Journal of Southeast University (English Edition) Vol. 22, No. 3, pp. 324 – 329 Sept. 2006 ISSN 1003—7985

Concept-based approach for information retrieval

Wu Chen[1,2] Zhang Quan[1] Jia Ning[1,2]

([1] Graduate School, Chinese Academy of Sciences, Beijing 100039, China)

([2] Institute of Acoustics, Chinese Academy of Sciences, Beijing 100080, China)

Abstract: A concept-based approach is expected to resolve the word sense ambiguities in information retrieval and apply the semantic importance of the concepts, instead of the term frequency, to representing the contents of a document. Consequently, a formalized document framework is proposed. The document framework is used to express the meaning of a document with the concepts which are expressed by high semantic importance. The framework consists of two parts: the "domain" information and the "situation & background" information of a document. A document-extracting algorithm and a two-stage smoothing method are also proposed. The quantification of the similarity between the query and the document framework depends on the smoothing method. The experiments on the TREC6 collection demonstrate the feasibility and effectiveness of the proposed approach in information retrieval tasks. The average recall level precision of the model using the proposed approach is about 10% higher than that of traditional ones.

Key words: information retrieval; concept; semantic knowledge; content representation

Since the first formal model for IR[1] came into being in 1968, a number of information retrieval (IR) models have been developed, such as vector spaces, probabilities, fuzzy logics, language and so on[2–5]. Most IR models are based on matching the query keys which are regarded as the retrieval needs of information seekers to a document collection's representation of content and then presenting the retrieval set as a ranked hierarchical list. The document collection's representation of content is always represented based on the term frequency (TF) and seldom on the content of the documents. These models have generated many useful systems, but they are essentially lacking due to the fact that they disregard the context and do not attempt to resolve the meaning of the terms. The ubiquitous existence of word sense ambiguities clouds the behavior of IR systems which disregard the semantic differences in words. Therefore, how to determine the sense of the word and the meaning of the context becomes very important. Meanwhile, the studies on cognitive science show that people understand entities by comprehending the concepts represented by the entities. The language works in the same way[6]. During these years, the conceptual expressions of language have been investigated[7–8] and a symbolic system has been formed in order to express words and sentences[6,9–10]. Based on this method, words and sentences in the language space can be translated into their conceptual forms represented by a set of meaningful character strings. Consequently, the meaning of the word and the sentence are evaluated. However, this theory does not address how to express the meaning of a sentence group via a formalized framework, which is necessary in order to express the meaning of a document and to serve the IR task. The problem of how to match the queries to the documents via their conceptual forms is not solved either. This paper will discuss these issues.

1 Related Work

In the studies of using the conceptual expressions to express words and sentences, the HNC (hierarchical network concept) theory[6,9] has been introduced, which involves some new concepts presented below.

One important concept is the semantic chunk[6] which is a semantic unit between words and sentences. But it is different from phrases and other traditional chunks. It is a semantic unit expressing a comprehensive concept. We classify the semantic chunk into two types: the main chunk and the supplementary chunk. The main chunk contains the necessary parts of a sentence. They are used to describe the objects and their actions in the sentence. The supplementary chunk provides the background knowledge of a sentence, such as time, place, etc. Sentence category is another concept. It includes a set of symbolic expressions. These expres-

Received 2006-04-10.

Foundation items: The National Basic Research Program of China (973 Program) (No. 2004CB318104), the Knowledge Innovation Program of Chinese Academy of Sciences (No. 13CX04).

Biographies: Wu Chen(1979—), male, graduate; Zhang Quan (corresponding author), male, doctor, professor, zhq@mail.ioa.ac.cn.

sions are named sentence category expressions. Each expression contains the expressions of chunks and some conjunctive symbols which are used to describe the relationships among chunks. These expressions are designed in advance and able to describe not only the meaning but also the structure of the sentence. Huang concluded 57 types of primitive sentence category expressions and 57×56 compound ones[6, 10].

The HNC presents a sentence category analysis technique[6, 11], by which the conceptual forms (sentence category expressions with regard to sentences; word concepts to words) of a sentence and its words can be automatically extracted. The word concept is the conceptual expression of a word sense. It is expressed by a meaningful character string.

2 Content Framework

Content framework is a formalized structure which is used to express the content of a sentence group. A sentence group is defined as a group of sentences, which focus on one subject. A document can be represented by a set of content frameworks which are derived from the sentence groups within the document.

We regard full stop as the signal for the end of the sentence group. This mechanical method can predigest the difficulties in measuring off the sentence groups. Apparently, a complex sentence will be always regarded as a sentence group. The content framework involves three parts: domain (DOM), situation (SIT) and background (BAC).

Domain describes the types of an event, or its characters. It is classified into 108 categories[9]. This classification can be used to classify the texts.

Situation explains the domain in detail. It denotes, from the perspective of content, the relationship among the objects involved in the event. Situation contains, from the perspective of form, the expressions of the semantic chunks. The semantic chunks are derived from the sentence category expressions of the sentences within the sentence group. The sentence category expression which contains the domain-related concepts are adopted.

Background explains the time, place etc. in which the event occurs. The information of the background is provided by the supplementary chunks in the sentence.

Considering the three parts of the content framework, we design a structure with two n-dimensional vectors to express the document framework. These vectors are defined as

$$\begin{aligned} \mathrm{DOM} &= \{\mathrm{dom}_1, \ldots, \mathrm{dom}_n\} \\ \mathrm{SIT}\ \&\ \mathrm{BAC} &= \{\mathrm{cs}_1, \ldots, \mathrm{cs}_m\} \end{aligned} \tag{1}$$

where dom_k is the expression of domain k the sentence group covers, which is consistent with a certain category of the domain; cs_k is the k-th word concept. We can see that the contents of SIT and BAC are combined into one vector since there is no difference between these two parts in the probability of content to be retrieved.

3 Translating Documents to Their Formalized Content Frameworks

The main idea is to obtain the content framework of each sentence group in the document via some semantic methods, and then obtain the document framework via some statistical methods based on the content framework.

The content framework is proposed to formalize the content of a sentence group. Before we can obtain the content framework, the sentence category expressions and word concepts should be extracted first. The whole procedure to fulfill these tasks can be divided into three stages: the "word sense and sentence category extracting", the "content framework generation" and the "document framework generation". This is shown in Fig. 1.

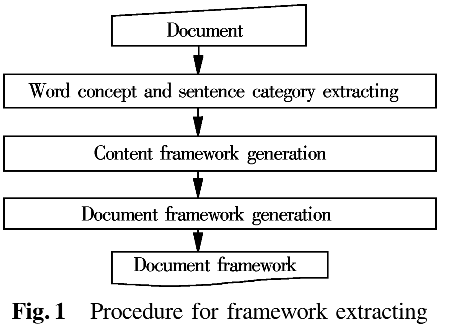

Fig. 1 Procedure for framework extracting

The semantic approach is adopted in the first two stages. The statistical approach is adopted in the last stage. Through the first stage, the sentence category expression and the word concept can be obtained. The second stage generates the content frameworks based on the results the first stage provides. The last stage combines the content frameworks as the document framework.

3.1 Word sense and sentence category extracting

This extracting approach adopts the HNC sentence category analysis technique[6, 11]. The approach mainly focuses on the processing logic which tells the computer how to obtain the sentence category expression and the word concept through some specific procedures. Restricted by the length of the article, we only give one example to explain the function of the approach. The processing details can be referred to Ref. [11].

Example 1 For six months ~ ‖I‖ lived ‖ ~ without a job ~ ‖ in New York.

After processing, we can obtain: ① The sentence category expression of this sentence is "CN1CN2S03J", where S03 means this sentence is a transposition state sentence due to "live" being a certain kind of "state". The structure of S03 (defined in advanced) is SB + S + SC. ② The main trunks of this sentence are "I", "lived in" and "New York". "lived in" is the kernel of the main chunk (named S under the structure of S03), "I" and "New York" are two general main chunks (named SB and SC, respectively). The supplementary chunks are "for six months" and "without a job", which act as conditions (CN). Hence we add signs CN1 and CN2 in front of the sentence category expression. ③ The conceptual expression of "I" in this sentence is "p4001"; "live" is "v65500214"; "New York" is "fpj2 ∗ 304/fpwj2 ∗ m1"; "six months" is "j3080c06/wj10-0"; "without a job" is "jlvu116 | ra00e21".

3.2 Content framework generation

The generation approach can be divided into three sub-stages. Including:

1) Measure the DOM based on the categories of domain discussed in section 2. The value of the domain is determined by the occurrences of the word concept in the main chunks, if the word concept is related to a certain category of the domain. If more than one concept has the same number of occurrences, the priority of main chunks to determine the value of domain is as follows:

$$\mathrm{Eg} > \mathrm{El} > \mathrm{C} > \mathrm{B} \tag{2}$$

where Eg is the kernel of the main chunks in the main sentence, or in the main clause if the sentence group is a complex sentence; El is the kernel of the main chunks in the subordinate sentence, or in the subordinate clause; C and B are the two general main chunks. C means the content; B means the object. For example, in the sentence, I have dinner, I is the object, dinner is the content.

2) Obtain the SIT based on the sentence category expression. The SIT consists of a set of word concepts. These word concepts are derived from the main chunks of the sentences which contain the word concepts that strengthen the current DOM.

3) Obtain the BAC via integrating the supplementary chunks.

We use example 1 again to explain this approach.

Because example 1 is a simple sentence, occurrences of each concept in the sentence are the same. Consequently, we should take advantage of Eq. (2) to determine the DOM. In example 1, the kernel of main chunks, "lived in", should be considered first. The concept of "live" is "v65500214" which is related to the 8th category. So the DOM of this sentence can be ascertained, and the value of DOM is 65.

There are three main chunks in the sentence, so the SIT should contain three elements. There are "v65500214" derived from chunk S, "fpj2 ∗ 304/fpwj2 ∗ m1" derived from chunk SB and "p4001" derived from chunk SC.

The BAC should contain two elements according to the supplementary chunks of the sentence. They are "j3080c06/wj10-0" derived from chunk CN1 and "jlvu116 | ra00e21" from chunk CN2.

Consequently, the content framework of the sentence in example 1 is produced as follows:

DOM = {65}

SIT & BAC = {v65500214, fpj2 ∗ 304/fpwj2 ∗ m1, p400, j3080c06/wj10-0, jlvu116 | ra00e21}

3.3 Document framework generation

Document framework is a vector space model with two n-dimensional vector spaces. A document can be expressed as a combination of two vectors:

$$V_{\mathrm{DOM}}(d) = \{\mathrm{DOM}_1, W_{D1}(d); \ldots; \mathrm{DOM}_n, \mathrm{W}_{Dn}(d)\}$$

$$V_{\mathrm{SIT\,\&\,BAC}}(d) = \{C_1, W_{C1}(d); \ldots; C_n, W_{Cn}(d)\}$$

where $V_{\mathrm{DOM}}(d)$ means the DOM vector of the document; $V_{\mathrm{SIT\,\&\,BAC}}(d)$ refers to the SIT & BAC vector of the document; DOM_n denotes the n-th domain appearing in the document collection; $W_{Dn}(d)$ is the weight of the DOM_n in the document, a measurement of occurrences of the DOM_n in the content frameworks within the document; C_n means the n-th word concept appearing in the SIT or BAC; $W_{Cn}(d)$ is the weight of concept C_n in the document d, a measurement of occurrences of C_n in the SIT and BAC of the content frameworks within the document. We always regard the DOM_k which has the highest $W_{Dk}(d)$ as the domain category that the document belongs to.

4 Computation of Document Query Similarity

We propose a domain smoothing method to complete the searching task. In the method, the domain likelihood is computed. The method is based on the two-stage smoothing method[12]. Using of an explicit and uniform domain language model makes it possible to make the matching become more accurate and assign the extra probability mass to the queries which do not contain the information of domain or appear in a certain document.

The method needs to transfer the queries into their conceptual forms. Users are always used to entering a set of separate query keys as the query. It is difficult to obtain the conceptual expression of each query key exactly since separate query keys always lack strong phrase ological relationships with each other. Therefore, we use the proportion of each concept candidate of a word to weigh the probabilities caused by each candidate.

The target function of the domain smoothing method is as follows:

$$P(Q \mid d_i) = \prod_{q_j \in Q} \sum_n P(q_j, w_{jn}) \left[\lambda P(w_{jn} \mid Q_\mathrm{S}) + (1-\lambda) P(w_{jn} \mid Q_\mathrm{D}) \right] \tag{3}$$

$$P(w_{jn} \mid Q_\mathrm{S}) = \frac{n_{\mathrm{SIT\,\&\,BAC}}(w_{jn}, S)}{\sum_w n_{\mathrm{SIT\,\&\,BAC}}(w, S)} \tag{4}$$

$$P(w_{jn} \mid Q_\mathrm{D}) = \frac{n_{\mathrm{SIT\,\&\,BAC}}(w_{jn}, d_i) m(\mathrm{DOM}(w_{jn}), d_i) + \mu P(w_{jn} \mid Q_\mathrm{DOM})}{\sum_w n_{\mathrm{SIT\,\&\,BAC}}(w, d_i) \sum_l m(\mathrm{DOM}_l, d_i) + \mu} \tag{5}$$

$$P(w_{jn} \mid Q_\mathrm{DOM}) = \sum_k P(w_{jn} \mid \mathrm{DOM}_k) P(\mathrm{DOM}_k \mid d_i) \tag{6}$$

where $P(w_{jn} \mid Q_\mathrm{S})$ is a collection language model; $P(w_{jn} \mid Q_\mathrm{D})$ is a document language model containing the domain language model $P(w_{jn} \mid Q_\mathrm{DOM})$; $P(q_j \mid w_{jn}) = \frac{n(w_{jn}, q_j)}{n(w_{jn}, S)}$, $P(q_j \mid w_{jn})$ is the probability of word q_j translating to concept w_{jn}; $n(w_{jn}, q_j)$ is a measure of how many concepts w_{jn} are generated by word q_j; $n(w_{jn}, S)$ is the number of concepts w_{jn} in the document collection S; $n_{\mathrm{SIT\,\&\,BAC}}(w_{jn}, d_i)$ is a measure of the number of occurrences of concepts w_{jn} in the SIT and BAC of the content framework of the document d_i; $\sum_w n_{\mathrm{SIT\,\&\,BAC}}(w, d_i)$ is the number of concepts in the SIT & BAC of the content framework of the document d_i; $m(\mathrm{DOM}(w_{jn}), d_i)$ is a measure of the number of the content frameworks in which the DOM value is supported by the concepts w_{jn}; $\sum_l m(\mathrm{DOM}_l, d_i)$ is the number of the DOM in documents d_i; $P(\mathrm{DOM}_k \mid d_i) = \frac{m(\mathrm{DOM}_k, d_i)}{\sum_l m(\mathrm{DOM}_l, d_i)}$, $P(\mathrm{DOM}_k \mid d_i)$ is the probability of document d_i belonging to a certain domain DOM_k under the designed distribution of the domain; $P(w_{jn} \mid \mathrm{DOM}_k) = \frac{n_{\mathrm{SIT\&BAC}}(w_{jn}, \mathrm{DOM}_k)}{\sum_w n_{\mathrm{SIT\&BAC}}(w, \mathrm{DOM}_k)}$, $P(w_{jn} \mid \mathrm{DOM}_k)$ is a measure of the frequency of the key concept w_{jn} in the documents which belong to domain DOM_k. The belongingness is determined by occurrences of the DOM_k within the document.

5 Experimental Results

We now provide experimental results to illustrate the behavior of our model (namely Model X). Due to the restriction of the word knowledge base, Model X can only serve the Chinese information retrieval at present. Our experiments are made in a Chinese circumstance.

The test collections are chosen from TREC6, which are originally designed for both the Chinese monolingual information retrieva1 and English-Chinese cross-language information retrieval. It is 170 MB as raw text. There are 26 topics constructed, and the topics are supplied in both English and Chinese, In our experiments, we just use the Chinese topics.

The value of parameters μ and λ should be evaluated. The results obtained by different queries are affected by different parameters. We test a set of values of μ, including 300, 500, 800, 1 000, 1 500, 3 000, 5 000, 8 000 and a set of values of λ, including 0.000 1, 0.001, 0.005, 0.01, 0.04, 0.1, 0.4. The highest performance of the experimental system occurs when μ is equal to 3 000 and λ is equal to 0.01, so we choose 3 000 and 0.01 as the standard value of μ and λ.

Two other IR models are also tested in our experiment for comparison. One is the Jelinek-Mercer model-based IR system based on term frequency, and the other is the Bayesian model-based IR system based on term frequency. The test results of these three approaches and the average precision over all relevant docs are given in Tab. 1.

The results are compared in Fig. 2, which is similar to ROC (receiver operating characteristic). A larger area below the curve indicates the higher performance of the system.

Tab. 1 Recall level precision %

Recall	Precision		
	Jelinek-Merce	Bayesian	Model X
0	70.00	70.00	82.00
0.1	64.16	65.21	71.02
0.2	58.12	52.56	59.41
0.3	50.28	52.29	57.43
0.4	48.37	50.42	55.76
0.5	35.86	37.76	47.89
0.6	31.03	35.21	45.42
0.7	22.27	27.31	32.58
0.8	18.34	22.37	22.37
0.9	14.91	16.87	19.12
1.0	6.12	6.10	6.16
Average	34.95	36.61	45.38

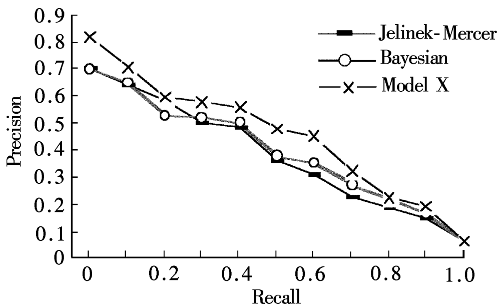

Fig. 2 Comparison of results

The experiment shows that the precision of Model X increases more evidently than the other two systems when the recall decreases. The reason for obtaining the superior results of Model X is that the word sense ambiguities are better solved by translating the words into their concept forms via the strategies of word sense and sentence category. Another reason is that Model X distinguishes the concepts which have different semantic importance in the formalized content framework. The content framework can greatly help the searching method to discriminatingly process every concept which appears in different positions of the sentences and disregard the isolated word concepts.

We also find that some inappropriate documents that consist of many complex sentences are inevitably involved in the query results of Model X due to the fact that the complex sentences cannot be analyzed accurately. It is an accessional factor that depresses the system behavior. One factor is the performance of the sentence category analysis system; the other is the approach used to tackle the new words in the documents and queries. The latter has been settled by finding the words via concept extracting strategies and processing the new words as integrated concepts.

6 Conclusion

Formalizing the content of a document and resolving the word sense ambiguity are two important issues in information retrieval. The IR approach we present in this paper is a combination of, and takes full advantage of, the semantics and the statistics. This approach can well resolve word sense ambiguity and solve the problem concerning equivalent words. This approach abstracts the meaning of a document using two concept vector spaces based on the content framework. One vector focuses on the domain category of the document. Another focuses on the content detail. A smoothing method considering the two vector spaces is presented in this paper. A domain language model is introduced into the smoothing method for bridging the gap between the document model and the collection model. Consequently, each document can be used to compute the query likelihood via the domain language model.

Acknowledgement The authors would like to thank Wei Xiangfeng, Miao Jianming, George Qian, Huang Zengyang, Anette Frank and the anonymous reviewers for helpful comments on this work.

References

[1] Salton G. *Automatic information organization and retrieval* [M]. New York: McGraw-Hill, 1968.

[2] Crestani F, Pasi G. *Soft computing in information retrieval* [M]. Germany: Physica Verlag and Co, 2000. 102 − 121.

[3] Lalmas M. Logical models in information retrieval: introduction and overview [J]. *Information Processing and Management*, 1998, **34**(1): 19 − 33.

[4] Miyamoto S. *Fuzzy sets in information retrieval and clustering analysis* [M]. Kluwer Academic Press, 1990.

[5] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval [J]. *Information Processing and Management*, 1988, **24**(5): 513 − 523.

[6] Huang Zengyang. *HNC (hierarchical network concept) theory* [M]. Beijing: Tsinghua University Press, 1998. (in Chinese)

[7] Schank R. Identification of conceptualizations underlying nature language [A]. In: Schank R, Colby K, eds. *Computer Models of Thought and Language*[C]. San Francisco, CA: W H Freeman Company, 1973. 187 − 247.

[8] Schank R. *Conceptual information processing*[M]. Amsterdam: North Holland, 1975.

[9] Huang Zengyang. *Mathematics and physics symbol system of language in language concept space* [M]. Beijing: Ocean Press, 2004. (in Chinese)

[10] Miao Chuanjiang. *Guide of HNC (hierarchical network concept) theory* [M]. Beijing: Tsinghua University

Press, 2005. (in Chinese)

[11] Wei Xiangfeng. The software platform for expanded sentence category analysis based on the HNC theory [D]. Beijing: Chinese Academy of Sciences, 2005. http://www.hncnlp.com/Abs/absEwxf.htm. (in Chinese)

[12] Zhai C, Lafferty J. Two-stage language models for information retrieval [A]. In: *Proceedings of the 25th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* [C]. Tampere, Finland, 2002. 49 - 56.

一种基于概念的信息检索方法

吴 晨[1,2] 张 全[1] 贾 宁[1,2]

([1] 中国科学院研究生院,北京 100039)

([2] 中国科学院声学研究所,北京 100080)

摘要:为了获取词语在文章中的语义权重,解决词语的同义、多义模糊问题,提升信息检索的效率,提出了一种基于概念的检索模型,模型中设计了一种形式化的文本内容表示框架,框架由2部分构成:文章的“领域”以及“情景与背景”信息,并由概念(形式化语义)加以表示. 同时,提出了提取该概念框架的方法,给出了用于框架与检索要求间匹配的两阶段平滑算法. 实验表明,在TREC6提供的小规模语料集下,采用所提出方法的信息检索模型与传统模型相比,平均召回准确率提升了约10%,效果显著,充分说明了基于本文描述方法构建的、以概念作为处理中介的信息检索系统的有效性和可行性.

关键词:信息检索;概念;语义知识;内容表示

中图分类号:TP393