit is achieved with high-recall searches that depend on matches of single concepts. With query expansion, the user is guided to formulate queries which enable useful results to be obtained [10].

This paper realizes query expansion according to semantic similarity of Chinese words. But quantifying semantic similarity of Chinese words is deemed to be difficult and more difficult task than that of English words because of the nature of Chinese language and lacking large-scale hierarchy organized language resources like Word Net. As a matter of fact, efforts in building Chinese semantic lexicon started in 1980s, some achievements on semantic classification system of Chinese words have been made, such as《同义词词林》(An electronic dictionary of synonymy), etc. Our synonymy lexicon is obtained from such dictionary which consists of 53,859 Chinese words. Some words not used very common are deleted from the synonymy lexicon and some common words are added to the lexicon manually resulting in a lexicon of 75369 Chinese words. The lexicon is used to expand query automatically. Every line in the lexicon gives a listing of words with similar meanings

and these words use space as word delimiter. In the searcher module, while the searcher segments the user query, every segment of the user query will be checked whether it exists in the synonymy lexicon or not. If it can be found in a certain line in the lexicon, all the words in this line are viewed as expanded queries. Otherwise, the word will not be expanded. The procedure continues until all the segments are expanded. Let's use the query in Chinese 计算机软件 as an example where the synonymy lexicon contains the word 计算机 电脑 微处理机 微机 微型机 微处理器 处理器 in a certain line. After segmenting, the query is divided into two words including 计算机 and 软件. The word 计算机 is contained in the lexicon, so it is expanded to 计算机 电脑 微处理机 微机 微型机 微处理器 处理器. As there are no words in the synonymy lexicon that match the word 软件 , this word is not expanded at all. As a result 计算机 电脑 微处理机 微机 微型机 微处理器 处理器 软件 are viewed as the final expanded query. Finally, the searcher module will search relevant documents from index files according to the expanded query and the document containing one of the segments in the expanded query will be hit. The returned documents will be sorted according to the relevance with the expanded query. The results show that query expansion using synonymy lexicon has better performance in recall. It can also be found that the correctness of word segmentation has a great impact on the quality of query expansion.

3 Query Recommendation Based on Query Logs

Providing related queries for search engine users can help them quickly find the desired content. Recently, some search engines start showing related search keywords in the bottom of the result page. Their main purpose is to give search engine users a comprehensive recommendation when they search using a specific query. Recommending the most relevant search keyword set to users not only enhances the search engine's hit rate, but also helps users to find the desired information more quickly. Also, for some users who are not very familiar with a certain domain, we can use the queries that are used by previous similar searchers who may have gradually refined their query, hence turning into expert searchers, to help guide these novices in their search [11].

There are many methods to get query recommendation, such as the method based on a query clustering process in which groups of semantically similar queries are identified, the method utilizing the click-through data, the method mining the search engine logs which contain abundant information on past queries, etc. This paper takes an easy and efficient method by searching user query logs to get query recommendation. The algorithm considers only queries that appear in the user logs. All the user queries are stored as structured data in the database. The same query will be stored only once, but the times of the query will be added one each time. And the returned query recommendations will be sorted according to the times of related queries. The query data in the database will be indexed by indexer module at interval time and stored as query index files. The indexing process includes filtering stop-word, Chinese segmentation, etc. When a user inputs query, the searcher module will search relevant results, meanwhile it will search related queries from query index files and return related keywords as query recommendation. Let's take the query in Chinese 大连酒店 as an example. The query recommendation can be 大连富丽华酒店, 大连海景酒店, 大连瑞士酒店 , etc. The user can click these related keywords to search the desired information.

4 Experimental Results and Discussions

The three measures corresponding to precision, recall and speed are widely used in search engine performance evaluation. Speed is the response time of the system for the user's query. Precision and recall respectively are defined as

$$\text{Precision} = \frac{r}{r+l} \times 100\% \qquad (1)$$

and

$$\text{Recall} = \frac{r}{r+b} \times 100\% \qquad (2)$$

where r is the number of relative documents in the result list, and l is the number of irrelative documents in the result list, and b is the number of relative documents but not in the result list.
These metrics interplay each other. For example, in a specific system, both precision and speed drop off with the increasing of recall [12].

In order to compare the performance between keyword-based and query-expanded methods, a

case is illustrated. This case uses ten queries to measure the recall and precision. It yields different results between the two methods. Figure 2 shows the return results from both keyword-based and query-expanded searching methods using 中山区饭店 as the query. The precision and recall of the searching system are shown in Figure 3.

Figure 2. The results of the keyword-based vs. the expanded query searching method

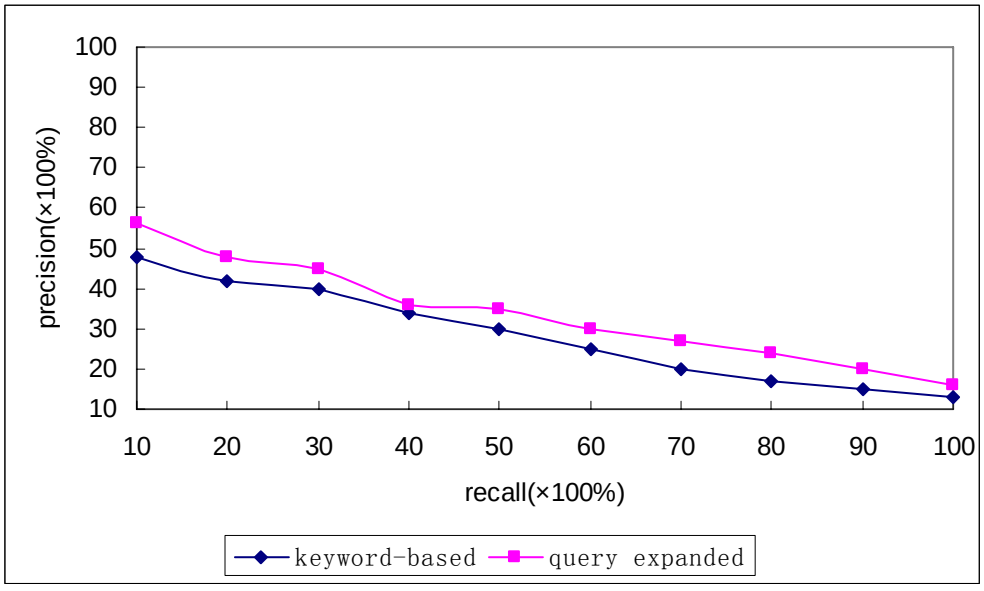

Figure 3. Comparison of search performance of two methods

Figure 3 shows that the accuracy increases with the recall decreasing. It also shows that the recall of the query-expanded method is higher than that of the keyword-based one.

Another important aspect is the response time. Although the expanded query method will spend some extra time on expanding users' queries, current response time of the developed search engine is less than 1 second, which is short enough for practical use.

5 Conclusions and Future Work

This paper presents the architecture of a specialized search engine and introduces the main modules. Query expansion and query recommendation are two features of the developed search engine. The experimental results show that the query expansion technique can yield performance that matches that of conventional query techniques. It can also be found that query recommendation brings good experiences to search engine users.

One direction for future research is the clustering of search results. Clustering can be defined as a process of organizing pieces of textual information into groups whose members are similar in some way, and groups as a whole are dissimilar with each other. We will try to devise a method to yield comprehensible and grammatically descriptions of clusters, and the method should assure a clear, transparent relationship between the cluster label and its contents. Finally, we plan to carry out several more experiments to test the performance of the method and how the method affects the quality of document assignment.

Acknowledgments

The work reported in this paper is sponsored by Dalian Shanyou Technology Company and National Natural Science Foundation of China (NSFC) under grant Nos. 70431001 and 70620140115.

References

[1] Google Search Technology Online at <http://www.google.com/technology/index.-html>.

[2] G.Almpanidis, C.Kotropoulos, I.Pitas, Combining text and link analysis for focused crawling-An application for vertical search engines. Information System, 32:886-908, 2007..

[3] R.Steele, Techniques for specialized search engines. Proceedings of the Internet Computing, 01:25-28, Las Vegas, 2001.

[4] Knut Magne Risvik, Rolf Michelsen, Search engines and Web dynamics. Computer Networks, 39:289-302, 2002.

[5] Jianfeng Gao, Mu Li and Chang-Ning Huang, Improved Source-Channel Models for Chinese Word Segmentation. ACL, 2003.

[6] Schubert Foo, Hui Li, Chinese word segmentation and its effect on information retrieval. Information Processing and Management, 40:161-190, 2004.

[7] Pak-Kwong Wong and Chorkin Chan, Chinese Word Segmentation based on Maximumu Matching and Word Binding Force. COLING,96:200-203, 1996.

[8] David D. Palmer, A Trainable Rule-based Algorithm for Word Segmentation. ACL, 1997.

[9] Atsushi Sugiura, Oren Etzioni, Query routing for Web search engines: architecture and experiments. Computer Networks, 33:417-429,2000.

[10] J.Bhogal, A.Macfarlane, P.Smith, A review of ontology based query expansion. Information Processing and Management, 43:866-886,2007.

[11] Zhiyong Zhang, Olfa Nasraoui, Mining Search Engine Query Logs for Query Recommendation.Proceeding of the 10th international conference on World Wide Web, Edinburgh, Scotland, 2006.

[12] Wen Yue, Zhiping Chen, Xinguo Lu, Feng Lin, Juan Liu, Using Query Expansion and Classification for Information Retrieval. ACM, Edinburgh, Scotland, 2006.