Journal of Computational Information Systems 8: 20 (2012) 1–8
Available at http://www.Jofcis.com

Complex Query Recognition Based on Dynamic Learning Mechanism ⋆

Qingshan LI[∗], Yingcheng SUN, Baoye XUE

Software Engineering Institute, Xidian University, Xi'an 710071, China

Abstract

Complexity of queries is an important factor which affects retrieval quality of searching engines. If the complexity of search criteria can be analyzed in advance and the search criteria is pre-processed with corresponding mechanisms, the retrieval results will be greatly improved. This paper proposes an identification method of complex retrieval criteria based on the dynamic learning mechanism. Then the definition of complex retrieval criteria and the influential factors, i.e. the retrieval criteria complexity and complex semantic structure are given, and the search criteria is identified by using dynamic learning mechanism. The experiments show the necessity and effectiveness of the proposed retrieval criteria complexity and complex semantic structure, and thus of the complex retrieval criteria. The proposed classification algorithm and dynamic mechanism of learning are also proved effective by the experiments.

Keywords: Query Analysis; Complexity Recognition; Dynamic Learning

1 Main Text

With the quick spread of the Internet and rapid growth of network information resources, the main information retrieval tool search engine is getting more and more applications. However users do not always find exactly what they're looking for after they input a single query in search engines. The main reason is the retrieval condition form problem except for the factor that the needed information does not existed on the Internet or does not recorded by search engines even it exists.

If the users' search queries do not match the target information well, the returned results might deviate from the real needs, although search engines may have collected the needed information. And Chinese expressions vary greatly, which increase the difficulty for search engine to understand users' search intent. For example, if the search query is "stock now should buy or should sell", then the related information "it is not the right time to enter stock market now", "you don't have

⋆Project supported by the National Natural Science Foundation of China (61173026), the National High Technology Research and Development 863 Program of China (2012AA02A603), the Fundamental Research Funds for the Central Universities of China, and the Defense Pre-Research Project of the 'Twelfth Five-Year-Plan' of China (513***301).

[∗]Corresponding author.
Email address: qshli@mail.xidian.edu.com (Qingshan LI).

1553–9105 / Copyright © 2012 Binary Information Press
October 15, 2012

to rush to sell shares in his hand" and so on cannot be easily retrieved. That's because the most of traditional search engines are based on keyword matching, they don't have much capacity of understanding the semantics of the search criteria. If the input query keywords do not or seldom appear in the web pages, the web pages are likely to be considered as unrelated content and will not be returned to users. Actually this information may be closely related to the users' intention, thus it results in low search efficiency.

For the above-mentioned problems, this paper puts forward a dynamic analysis mechanism, which can be used to analyze the retrieval conditions. The search queries that may result in bad retrieval results are defined as complex queries and will be identified, while the others are defined as simple queries. By further processing of complex queries, such as helps to modifying the queries by using question-answering systems or providing friendly keyword guidance and so on, the search efficiency of complex queries can be greatly improved. The paper proposes that in information retrieval, complexity and sophisticated semantic structure of search conditions have significant impacts on the performance and complexity of query analysis.

The organization of this paper is as follows: Section 1 is a brief introduction of the problems studied in this paper. Section 2 introduces some related studies. Section 3 introduces the concepts and definitions of complex search conditions. Section 4 gives the complexity analysis strategy of search conditions based on dynamic learning mechanism. Section 5 gives experiment. Conclusion and future work followed in section 6.

2 Related Studies

Research issues related to the subject of this paper mainly focuses on the following three aspects. First is the research on Chinese sentence pattern recognition. Shi Xifan researched paratactic structure identification method of complex questions [1]. The second aspect is the recognition of phrases. Xi Jianqing etc [2] researched HMM-based Chinese prepositional phrases automatic identification method, and Xu Jian [3] discussed phrases recognition using the improved HMM model. The third aspect is research on query classification methods. Harabagiu S and others proposed that complex conditions often contained connection phrases, continuous list, or embedded condoms, which can be decomposed through grammar and semantic analysis [4]. G. Kumaran and J. Allan researched characteristics of long query and proposed selective interactive reduction and expansion (SIRE) technique to capitalize on the strengths of the interactive query reduction (IQR) and interactive query expansion (IQE) techniques [5]. G. Kumaran and Vitor R. Carvalho present a way to automatically reduce long queries to shorter, more effective ones, involving transforming the reduction problem into a problem of learning to rank all sub-sets of the original query (sub-queries) based on their predicted quality, and selecting the top sub-query [6].

This paper mainly focuses on the recognition of complex retrieval conditions, and the retrieval conditions are classified into simple search conditions and complex ones, and the recognition process is based on similarity matching algorithms and dynamic learning mechanism.

3 Definitions of Complex Query

According to the actual condition of received requests of search engines, queries can be classified into two types of forms: simple search condition and complex search queries. Simple queries cover a wide range of expression forms, including phrases, combination of phrases and simple sentences with clear and intuitive intentions. This kind of queries can result in comparatively good search effect after word segmentation. The other kind of queries-complex queries, mainly refer to the ones that are composed by sentences with complicated semantic structures or quite many adjectives. This kind of queries is too complicate for search engine to understand. Query Complexity and Complicate Semantic Structure are two key elements of the complex queries.

3.1 Query complexity

Definition 1 *Query Complexity (QC), refers to the number of words of a query after word segmentation and stop-word removal such as "的" (of), "了" (the) and other auxiliary words, as well as "哼" (hum), "啊" (ah) and some other interjection words.*

We obtained part of query logs from a widely-used Chinese search engine Sogou. By word segmentation, we removed part-of-speech tagging and stop-word, the QC values are counted and analyzed. The analysis finds that about 3% of user search queries have QC value higher than 6. 360 queries with QC value taken from 1 to 12 are selected, and each QC value corresponds to 30 queries. The 360 queries are searched in Baidu, Google, Sogou and Yahoo!, and the average similarities of queries with the same QC for every search engine are calculated and recorded. The similarity is calculated from the top 30 organic results of each search engine according to the conclusion that most users only browse the top 30 results from research [7]. Figure 1 gives the effect of QC on the precision of searching.

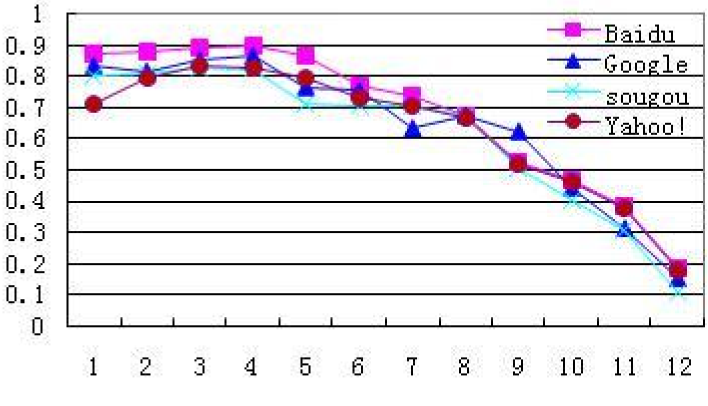

Fig. 1: Effect of QC to search results' relevance

From figure 1, it is showed that the queries which have QC within 3 and 5 can get the most correlated search results. The similarities of returned results from the four search engines are all higher than 0.7, among which the results from Baidu have the highest similarity of close to 0.9 with the queries. This means that 3 to 5 keywords are most likely to match the pages that users need, and with the best search effect. The figure also shows that when QC increases, the similarity is declined. And all the four search engines return less than 0.5 similarity correlated pages as the QC is higher than 10, which means the query search effect is obviously becoming poor. From all above, it is found that QC value is a significant factor that affects the searching results to the search engines.

3.2 Complicate semantic structure

Definition 2 *Complicate Semantic Structure (CSS), refers to the query structures which include one or more elements in structure set S = "V+Dno/Dnot/Dornot+V", "A+ Dno/Dnot/Dornot +A", "VV", "AA", in which V means verbs, A for adjectives, and Dno refers to "不", Dnot for "没", Dornot for "还是不" and "+" refers to connection between words. Each sign in a structure refers to a same word.*

The necessity of the proposed CSS is investigated. Through experiments of analyzing the query logs from Sogou lab, it is shown that near 2% of queries include CSS, for example "好不好" (good or not) or "研究研究" (have a research), etc. In order to eliminate the disturbance of QC, the experiment chooses 50 queries that QC is smaller than 6. The parts of queries that are CSS structures are replaced by synonyms which are not CSS, for example, the phrase "好不好" (good or not) replaced with "怎样" (how) or "情况" (what thing), etc. The searching experiments are carried out for queries that include CSS and do not include CSS respectively. And the similarities between queries and corresponding returned top 30 results are recorded, and the average similarities are calculated. Figure 2 shows the effect of CSS on the search results.

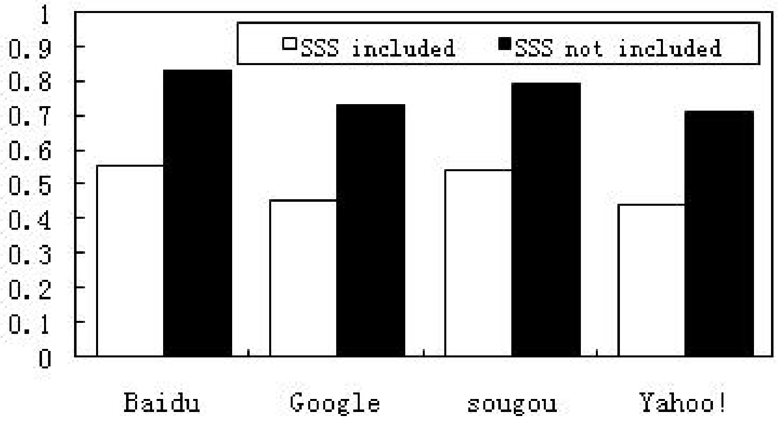

Fig. 2: Effect of SSS to search results' relevance

It can be seen from the figure that the queries that don't include CSS can result in higher correlation searching results than those include CSS. And the four search engines return higher correlated results by more than 30%, among which the average result correlation of Baidu increases by 33.3%, Google by 37.8%, Sogou by 31.2% and Yahoo! by 37.5%. So the complex semantic structure CSS has a certain influence on the searching effect.

According to the above analysis, it can be found that query complexity QC and complex semantic structure CSS are two main factors which influent queries' search effects. Because the complex query have one or both of the two elements, the searching results are likely to be worse. So it is of significant to identify complex queries for search engines further processing.

3.3 Complex queries

Definition 3 *Definition 3 Complex Queries, which matches one or more elements of the set CSS and whose QC is larger than N. The above definitions show that QC and CSS can be used* *as the judgment standards of complex queries. This paper starts with complex queries and aims to recognize them. The ones which do not satisfy the judgment standards are classified as simple queries.*

4 Query Complexity Analysis Based on Dynamic Learning

The query analysis process consists of three steps: Chinese word segmentation, part-of-speech tagging and query type identification. Among them the first two steps use ICTCLAS software toolkit version 2.0 developed by Institute of Computing Technology (http://ictclas.org/). It cannot only support word segmentation function, but also can give part-of-speech tagging and unknown word recognition. Its segmentation precision can reach as high as 97.58%. Besides, it supports custom segmentation and part-of-speech tagging rules, which can achieve effective segmentation results. In this paper, the adverb “不” is defined as Dno, “没” as Dnot, and “还是不” as Dornot. After word segmentation and part-of-speech tagging steps, the next is query type identification. In this step, the QC value N is computed first after removing the words which are auxiliaries or interjections in the query. And the segmentation result is called query_div. Then the similarity between query_div and complex semantic structure CSS can be worked out using formula 1.

$$sim(query\_div) = \frac{\sum\limits_{term_j \in query\_div} (\sum\limits_{s_i \in SSS} same(s_i, term_j))}{|\sum\limits_{term_j} term_j| * |\sum\limits_{s_i \in SSS} s_i|} \tag{1}$$

$$in\ which \quad same(s_i, term_j) = \sum_{k=1}^{n} (tw_{ik} * tw_{jk})/(\sqrt{\sum_{k=1}^{n} tw_{ik}^2}\sqrt{\sum_{k=1}^{n} tw_{jk}^2}) \tag{2}$$

In the above formulas, s_i is one of the elements in the collection CSS, and $term_j$ is part of query_div. The similarity between s_i and $term_j$ $same(s_i, term_j)$ is evaluated by using cosine of vector inclination method in formula 2. If s_i and $term_j$ are exactly the same, the $same(s_i, term_j)$ value is 1. In the formula, tw_{ik} is the weight of the kth words of s_i, and n is the size of the word vector space. In CSS, the using frequencies of all elements are different, so the weights of different elements are needed to be calculated. The QC should be calculated according to the weight from high to low with matching method. The formula is as formula 3 below.

$$tw_i(s_i) = \frac{n(s_i)}{\sum\limits_{s_i \in SSS} n(s_i)} \tag{3}$$

In the formula, $n(s_i)$ is the number of times that s_i has been input by users, and is the number of sum of the times of all elements s_i, and $f(s_i)$ is the using frequency of s_i. The value of $sim(query\_div)$ is restricted. The least similarity between query and CSS is query_limit. And if the query query_div includes one of the element of CSS, the $sim(query\_div)$ is larger than query_limit, otherwise the value is smaller than query_limit.

The least similarity between s_i and $term_j$ is term_limit. The query structures which have larger similarities than term_limit but smaller than 1 and are also excluded in CSS are recorded. And its frequency *Ftemp* is increased. Then the query is checked whether it can result in good feed back by search engines. For the structures which have low search similarities, they are added into CSS dynamically. *Ftemp* is calculated by formula 4.

$$Ftemp(y) = n(y)/N \tag{4}$$

In the formula, y is a substructure of query_div, $n(y)$ is the number of using times of y in a certain period of time. N is the number of times a user searches in a certain period of time. The process of the query complexity analysis is shown as figure 3 below.

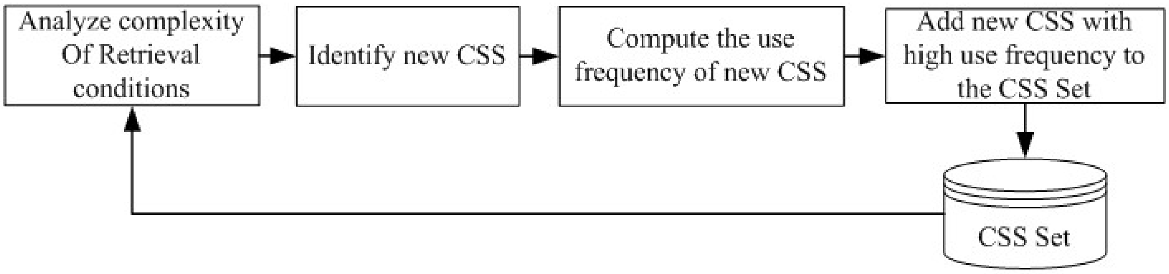
Analyze complexity Of Retrieval conditions
Identify new CSS
Compute the use frequency of new CSS
Add new CSS with high use frequency to the CSS Set
CSS Set

Fig. 3: Dynamic learning process of new CSS

Because of the complexities of the Chinese sentence patterns and language development, some elements that don't exist in CSS may become quite frequently used query inputs by new users. So it is necessary to update CSS dynamically using dynamic learning mechanism. With the above algorithm, the query complexity is analyzed, and new structures that have high similarities are recorded together with their frequencies of occurrence. And then the search effects of the new structures are examined mainly by similarity measurement. The structures that have low similarities and bad search effects are added into CSS. Through the steps above the dynamic learning of complex query structures is accomplished.

5 Experiment Analysis

The query logs from Sogou lab are manual labeled according to the definition of the two kinds of queris, simple and complex ones. From the labeled queries, 50, 100, 150, 200, 250 and 300 are randomly selected respectively for the two kinds, and they are mixed as the experimental data for complex query identification. The identification rates are recorded as shown in figure 4.

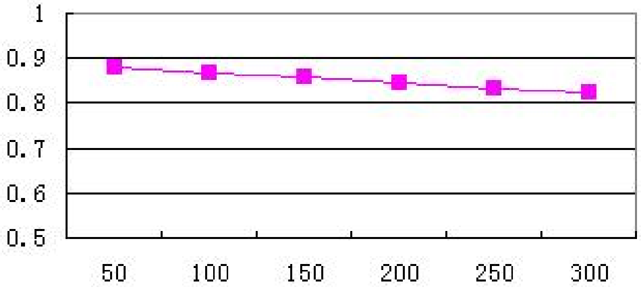

Fig. 4: Recognition accuracy of complex query

It can be seen from the figure that the identification effect is the best at the rate of close to 90% for 50 groups of the two kinds of queries. With the incensement of the test data, the identification rate is gradually decreasing. This is mainly because that some queries are proper nouns and idioms which are segmented incorrectly, and through complementing of the dictionary the effect may be improved. However, the whole identification effect is ideal with the identification rates of 80% or more for each group.

Next is the verification of the learning effect of new complex query structures that are not included in the current CSS. 300 queries are randomly chosen from the query logs as the test set and new queries with new uncommon complex query structures are added into it. Although this kind of queries has low appearing frequency, the total query numbers are quite large for the enormous quantity of search requests. In the experiment, 5%, 10%, 15%, 20%, 25% and 30% new queries with complex query structures are added into the test set respectively. The experimental results are as shown in figure 5.

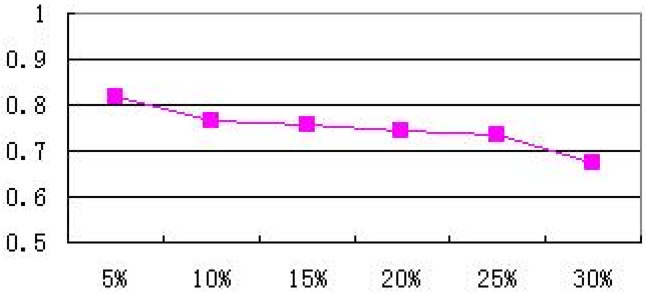

Fig. 5: Learning efficiency of new CSS

The figure shows that the learning efficiency of new complex semantic structures for each proportion. It is found that the learning rate can achieve up to 70% when new complex semantic structures have 5% to 25% proportion in the test sets, while the rate falls to 68.3% at a 30% proportion. That is mainly because that the proposed learning mechanism is based on the original CSS set which is expanded with similar new structures. For structures that have low similarities with the original CSS, it is some difficult to be identify them, and the identification efficiency is likely to decline. However, this kind of query structures has a comparatively low appearing rate, so it causes minor affections to the learning efficiency.

6 Conclusion

Research on factors that influents search effect of search engines is always one of the hot topics in information retrieval. And the processing ability of complex queries is an important factor that affects the searching results. This paper puts forward an identification method of complex queries base on dynamic learning mechanism. It first gives the definition of complex queries and then proposes two searching efficiency influence factors, complexity and structures of complex queries. Using dynamic learning mechanism and similarity matching algorithms, it can identify complex queries. The necessities and necessity of proposing query complexities and complex semantic structures are investigated. At last the propose query classification algorithm and dynamic learning mechanism prove to be effective and feasible.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (61173026), the National High Technology Research and Development 863 Program of China (2012AA02A603), the Fundamental Research Funds for the Central Universities of China, and the Defense Pre-Research Project of the 'Twelfth Five-Year-Plan' of China (513***301).

References

[1] ShiXi Fan. Combination of rough set theory and maximum entropy model for conjunctive structure detection in QA system, Proceeding of the Sixth International Conference on Machine Learning and Cybemetics, pp. 3051 – 3056. IEEE Press, Hong Kong, 2007.

[2] XI Jianqing. Research on Automatic Identification for Chinese Prepositional Phrase Based on HMM, Computer Engineering, 33 (2007), 172 – 173.

[3] Xu jian. Syntax Parsing of Contemporary Chinese Based on Hidden Markov Model, Computer Engineering and Applications, 27 (2003), 09 – 112.

[4] Harabagiu S. Answering complex questions with random walk models, Proc of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 220 – 227. ACM, Seattle 2006.

[5] G. Kumaran. Effective and efficient user interaction for long queries, In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information, pp. 11 – 18. ACM Press, Singapore 2008.

[6] G. Kumaran. Carvalho, Reducing long queries using query quality predictors, Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 19 – 23. ACM Press, Boston, MA, USA 2009.

[7] Jansen BJ. Real Life Information Retrieval: A Study of User Queries on the Web. ACM SIGIR Forum, pp. 5 – 17. ACM Press, 1998.