A Real-time Integration of Concept-based Search and Summarization on Chinese Websites

Joe F Zhou and Weiquan Liu
Intel China Research Center
100020 Beijing, China
{joe.f.zhou; louis.liu}@intel.com

Abstract

This paper introduces an intuitive search environment for casual and novice Chinese users over Internet. The system consists of four components, a concept network, a query reformulation model, a standard search engine, and an automatic summarizer. When the user enters one or more fairly general and vague terms, the search engine returns an initial answer set, and at the same time pipes the query to the concept network that connects thousands of conceptual nodes, each referring to a specific concept for a given domain and pointing to a number of associated conceptual terms. If the concept is located in the network, the related conceptual terms are displayed. The user has the option of using one or more of these specific terms to reformulate the next round of searches. Such search iterations continue until the user's ultimate information seeking goal is reached. For each search iteration, auto summarizer presents the main theme of the document retrieved and an optional text-to-speech engine can read out the output summary if the user prefers.

1. Introduction

Internet is changing the world, and at the same time changing people's information seeking behaviour. Traditionally, information searchers are trained professionals working in libraries or other special technical or scientific fields. They have developed a variety of techniques and heuristics for addressing information seeking difficulties in the environment typically dominated by Boolean query formula. These Boolean information retrieval systems are normally commercial and non-interactive systems and the searches conducted in such settings are exact-match and set-based retrieval from databases of indexed citations and abstracts of documents (Koenemann and Belkin, 1996). With the dramatic explosion of information sources over Internet, current user population is no longer restricted to professional searchers. Practically it includes everyone in life. The majority of these users, however, are either casual or novice or both. Casual users, such as browsers of news stories, look for interesting information rather than information relevant to a specific need (Stadnyk and Kass, 1992). Novice users may have a specific information topic, but due to little or no training in search and retrieval, they don't know how to make best use of the available operators and tools. On the whole, query formulation is one problem facing both types of information users (Turtle, 1994). Internet users all have difficulty in mapping their intent to any logical query structure. They prefer limiting their searches to one or a simple list of terms, while seeking help from the system to guide them to achieve their ultimate information goal.

In this paper we introduce an integrated system that combines a standard search system with a query reformulation model, a pre-constructed concept network, an automatic document summarizer, and an optional text-to-speech (TTS) engine. Together, these intelligent components provide intuitive human intelligence to the Chinese users over Internet. The integrated system not only guides and navigates the user to perform searches in a humanly conversational way, but also makes and delivers the retrieved information back to the user in a simple, succinct, and easily comprehensible manner.

2. System Configuration

Figure 1 and Figure 2 present the overall system configuration and data flow of the integrated system. The system consists of four main components: a concept net, a query

reformulation model, a standard search engine, and a summarizer. There is also an optional component, i.e., if the user chooses, she can launch a text-to-speech (TTS) engine to read out the automatically generated summary.

The concept net is a network of conceptual terms. It is normally constructed for a specific domain with certain amount of human intervention. A link connects each pair of related concepts in the network, specifying the semantic relationship between them. For the current economic news domain, the main types of relationships include, but not limited to, canonical form of, synonym of, hyponym of, hypernym of, part of, product of, member of, etc.

The system accepts users' queries expressed in Chinese natural language. The query may contain the terms that are already stored in the concept network. Unlike professional searchers, Most of Chinese web users have little training in information retrieval or have no prior knowledge or experience in it. They may not even know where to search and how to search. To perform an initial search, they tend to use one or more very general or vague terms. Under such circumstances the system guidance and navigation are extremely important. One unique functionality of this integrated system is to intuitively lead a casual or novice user from a more general search to a more specific search until the user becomes satisfied with the returned information.

For each search conducted, the query reformulation model looks up the concept network for more specific terms that are relevant to the more general terms in the earlier query. For example, if a general term is a company's name, then its subsidiaries, its products, its stock symbol, its industrial code, etc. are considered to be specific information about the company. The query reformulation model either replaces or expands the original query with these related terms and formulates them into a standardized format. Search operators, such as AND, OR, NOT, NEAR, etc. are used to connect the terms selected. Assigned to each individual term is a different weight so as to reshape a new search emphasis.

The standard search engine performs the search against the targeted database using the reformulated query with N relevant documents returned in an order of the relevance to the query (N is a number defined by the user and it is 10 by default). At the same time, the concept terms in the original query and the corresponding specific terms extracted from the concept network are also displayed. The web interface is designed in a way that makes each of the specific terms searchable. The user has the choice to select any of these specific terms to form a new query. The search engine takes the new query to perform the next round of search, actually a more specific search based on the user's intention. The iteration continues - the more specific search while using more specific terms, the closer the user will be to his desired information - though he can stop anytime after each search iteration to examine the retrieved documents.

A text summarizer automatically generates the summaries for the documents each search iteration returns (Liu and Zhou, 2000). Together with the output summary, a selection panel that includes the key-word list (average 3 to 6 words), the headline, and the leading text (usually the first 100 characters) of the document is displayed. The selection panel provides the user with an ability of examining the retrieved information more efficiently. By glancing over the key words, the user should be able to grasp the main idea of the document. He can make a decision whether to skip the document or continue to learn more about it. If his choice is the latter, he can move up to the headline or the leading text that offer more about the document content. He can, of course, move further to look at the summary that is supposedly a mini-document of the original. If the user decides to move further to read the entire document, then he will click on the hyperlinked title or headline of the document. The integrated system will go directly online, usually a specific website, to grab the document for the user.

Associated with the summarizer is an optional text-to-speech (TTS) engine. The user can choose either to read the document himself or get relaxed by simply listening to the system's voice output.

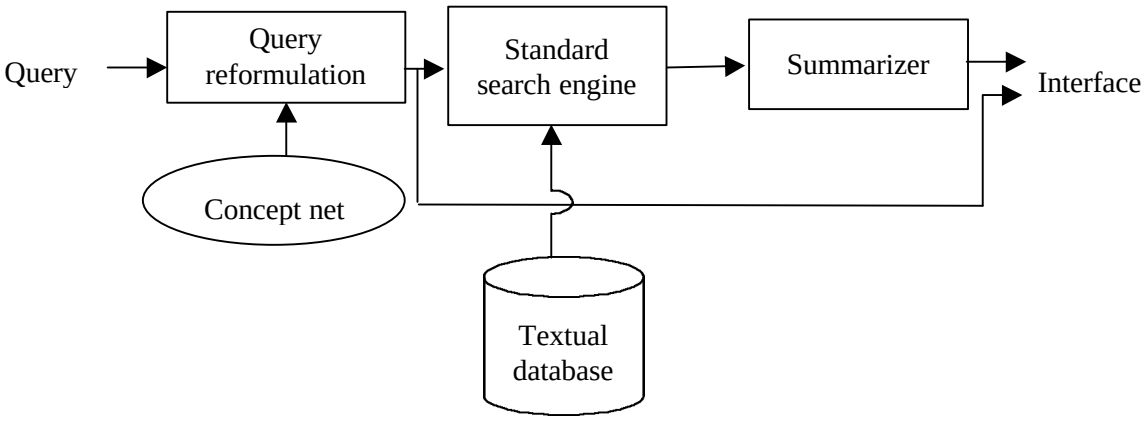


Fig. 1 System configuration

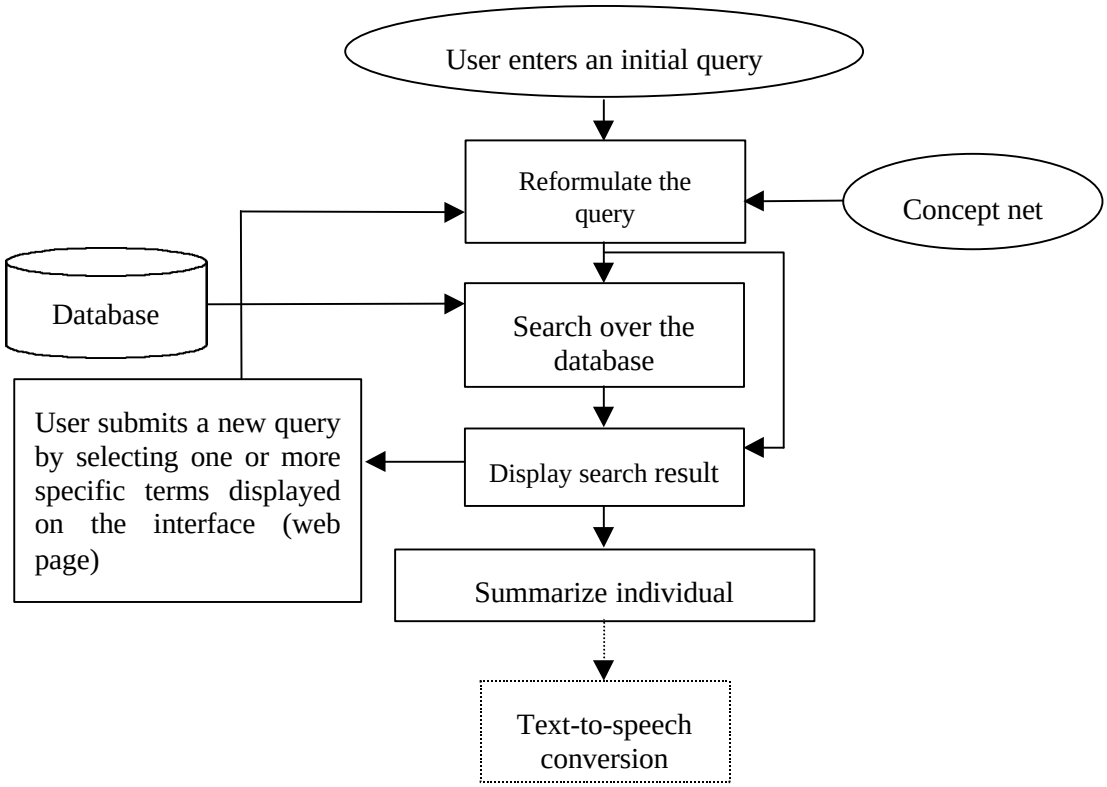


Fig. 2 System Data Flow

3. An Example Search and Summarization

In this section we use an example to illustrate how the integrated system works.

Suppose that there is a Chinese user who has no experience in search over Internet, neither has any idea regarding where or how to do it. He just heard some news about 联想 (*Legend*, one of the biggest computer manufacturing companies in China) and wants to find it out from Internet. When we make our integrated search and summarization system available to him, the only word in his mind is the name of the company 联想 (*Legend*). So, he enters this fairly general term as his first search query and presses the *Go* button (see Figure 3 (a)).

Figure 3 (b) shows what the integrated system returns to the user in response to his initial

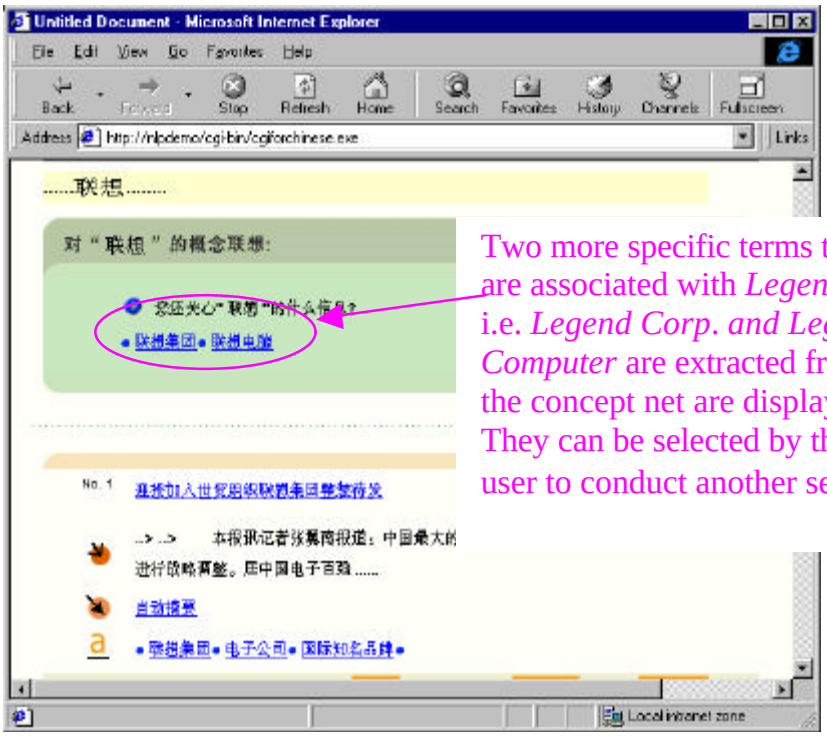
search. In addition to the top 10 most relevant documents, two more specific terms are extracted from the pre-constructed concept network, i.e., 联想集团 (*Legend Corporation*) and 联想电脑 (*Legend computer*). In the concept network these two terms are found to be closely associated with the initial search term联想 (*Legend*). When the user examines the top relevant document returned, he finds that a selection panel is displayed on the interface that is associated with this (or each) returned document. If looking up from the last row, he notices three key words that are extracted from the document, i.e., 联想集团 (*Legend Corporation*), 电子公司 (*electronic company*), 国际知名品牌 (*internationally well-known brand*). By putting these three key words together the user should be able to grasp the main idea expressed in the document (something like "Legend Corporation is an electronic company who has some internationally well-known brand names"). If the user wants to know more about the document, he can move up to click the button 自动文摘 (automatic summarization) or read the leading text (in this case the first 50 characters from the document are extracted). If the user wants to refer to the entire document, he can do it by clicking on the headline. The integrated system will go live to Internet to retrieve the entire original text.

But, what if the user is not satisfied with the current answer set? Then, he has an option of kicking off another round of search. To do that he may want to narrow down his search by choosing between 联想集团 (*Legend Corporation*) and 联想电脑 (*Legend computer*) - the two relatively speaking, more specific terms associated with his initial query 联想 (*Legend*). Let us suppose that the user decides to select 联想集团 (*Legend Corporation*) to conduct a new but more specific search. The search can be activated by simply clicking on this selected term. Figure 3 (c) shows the results of this search iteration.

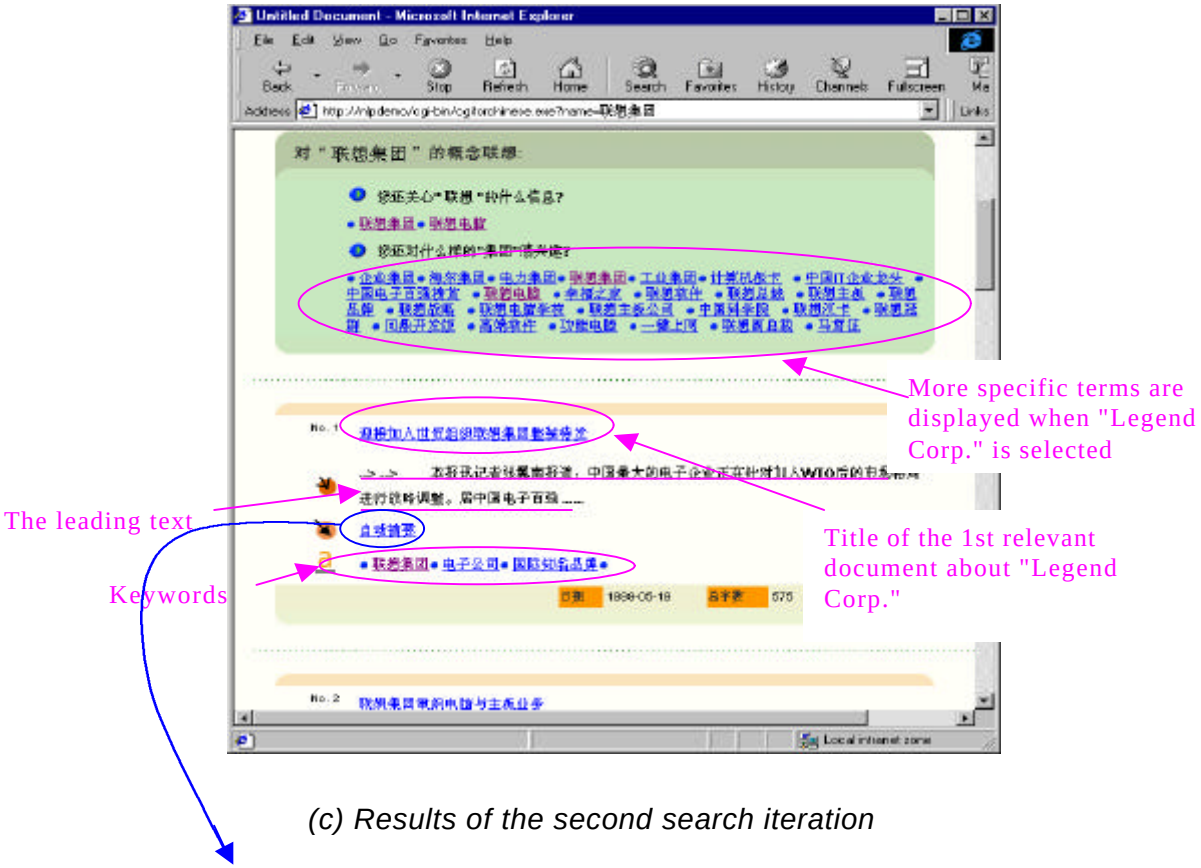
Again, in addition to the top 10 most relevant documents returned, more than 20 related terms are extracted in this round from the concept network that are considered to be conceptually relevant to the search term联想集团 (*Legend Corporation*). These terms represent even more specific concepts comparing to the two specific terms returned in the first round. A detailed examination of these terms reveals that they represent the following conceptual categories: first, more specific news or information about the company, such as 中国IT龙头 (*the head of Chinese IT*), 中国电子百强榜首 (*No. 1 of the first 100 most powerful electronic companies in China*), 联想冠群 (*Joint venture of Legend and CA*) and 幸福之家 (Happy Family, a software kit); the products of Legend Corporation, such as 联想主板 (*Legend motherboard*), 联想软件 (*Legend software*), 联想汉卡 (*Legend Chinese card*); the management of Legend company, such as 联想总裁 (*president of Legend*), 联想副总裁 (*vice-president of Legend*); and other related corporations, such as 企业集团 (*business corporation*), 电力集团 (*electronic corporation*), 工业集团 (*industrial corporation*), and 海尔集团 (*Haier Corporation*). At this point, the user has another option. He can go through the selection panel associated with each retrieved document to see which document contains the information he wants. If so, his search iteration will terminate. If not, he can reformulate his next search by selecting any one or more terms from the specific term list. In this particular case, the user may go for the first choice since the top document retrieved turns out to be identical for these two search iterations conducted so far. The document entitled 迎接加入世贸组织联想集团整装待发 (*Reorganization of Legend Corporation to welcome the joining of WTO*) receives the relevance rate of 100% in both searches. This is probably the news about Legend Corporation the user is looking for.

As mentioned above, even if the user has obtained the information he desires, he still can reformulate or expand his next search for more specific news about 联想 (*Legend Corporation*). Also, the user can launch a completely new search for another company. Remember 海尔集团 (*Haier Corporation*), another most powerful company in China, appears in the specific term list. By clicking on this term, the user will be able to locate the latest news or other information about a new company.

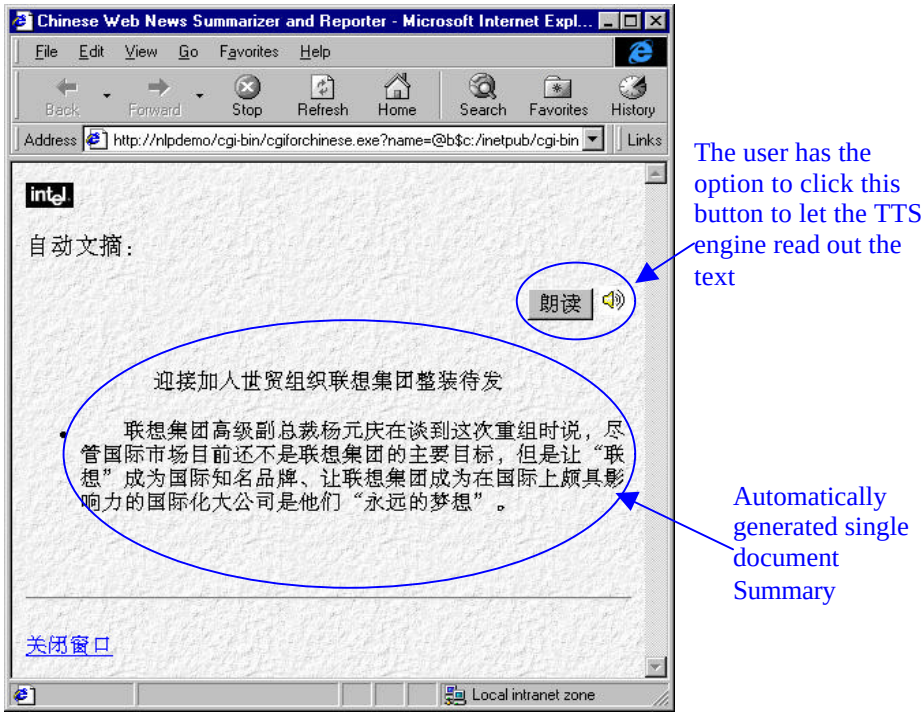
(a) User enters initial query



(b) Results of the first search iteration



(c) Results of the second search iteration



(d) summary for a single document

Fig. 3 Snapshots of the interactive interface

4. Conclusion

This paper presents an intuitive search system for casual or novice Chinese users on specific websites. Instead of describing the algorithmic portions of each component, we focus on illustrating the overall system design and how the integrated system works and serves the end-users. The presented system provides human intelligence to the user who desires to seek relevant information over Internet. It does not assume that the user receives any training in search and retrieval or has any prior experience in using Internet. The user can start his search with a general and vague term or idea. The integrated system will guide the user from a general search to a series of more specific searches until he is fully satisfied with the information returned. Our survey states that the existing search systems, especially those in the Chinese market, are still dominated by key term search mechanism. This integrated system presents a new paradigm. It provides a concept-driven search environment that allows the user to manipulate the semantic relationships between the original query terms with its associated terms. We have so far built two applications based on two specific domains. The preliminary results have demonstrated that the integration of a concept network, a query reformulator, a standard search algorithm, an auto summarizer, and an optional TTS engine indeed suits the current information seeking behavior and make search activities in websites more intuitive, as well as productive.

References

Brandow *et al*, Brandow R. Mitze K. and Rau L F. *Automatic Condensation of Electronic Publication by Sentence Selection.* Information Processing & Management, 31(5): 675-68, 1995

Koenemann J. and Belkin N., *A Case For Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness.* In the proceedings of ACM/SIGIR, 1996.

Stadnyk I. and Kass R. *Modeling Users' Interests in Information Filters.* Communications of the ACM, 35(12):49-50, 1992.

Turtle, H., *Natural Language vs. Boolean Query Evaluation: A Comparison of Retrieval Performance.* In the proceedings of ACM/SIGIR, 1994.

Liu, W. and Zhou J. *Building a Chinese Text Summarizer with Phrasal Chunks and Domain Knowledge*. In the Proceedings of Rocling' 2000. Taipei, 2000.