주 키워드와 부 키워드를 이용한 자연언어 정보 검색 모델

강현규[†] · 박세영[††]

요 약

정보 검색이란 사용자의 정보 요구를 만족하는 관련 정보를 검색하는 것이다. 그러나 정보 검색 시스템의 하나의 역할은 관련 정보의 집합들을 단순히 제시하는 것이 아니라 주어진 요구 사항에 가장 가까운 문서를 결정하는데 도움을 주는 것이다.

최근에 여러 가지 텍스트 분석 시스템들에서 내용을 인식하기 위해 구문 분석 방법 사용이 시도되고 있다. 불행히도 단독의 구문 이해 방법으로는 임의의 텍스트 예들을 완벽하게 분석하기 위해 불충분한 것으로 알려지고 있다.

이 논문에서는 2단계 문서 순위에 기반한 문서 순위 결정 방법에 대하여 논한다. 1단계는 문서를 검색하기 위해 사용하고 2단계는 검색된 문서를 재순서화하는데 사용한다. 1단계에서 이용된 주키워드는 문서를 구별할 수 있는 좋은 능력을 가지는 명사나 복합명사로서 정의될 수 있다. 2단계에서 이용된 부 키워드는 주키워드나 기능어가 아닌 형용사나 부사 또는 동사로 정의 될 수 있다.

실험은 23,113 항목을 가지는 한국어 백과사전과 일반 사용자들로부터 수집된 161개의 한국어 자연언어 질의로부터 이루어졌다. 자연언어 질의의 85%가 부 키워드를 가지고 있었다. 2단계 문서 순위 방법은 일반 문서 순위 방법보다 현격한 검색 효율의 향상을 제공한다.

A Model of Natural Language Information Retrieval Using Main Keywords and Sub-keywords

Hyun-Kyu Kang[†] · Se-Young Park[††]

ABSTRACT

An Information Retrieval (IR) is to retrieve relevant information that satisfies user's information needs. However a major role of IR systems is not just the generation of sets of relevant documents, but to help determine which documents are most likely to be relevant to the given requirements.

Various attempts have been made in the recent past to use syntactic analysis methods for the generation of complex construction that are essential for content identification in various automatic text analysis systems. Unfortunately, it is known that methods based on syntactic understanding alone are not sufficiently powerful to produce complete analyses of arbitrary text samples.

In this paper, we present a document ranking method based on two-level ranking. The first level is used to

[†] 종신회원: 한국전자통신연구원 자연어처리연구실
[††] 정 회 원: 한국전자통신연구원 자연어처리연구실

논문접수:1997년 5월 28일, 심사완료:1997년 9월 18일

retrieve the documents, and the second level to reorder the retrieved documents. The main keywords used in the first level can be defined as nouns and/or compound nouns that possess good document discrimination powers. The sub-keywords used in the second level can be also defined as adjectives, adverbs, and/or verbs that are not main keywords, and function words.

An empirical study was conducted from a Korean encyclopedia with 23,113 entries and 161 Korean natural language queries collected by end users. 85% of the natural language queries contained sub-keywords. The two-level document ranking methods provides significant improvement in retrieval effectiveness over traditional ranking methods.

1. Introduction

The basis for IR systems consists of determining the absence or presence of keywords (index terms or terms) in conjunction with their counting and distribution information. However a major role of IR systems is not just to generate a set of relevant documents, but to help determine which documents are most likely to be relevant to the given requirements. IR systems should present to users a sequence of documents which are ranked in decreasing order of query-document similarity. Users are able to minimize their time spent to find useful information by reading the top-ranked documents first. Therefore, the document ranking method is an important component of IR systems [4].

Index tasking is obviously crucial for retrieval, because failures in the indexing policies immediately lead to retrieval failures. Indeed, if the indexing is insufficiently exhaustive - that is, if the chosen index terms do not properly reflect all the subject areas covered by a given document - it may not be possible to retrieve a document when it is needed. Various attempts have been made in the recent past to use syntactic analysis methods for the generation of complex construction, such as noun and prepositional phrases, that are essential for content identification in various automatic text analysis systems [11-14]. Unfortunately, it is known that methods based on syntactic understanding alone are not sufficiently powerful to produce complete analyses of arbitrary text samples. No matter how the problem is simplified, the analysis of noun phrase constructions, which is chiefly needed in information retrieval, is especially difficult, and all the various attempts to come up with general rules for noun phrase understanding have been unsuccessful [15].

We propose a ranking optimization of IR systems for large data banks of collected and stored information. In this paper, we focus on "two-level document ranking (two-level ranking)," through which documents are retrieved using main keywords first, and then documents (re-ranking) reordered by search retrieved documents with sub-keywords.

Main keywords are nouns, proper nouns or compound nouns, which offer good document discrimination powers. Discrimination power is to compute the "discrimination value" of a keyword. It measures the degree to which the use of the keyword will help to distinguish the documents from each other [16]. The good discriminators whose introduction for indexing purpose decreases the space density. The poor discriminators whose utilization renders the documents more similar. In particular, when a good discriminator is assigned to the documents of a collection, the few items to which the keyword is assigned will be distinguished from the rest of the collection. Sub-keywords are adjectives, adverbs, and/or verbs which are not main keywords, and function words.

So, we are going to improve the retrieval effectiveness by reordering of ranking using two-level ranking method. An empirical study was performed using a Korean encyclopedia with 23,113 entries [17] and 161 natural language queries collected by end users. It has been implemented using Microsoft

Viaual C^{++} (MS VC^{++}) language for indexing and retrieval in PC Windows NT environment.

In this paper, we give a motivation first. In section 3, we present the features of main keywords and sub-keywords, and analyze natural language queries. The two-level ranking method is explained in section 4. In section 5, we describe our test collection, give an example that how to operate our system, and some experiment results are presented therein. Finally, we conclude with our main findings, and point out future research directions.

2. Motivation

Boolean search strategies are used in most commercial text retrieval system, but their drawbacks are well known [1, 18]. For example, formulation of useful Boolean queries is not easy to learn, and users often have difficulty in controlling the output size. To improve retrieval effectiveness, IR system allows a query to be expressed as a natural language describing the user's information need. Thus the description can be treated as a short document. The main reason the natural language/ranking approach is more effective for end-user is that all the terms in the query are used for retrieval, with the results being ranked based on co-occurrence of query terms, as modified by statistical term weighting. But the primary difficulty in natural language processing is due to the flexibility, and by extension the ambiguity of most languages [9]. Unfortunately, it is impossible to go beyond the general realization that the language analysis task is difficult.

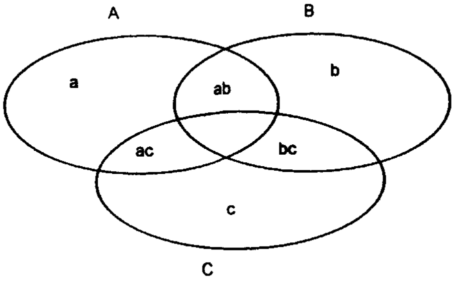
In the Korean language (Hangul), keywords with good document discrimination powers are nouns, proper nouns, and compound nouns. Numerals, pronouns, prepositions, adverbs, adjectives, and verbs do not distinguish documents alone. Additionally, declinable word like verbs or adjectives have many irregular and inflections forms. In the Korean language, especially, we sometimes find that one word has 20-30 dictionary meanings. Because there are many inflected forms and ambiguities, Korean morphological analysis is not easy. So, most Korean IR systems employ keywords which are noun-group, including nouns, proper nouns and compound nouns. Therefore, these IR systems may give improper and inaccurate information to the users. This happens by reason of homonyms or frequency counts of meaningless keywords. The IR systems ignore the characteristics and constraints of keywords. From these observances, we have introduced the main keyword and sub-keyword concept; one that has changed retrieval concept by using two-level ranking method.

Current information retrieval systems calculate the frequency of the keywords. Therefore, the systems may give improper results to users. For example, if a user poses the query, 〈〈남극에 사는 동물은 ?〉〉 (Nam-Guk-Ye (noun +post-position) Sa-Nun (verb +post-position) Dong-Mul-Un (noun +post-position)? : in-the-Antarctic live what-animals ? : What animals live in the Antarctic ?), the IR system extracts the keywords 〈〈남극〉〉 (Antarctic), and 〈〈동물〉〉 (animal) calculates their weights and selects documents according to the keyword weights. The IR system might have to give context documents, 〈〈펭귄〉〉 (penguin), 〈〈고래〉〉 (whale), 〈〈남극〉〉 (Antarctic), and so on, to users. The system presents, 〈〈동물〉〉 (animal), 〈〈식물〉〉 (plant), 〈〈아문센〉〉 (Amundsen), and 〈〈남극 조약〉〉 (Antarctic treaty) context documents, as of higher rank to the user. Because there are many 〈〈동물〉〉 (animal), or 〈〈남극〉〉 (Antarctic) words in their documents, the important information, 〈〈사는〉〉 (live), is ignored. And, if a user poses another query, 〈〈배가 재배되는 시기는?〉〉 (Bae-Ga (noun +post-position) Jae-Bae-Dae-Nun (verb +eomi +post-position) Si-Gi-Nun (noun +post-position)? : pears do-grow when (time) : When do pears grow?), the IR system extracts 〈〈배〉〉 (pear) and 〈〈시기〉〉 (time), and calculates their weights. However, the word 〈〈배〉〉 (pear) has three major homonyms, ship, pear, and belly in the Korean language. The user wants to gain an information on pear, but the system presents documents that have large numbers of the words, ship, belly, and also for 〈〈시기〉〉 (time) in response to the pear keyword. If the system had used the verb information, 〈〈재배되는〉〉 (grow), it might have presented documents related to pears as the highest rank.

In order to provide more precise information, we have introduced the main keyword (nouns, proper nouns, compound nouns) and sub-keyword (verbs, adjectives, adverbs, etc.) concept one that has changed the retrieval concept by using two-level ranking method. It distinguishes keywords according to keyword importance. The main keyword and sub-keyword concept can solve some of the ambiguities posed by nouns and compound nouns.

3. Keywords and Natural Language Queries

Almost all IR systems use keywords as document representatives. These keywords can be defined as sequences of words which have a minimal meaning. (Fig. 1) illustrates the features of keywords in documents. "A", "B", and "C" are documents. A sequence of strings, "a", "b", and "c", are the words in the documents. "a", "b", and "c" are keywords which offer good discrimination powers for each document. On the other hand, "ab", "ac", and "bc" do not discriminate any documents by themselves. But, if "ab" and "b" words are used simultaneously, then we can see that the words represent document B.



(Fig. 1) Documents and words in documents

Main keywords are nouns, proper nouns and/or compound nouns which offer good document discrimination powers. Sub-keywords are adjectives, adverbs, and/or verbs which are not main keywords, and function words. Sub-keywords are not adequate keywords alone. But, if sub-keywords are used together with main keywords then required documents can be retrieved accurately.

To gain empirical experience, we collected about 260 natural language queries from end-users who wanted to extract the content from an encyclopedia. 85% of the collected natural language queries contain sub-keywords.

We observed 218 natural language queries which exclude queries with duplicate or no answers in an encyclopedia from 260 natural language queries posed. For example, excluded queries are "How much is the encyclopedia?", "What were the score of yesterday's professional Baseball games?", "What was the computer graphic technology used in the movie 'Jurassic Park'?", and so on.

We can classify natural language queries into 5 types. Classification is based on the answer characteristics for the natural language queries. Because the purpose of IR systems is to retrieve relevant documents, it is important that classification is based on answer. Type 1 is based on specific facts. It consists of main keywords, and function words. Type 2 is based on facts with attached conditions. Natural language queries are described by modifiers. Type 3 is based on descriptive form restrictions. Type 4 is based on questions and answers. Type 5 is based on yes or no questions. Classified natural language queries are classified as follows (m-k : main keyword, s-k : sub-keyword) ;

① A type of queries about specific facts.

파충류의 생활에 대하여? (Pa-Chung-Ryu-Ui (noun +post-position) Saeng-Whal-Ye (noun +post-position) Dae-Ha-Yeo (wh-adverb)? : of-the-reptiles life about? : About life(m-k) of the reptiles(m-k)?)

②A type of queries about specific objects with attached condition.
세계에서 가장 높은 빌딩은? (Se-Gae-E-Seo(noun +post-position) Ga-Jang (adverb) Nop-Un (adjective) Bil-Ding-Un (noun +post-position)?: in-the-world the-most tall the-building?: The building (m-k) of the most (s-k) tall (s-k) in the world (m-k).)

③A type of queries with descriptive form restriction.
지구가 도는 이유는 무엇인가? (Ji-Gu-Ga (noun +post-position) Do-Nun (adverb) Yi-You-Nun (noun +post-position) Mu-Yeok-In-Ga (wh-adverb)?: the-earth go-round the-reason what-is?: What is the reason(m-k) of go round(s-k) the earth(m-k)?)

④A type of queries for questions and answers.
유럽을 여행하는데 얼마나 돈이 드는가? (Yu-Rub-Ul (noun +post-position) Yeo-Hang-Ha-Nun-De (noun +post-position) El-Ma-Na (adverb) Don-Yi (noun +post-position) Du-Nun-Ga (adverb)?: Europe travel how-much money spends?: How much(s-k) money(m-k) spends(s-k) travel(m-k) Europe(m-k)?)

⑤A type of queries for yes or no questions.
고래는 포유류인가? (Go-Rae-Nun (noun +post-position) Po-You-Ryu-In-Ga (noun +wh-position)? : mammal is-the-whale?: Is the whale(m-k) mammal (m-k)?)

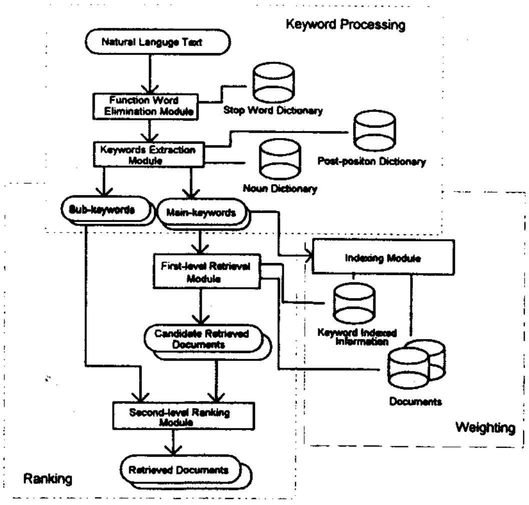
Types 4 and 5 require a more intelligent approach, such as a question and answer system or knowledge-based system. We have evaluated 161 Type 1, Type 2 and Type 3 natural language queries, excluding Type 4 and Type 5.

4. Two-level Ranking Method

Both query and document terms can be weighted, to distinguish terms that are more important for retrieval purposes from these less important. The similarity between a query and the documents can be computed, in order to rank the retrieved documents in decreasing order of the query-document similarity. Many similarity measures have been studied [5, 7-9].

Firstly, we can rank the document order by retrieving the documents from the indexed information, using main keywords (we have called this "First-level ranking"). Then, we reorder ranking by searching retrieved document using sub-keywords (we have called this "Second-level ranking"). At this time, we consider the lengths (window sizes) between main keywords and sub-keywords.

An overall block diagram is shown on (Fig. 2). For indexing, we should extract the keywords in the text documents. Then, we assign weights to the main keywords. For retrieval, we extract the keywords from natural language query first. Then, we rank the documents using the two-level ranking method.



(Fig. 2) Overall system block diagram

4.1 Keyword Processing

Identify individual words occurring in the natural language texts. Use a stop list of common function words to delete the high-frequency function words from the texts. Extracting keywords from large-text data has been an important issue among researchers in building IR systems.

Keywords can be defined as sequences of words which have a minimum meaning. Keywords are assumed to be nouns, proper nouns and/or compound nouns. Keywords can be extracted by referring to nouns and post-positions from noun dictionary and post-position dictionary. The process of keyword extraction produces nouns, proper nouns, compound nouns and keywords as output reference dictionaries. After that, keywords are selected among the nouns, proper nouns, and compound nouns by using keyword selection algorithms [19]. In Hangul (Korean language), a syllable including a noun consists of a noun together with noun post-position, or of a noun alone. Therefore, a routine to find nouns require syllable cutting rules.

4.2 Weighting

One commonly used function [1, 9, 10, 20] for assigning weight to main keywords in document or query x is the frequency-modified inverse document frequency, described by

$$w_{x,t} = k \cdot f_{x,t} \cdot \log_2(N/df_t).$$

Where k is a constant for adjusting the relative importance, $f_{x,t}$ is the number of occurrences of main keyword t in document x (the within-document frequency), N is the number of documents in the collection, and df_t is the number of documents containing main keyword t. This is commonly called by the $tf \cdot idf$ weighting. Note that this function includes both global information from the main keyword weight (N/df_t), and local information from $f_{x,t}$.

4.3 Ranking

The first-level retrieval technique used in this paper is the cosine measure, one of the most effective ranking technique [1, 9, 10]. A ranking operation, based on the cosine correlation used to measure the cosine of the angle between vectors, can then be used to compute the similarity between a document and a query, and documents can be ranked based on that similarity.

$$similarity(q, d) = \frac{\sum_t w_{q,t} \cdot w_{d,t}}{\sqrt{\sum_t w_{q,t}^2} \cdot \sqrt{\sum_t w_{d,t}^2}},$$

Where q is a query, d is a document, and $w_{x,t}$ is the weight of main keyword t in the document or the query x. The expression $\sqrt{\sum_t w_{x,t}^2}$ is a measure of the total weight or length of document or query x in the main keywords of weight and number of main keywords in x, and is denoted by w_d for document d, and by w_q for a query q.

In second-level rankings, we reorder rankings by searching retrieved documents using sub-keywords. Of course, we consider the lengths (window sizes) between main keyword and sub-keyword. At this time, if the sub-keyword exist in the retrieved document, the weights of the retrieved documents become doubled.

According to sentence and word units, we assign the weight and reorder rankings of the document.

- Sentence Units
  How many sentences are there taking into account main keywords and sub-keywords?
- Word Units
  How many words are there taking into account main keywords and sub-keywords?

5. Performance Evaluation

5.1 Test Collection (Database)

We used as our test collection a Korean encyclopedia published by the Kemong Company [17]. It is published in 6 volumes, with 500 pages per volume. The size of the text data is around 10 mega-bytes (10MB), and contains 23,113 entries. The content of each entry describes the concept of entry, using other entries or more fundamental words (or concepts). The word length (or document length) of the content to describe an entry varies from 20 words to 1,000 words. The average document length is 56 words. ⟨Table 1⟩ gives the statistics of the test collection (A

Korean encyclopedia).

The following is a tagging of the test collection. ~~〈doc〉 and 〈/doc〉 means a document. The pair, 〈id〉~~ and 〈/id〉, form the identification of the document. The pair, 〈entry〉 and 〈/entry〉, form an entry that is the entry title. 〈content〉 and 〈/content〉 are the content of the entry. The pair, 〈entry_ho〉 and 〈/entry_ho〉, denote a homonym, which means multipleentry meanings. 〈see〉 and 〈/see〉 mean a standard entry used for the entry. The pair, 〈see_also〉 and 〈/see_also〉, form a related entry for the entry.

〈Table 1〉 Test Collection (Database) Statistics

Entries	23,113
Queries	161
Average document length	56 words
Average query length	6 words
Average relevant documents	11

The evaluation of retrieval effectiveness has been resulted 161 natural language queries and relevance information. The natural language queries have used 161 natural language queries that are Type 1, Type 2 and Type 3 natural language queries, excluding Type 4 and Type 5 from collected 260 natural language queries that gathered 10 questions per person from 26 persons who wanted to extract the content from an encyclopedia. Relevance information selected entries per query that agree with at least 2 experts of 4 experts.

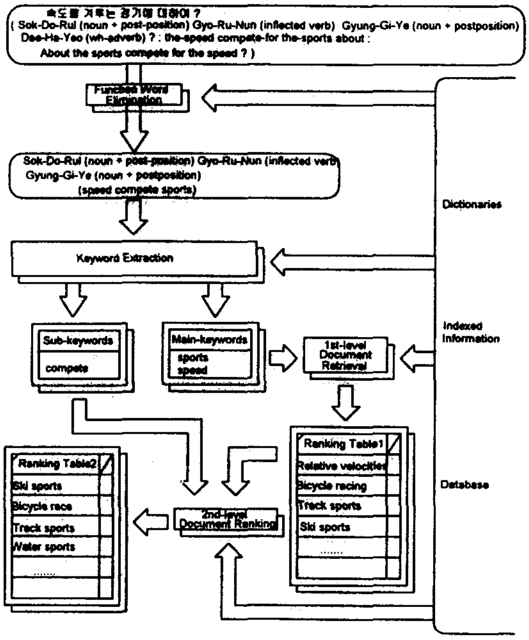
5.2 Evaluation Method

IR system performance is often measured by using "recall" and "precision" values, where recall measures the ability of the system to retrieve useful documents, while precision conversely measures the ability to reject useless materials [1]. The recall is the ability of the system to present all relevant items. The precision is the ability to present only the relevant items. Recall is defined as the proportion of relevant material retrieved, while precision is the proportion of retrieved ~~material that is relevant.~~

A good system is one which exhibits both a high recall and a high precision. The standard recall R and standard precision P may be defined as;

R = number of items retrieved, and relevant / total relevant in collection
P = number of items retrieved, and relevant / total retrieved

(Fig. 3) shows the process of the natural language query, "속도를 겨루는 경기에 대하여?" ("Sok-Do-Gyung-Gi-Ye (noun +post-position) Dae-Ha-Yeo (wh-adverb)?" : the-speed compete-for a-sports about? : About a sport the compete for the speed?), an example of our system. Firstly, we remove the stop word, for example, "Dae-Ha-Yeo (wh-adverb) : about". And then we extract the keywords, using noun and post-position dictionaries. We have a dictionary of around


속도를 겨루는 경기에 대하여 ?
(Sok-Do-Rul (noun + post-position) Gyo-Ru-Nun (inflected verb) Gyung-Gi-Ye (noun + postposition)
Dae-Ha-Yeo (wh-adverb) ? : the-speed compete-for the-sports about :
About the sports compete for the speed ?)
Funcbed Word Elimination
Sok-Do-Rul (noun + post-position) Gyo-Ru-Nun (inflected verb)
Gyung-Gi-Ye (noun + postposition)
(speed compete sports)
Dictionaries
Keyword Extraction
Sub-keywords
compete
Main-keywords
sports
speed
1st-level Document Retrieval
Indexed Information
Ranking Table1
Relative velocities
Bicycle racing
Track sports
Ski sports
....
Database
Ranking Table2
Ski sports
Bicycle race
Track sports
Water sports
....
2nd-level Document Ranking

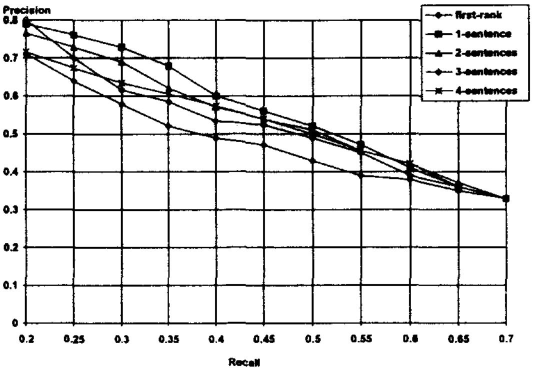
(Fig. 3) An example of natural language retrieval

100,000 nouns and a dictionary of around 3,000 post-positions. Using the main keywords, speed (m-k), sports (m-k), we firstly retrieve the documents. After that, using firstly retrieved documents and the sub-keyword, compete (s-k), reorder the retrieved documents.

We have shown that the 4th ranked "스키 경기" ("Ski sports") entry of the Ranking ⟨Table 1⟩ is top-ranked in the Ranking ⟨Table 2⟩. Additionally, the first-ranked entry, "상대 속도" ("Relative velocities"), of the Ranking ⟨Table 1⟩ disappears in the top 4 ranked table in the Ranking ⟨Table 2⟩.

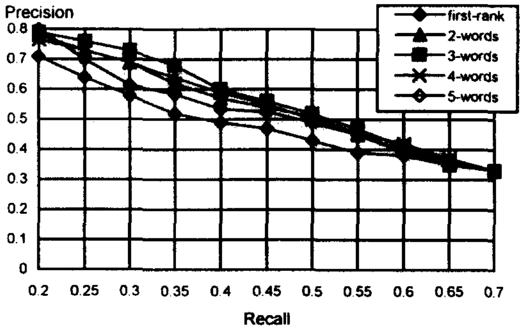
5.3 Evaluation Results

(Fig. 4) shows the results of retrieval effectiveness by sentence units. The first-rank refers to retrieval effectiveness results of first-level rank. The 1-sentence line is the result of retrieval effectiveness in the inner sentences. The 2-sentences line is the result of retrieval effectiveness within 2 sentences. The 3-sentences line is the result of retrieval effectiveness within 3 sentences, and so on. The 1-sentence case represents good retrieval effectiveness.



(Fig. 4) The results of retrieval effectiveness by sentence units

(Fig. 5) shows the results of retrieval effectiveness by word units. The 2-words line is 2, which is the length (window size) of between main keyword and sub-keyword. The 3-words line is 3, and so on. We can see that the 3-words case offers good retrieval effectiveness. The 3-words case has improved the retrieval effectiveness by 8.4%, rather than the retrieval effectiveness of general IR method (first-level ranking).



(Fig. 5) The results of retrieval effectiveness by word units

⟨Table 2⟩ shows the improvement of overall retrieval effectiveness by sentence units. The values are average improvement percentages, which are the sums of the different (improvement) precision values for each recall point divided by the recall point numbers.

⟨Table 2⟩ Improvement of retrieval effectiveness by sentence units(*first-rank* = *0.0%*)

Sentence units	Improvement in retrieval effectiveness
1-sentence	+8.4%
2-sentence	+6.5%
3-sentence	+4.5%
4-sentence	+4.7%

⟨Table 3⟩ shows the improvement of overall retrieval effectiveness by word units.

Furthermore, we show retrieval effectiveness improvement from 11% to 16% from the 0.3 to the 0.4 recall points. We have shown experimental results of context sensitive rankings with two-level ranking method. We have shown through two-level ranking ~~method that the proposed method achieves higher re~~trieval effectiveness than general IR ranking method.

〈Table 3〉 Improvement of retrieval effectiveness by word units(first-rank = 0.0%)

Word units	Improvement in retrieval effectiveness
2-word	+6.6%
3-word	+8.4%
4-word	+6.5%
5-word	+4.4%

First-level rankings (rankings based on main keywords) provide fast retrieval with inverted index files while second-level rankings (searches based on sub-keywords) provide time overheads for reorder of retrieved document with searching words in sentences. But the first-level ranking is provided index size overhead with inverted index file while second-level ranking is provides not increasing the index size for reordering of retrieved documents with search words in sentences.

For our empirical experience, the average number of relevant documents was 11. Because reordering only retrieved documents, we reordered only 30 retrieved documents. We could reorder in real-time, more or less. So we provided retrieval efficiency without hurting response times and increasing the index sizes.

6. Conclusions

An Information Retrieval (IR) is to retrieve relevant information that satisfies user's information needs. However a major role of IR systems is not just the generation of sets of relevant documents, but to help determine which documents are most likely to be relevant to the given requirements.

In this paper, we have defined "Main keywords", "Sub-keywords", and "Two-level rankings" and have ~~presented reordering method (re-ranking) using two-level~~ ranking method in vector space model. In natural language queries, main keywords and sub-keywords have important features. We collected 260 natural language queries from end users. 85% of the natural language queries contained sub-keywords.

In our experiments, we measured retrieval effectiveness by sentence units and word units. An 8.4% improvement in retrieval effectiveness measured in inner sentence and 3-word window sizes, was gained over that offered by general IR method.

In IR systems for large data banks of collected information and stored documents, indexing using main keywords is very simple and clearer. Furthermore, we can reduce the number of keywords and index sizes. The two-level ranking method can lead to retrieval of relative documents using main keywords and sub-keywords more precisely than retrieval using keywords in traditional IR method.

In the future, we will concentrate on more precise relationships between main keywords and sub-keywords. Additionally, we will observe additional weighting methodologies according to the presence or absence of sub-keywords in initially retrieved documents.

References

[1] G. Salton and M. McGill, ed., "*Introduction to Modern Information Retrieval,*" McGraw-Hill Book Company, New-York, 1983.

[2] M. Bartschi, "An Overview of Information Retrieval Subjects," *IEEE Computer*, Vol. 18, No. 5, pp. 67-84, 1985.

[3] C. Cleverdon, "Optimizing Convenient Online Access to Bibliographic Databases," *Information Services and Use*, Vol. 4, No. 1/2, pp. 37-47, 1983.

[4] T. Noreult, "Automatic Ranked Output from Boolean Searches in SIRE," *Journal of the*

American Society for Information Science, Vol. 28, No. 6, pp. 333-339, 1977.

[5] D. Harman, "An Experimental Study of Factors Important in Document Ranking," *ACM SIGIR Conference on Research and Development in Information Retrieval*, Pisa, Italy, pp. 186-193, 1986.

[6] D. Harman and G. Candela, "Retrieving Records from a Gigabyte of Text on a Minicomputer using Statistical Ranking," *Journal of the American Society for Information Science*, Vol. 41, No. 8, pp. 581-589, 1990.

[7] IJ. J. Aalbersberg, "A Document Retrieval Model Based on Term Frequency Ranks," *ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, pp. 163-172, 1994.

[8] J. S. Ro, "An Evaluation of the Applicability of Ranking Algorithms to Improve the Effectiveness of Full-Text Retrieval. II. On the Effectiveness of Ranking Algorithms on Full-Text Retrieval," *Journal of the American Society for Information Science*, Vol. 39, No. 3, pp. 147-160, 1988.

[9] G. Salton, ed., "*Automatic Text Processing: The Transformation, Analysis, and Retrieval of nformation by Computer*," Addison-Wesley Publishing Company, New-York, 1989.

[10] W. Y. P. Wong and D. L. Lee, "Implementation of Partial Document Ranking Using Inverted Files," *Information Processing and Management*, Vol. 29, No. 5, pp. 647-669, 1993.

[11] A. F. Smeaton, "Incorporating syntactic information into a document retrieval strategy: An investigation," *ACM SIGIR Conference on Research and Development in Information Retrieval*, Pisa, Italy, pp. 103-113, 1986.

[12] A. F. Smeaton and C. J. van Rijsbergen, "Experiments on incorporating syntactic processing of user queries into a document retrieval strategy," *ACM SIGIR Conference on Research and Development in Information Retrieval*, Grenoble, France, pp. 31-51, 1988.

[13] P. S. Jacob and L. F. Rau, "Natural Language Techniques for Intelligent Information Retrieval," *ACM SIGIR Conference on Research and Development in Information Retrieval*, Grenoble, France, pp. 85-99, 1988.

[14] G. Salton, C. Buckley and M. Smith, "On the Application of Syntactic Methodlogies in Automatic Text Analysis," *Information Processing and Management*, Vol. 26, No. 1, pp. 73-92, 1990.

[15] K. Sparck Jones and J.I. Tait, "Automatic Search Term Variant Generation," *Journal of Documentation*, Vol. 40, No. 1, pp. 50-66, 1984.

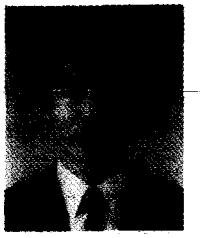
[16] G. Salton and C.S. Yang, "On the Specification of Term Values in Antomatic Indexing," *Journal of Documentation*, Vol. 29, No. 4, pp. 351-372, 1973.

[17] Kemong Company, ed., "*The Kemong Company New Encyclopedia*," Kemongsa Publishing Co., Seoul, 6 volumes, 1992.

[18] W.S. Cooper, "Getting beyond Boole," *Information Processing and Management*, Vol. 24, No. 3, pp. 243-248, 1988.

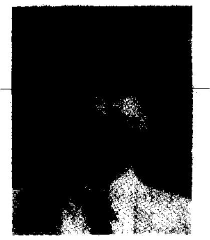
[19] H.K. Kang, C.Y.Lee, H.W. Jang and S.Y.Park, "An Implementation of an Automatic Keyword Extraction System," *The 3rd Pacific Rim International Conference on Artificial Intelligence*, Beijing, China, pp. 708-711, 1994.

[20] K. Sparck Jones, "A Statistical Interpretation of Term Specificity and Its Application in Retrieval," *Journal of Documentation*, Vol. 28, No. 1, pp. 11-21, 1972.



강현규

1985년 홍익대학교 전자계산학과(학사)
1987년 한국과학기술원 전산학과(석사)
1992년 정보처리기술사 자격 취득
1997년 한국과학기술원 전산학과(박사)
1987년~현재 한국전자통신연구원 선임연구원
관심분야:정보 검색, 자연언어 처리, 디지털 라이브러리



박세영

1980년 경북대학교 전자공학과(학사)
1982년 한국과학기술원 전산학과(석사)
1989년 프랑스 파리 7대학(박사)
1995년~1996년 미국 미주리 주립대 객원연구원
1982년~현재 한국전자통신연구원 책임연구원, 자연어처리연구실장
관심분야:자연언어 처리, 정보 검색, 디지털 라이브러리