Collected Papers for 1998 Autumn Academic Conference by Korean Institute of Information Scientists and Engineers, Vol. 25. No. 2

A Method for Improving Recall Precision on Information Retrieval Systems Using Multiple Terms

Jonghee Choi[†], Dongsi Choi[††], Seyoung Park[††], Heekuck Oh[†]
Dept. of Computer Science, Hanyang Univ.[†], ETRI[††]

Summary

Studies on information retrieval systems using multiple terms instead of single terms for precise information queries have been actively carried out. However, there are not many retrieval systems that use multiple terms. One of the examples of information retrieval systems using multiple terms are information retrieval systems using keyfacts. A keyfact is one of the multiple terms that includes not only the key words but also the related information. Information retrieval systems based on keyfacts create keyfacts with the same weighted value in the index process of current documents and the keyfact extraction process of the query language. However, a noun phrase creates different keyfacts according to its meaning, so there are many problems in applying existing information retrieval method to its results. Therefore, in this thesis we suggest a more precise information retrieval method by assigning appropriate weighted value to each keyfact created during the index process.

1. Overview

Most information retrieval systems retrieve information using a keyword, which is a single term. When retrieving information with keywords, the information to be retrieved can be ambiguous. Ambiguity arises when one word has different meanings or has too wide a range of meanings. One method to resolve this ambiguity is to use multiple terms. Multiple terms include the information related to the keyword as well as, unlike a single term, a keyword that has been used by the existing information retrieval system. Since multiple terms include related information as well, we can capture the meaning of the keyword precisely. Related information means information that describes the characteristics of keywords so that we precisely understand the meaning of keywords. So the accompanying noun when consisting of a compound noun, idiomatic language when having idiomatic language, and a verb or an adjective for a sentence acts as related information [1-3]. With these in mind, the concept of a multiple term is a keyfact.

In this thesis, a method to assign weighted value to each keyfact is suggested in order to find more precise information in the information retrieval system using keyfacts. Furthermore, precision when using this method vs. when the equal weighted value was assigned was compared. Chapter 2 contains an explanation of keyfacts, and Chapter 3 covers the method of extracting keyfacts. In Chapter 4, we will explain how we assign weighted value while indexing, and in Chapter 5, we will compare the precision of the retrieved results when assigning different weighted values to each keyfact vs. when assigning the equal weighted value through an experiment. Finally, in Chapter 6, our conclusion and the future research direction will be discussed.

2. Keyfacts

We can increase the search precision if we retrieve with keywords that include related information, not just the keywords when retrieving information. This is because it narrows down the scope of retrieval. Users create query language with multiple terms instead of single terms for more precision.

A word keyfact is originated from the concept that it is not a word but a fact that represents the document and the keyfact should have related information with the keyword. A keyfact consists of a central word and a subordinate word, which means the keyword is a central word and the related information is the subordinate word. There are

different ways to express things in a sentence but if it has the same meaning, it becomes the same keyfacts. So a keyfact can be the same in terms of meaning but it can be different grammatically, because there can be different ways to express one keyfact. Keywords can be extracted from a document with the existing method, and then you can infer the original document only using the keywords. In addition, noun phrases can be extracted from a document and you can infer the original document only using the noun phrases. It has been proven that the latter better expresses the original document [4].

As an index term, it should represent the document first, and then there is the possibility of it showing up again. Usual noun phrases are representative to some degree, but there is almost no possibility of them showing up again. So a noun phrase should be created with different keyfacts in a keyfact based information retrieval system.

3. How to extract keyfacts

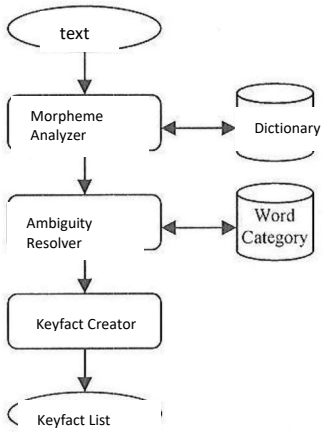

Figure 1 Keyfact Extractor Diagram

To extract keyfacts, we should go through the three step process. Firstly, the given sentence or word phrase should be analysed into morphemes and then secondly, in the analysed morphemes, ambiguities should be resolved. When resolving ambiguities, relevancy with other morphemes included in the same sentence or word phrase is compared. This can be done easily by using relevant nouns. However, there are difficulties in completely resolving the ambiguity. So, resolving ambiguity in this step is applied only to the very simple patterns and others are determined depending on its frequency in a corpus. Finally, keyfacts are extracted according to the keyfact generating rule with the morphemes after resolving ambiguity.

A morpheme, after going through the morpheme analyser and ambiguity resolver, has one part of speech and one meaning. Keyfacts may be extracted as follows [5].

- It may be a keyfact only with a central word. In other words, one noun (existing keyword) is used as a keyfact.
- When two keywords are connected with 'of' two keywords can form central words and the keyword after 'of' can be a subordinate word.
- When two keywords are connected with '와/과(and)' two keywords can form central words and the keyword after '와/과(and)' can be a subordinate word. In this case, it doesn't matter even if the position of the central word and the subordinate word can be swapped with each other.
- Derivative determiner, descriptive verb and non-descriptive verb can be used only as a subordinate word, a relative verb.
- When two keywords connected without proposition form one keyword, there is a sequence.

When creating a keyfact, it is not about making one noun phrase into one keyfact. Multiple keyfacts are created from one noun phrase. This is because it is possible to express a noun phrase as one keyfact but there is a problem in the partial matching with keyfacts generated by other patterns. Of these keyfacts generated in this way, there are not only the ones with both a central word and a subordinate word but also ones only with a central word. More precise searches will not be possible if the equal weighted value is assigned to those with both a central word and a subordinate word and those with only a central word when searching with these keyfacts.

4. Indexing process

This thesis suggests a different method from other existing information retrieval systems in regards to the fact that each keyfact has its unique weighted value during the indexing process. Since multiple keyfacts are generated from one noun phrase, there are problems in applying the equal weighted values to all keyfacts. For example, let's say

there is a noun phrase, “Essence of Appreciation.” Then in this word phrase the keyfacts below are created.

[Appreciation, NIL], [Essence, NIL], [Appreciation, Essence], [Appreciation, Essence, NIL]

In this case, [Appreciation, Essence] has more precise meaning than [Appreciation, NIL] and [Essence, NIL]. Thus, [Appreciation, Essence] should have more weighted value than [Appreciation, NIL] and [Essence, NIL].

Another example is a noun phrase, “Order of God and Nature.” Below The keyfacts below are extracted.

[God, NIL], [Nature, NIL], [Order, NIL], [God, Order], [Nature, Order], [God Nature,], [Nature God,], [God Nature, Order], [Nature God, Order]

In this case as well, the weighted value of the keyfacts with both a central word and a subordinate word should be higher than the one only with a central word.

Also, as you can see above, since each keyword “Appreciation,” “Essence,” “God,” “Nature,” and “Order” appears once in the body of the document, the sum of weighted values on each noun within the keyfacts generated from these keywords should be the same. In other words, of the keyfacts extracted from the noun phrase “Essence of Appreciation,” the “Appreciation” appeared in three keyfacts out of a total of four keyfacts. Of the keyfacts extracted from noun phrase “Order of God and Nature,” the “God” appeared in six keyfacts out of a total of nine keyfacts. So each keyfact with “Appreciation” should be assigned a weighted value of 1/3 and each keyfact with “God” should be assigned a weighted value of 1/6. However, more experiments and considerations are needed to calculate more precise values.

5. Search and experiment

All keyfacts are either [central word, subordinate word] or [central word, NIL]. [Central word, subordinate word] is more narrowed down, having a more precise meaning than [central word, NIL], thus being more helpful to retrieve appropriate documents and information. In this experiment, we targeted 23,112 documents from an encyclopaedia by Kyemong Co. and the total volume of data was approximately 12Mbytes. This experiment was conducted according to two cases: all keyfacts having equal weighted value and each keyfact having different weighted value. The formula below was used in the latter.

$$Weight = \frac{k}{N} \times p$$

Formula 1

In formula 1, N is a total number of keyfacts generated from one noun phrase. k is the number of keyfacts including the particular words. p is a correlation coefficient. In this thesis, 1 was used when both a central word and a subordinate word exist and 0.5 when only a central word exists.

In these two cases, the retrieved document actually matches. When the same queries are raised the same document is retrieved but its ranking in the two cases are different from each other. A vector space model was used as a rank assignment algorithm. Also, to measure the precision, answers to the queries were defined in advance. Precision represents how much the retrieved result matches with the pre-defined answers to the queries. However, the existing concept of precision was slightly extended. Precision was determined by emphasizing the ranking of the retrieved results. The number of appropriate documents included in the top 15 of the retrieved documents was compared. For instance, let’s say there are 10 documents with pre-defined answers. If 10 documents are included in the top 15 ranking of the retrieved documents, the precision is 100%. If 5 documents are included it becomes 50% and if there is no document, it is 0%.

Below is the comparison of the answers to the queries.

1) “What is the origin of Chuseok?”

	Keyword	Keyfacts with equal weighted value	Keyfacts with different weighted values
Retrieved document	466	314	314
Precision	75%	75%	75%

2) “What are Jang Yeong-sil’s achievements?”

	Keyword	Keyfacts with equal weighted value	Keyfacts with different weighted values

Retrieved document	352	258	258
Precision	77%	88%	88%

3) "What is the difference between ultraviolet rays and infrared rays?"

	Keyword	Keyfacts with equal weighted value	Keyfacts with different weighted values
Retrieved document	1075	738	738
Precision	30%	40%	60%

In the first example, the same precision was achieved for the three cases. For the second case, the precision result retrieved from keyfacts with equal weighted value and the result retrieved from keyfacts with different weighted values is the same. But this precision is the result of a simple investigation to see if it is included in the top 15. If we look into each ranking of the retrieved result, the ranking of the result retrieved from the keyfacts with different weighted values was the closest to the ranking of the pre-defined answers to the queries. In addition, the average precision in 40 pre-defined answers to the queries is as follows:

	Keyword	Keyfacts with equal weighted value	Keyfacts with different weighted values
Precision	69%	74%	78%

6. Conclusion and future research direction

When retrieving information using keyfacts – multi terms - more precise results were achieved by applying different weighted values to each keyfact, rather than applying equal weighted value to all the keyfacts. The difference in precision was not significant in the two cases but when it comes to the ranking of the retrieved result we could retrieve more appropriate documents when using the keyfacts with different weighted values applied. In the future research, a method or an algorithm to assign different weighted values suited for each case should be developed, rather than considering the weighted value as two cases only.

References

[1] C.J. van Rijsbergen, *Information Retireval*, Butterworths, London, Second Edition, 1979.

[2] "ETRI-NLPS natural language process form tag set for meaning based information retrieval" by Kyungtak Jung, Dongsi Choi, Miseon Jeon, Raewon Seo and Seyoung Park, Natural Language Processing Section, The Electronics and Telecommunications Research Institute, 1997

[3] G. Salton and M. G. McGill. *Introduction to Modern Information Retrieval*, McGrawHill, New York, 1983.

[4] "Contents based multimedia information retrieval technology development" The Electronics and Telecommunications Research Institute

[5] "Keyfact concept based information retrieval system" by Daisuk Jang, Department of Computer Science, Hanyang University, thesis for master's degree, 1997

CERTIFICATE OF TRANSLATION

TRANSPERFECT LEGAL SOLUTIONS

I, Yoon Hee Choi, am a professional Korean to English translator based in Texas. I am competent to translate from Korean to English, and I have 11 years of experience doing so. I hereby certify that the attached English document **A Method for Improving Recall Precision on Information Retrieval Systems Using Multiple Terms** is an accurate translation of the attached Korean document **다중단어를 사용한 정보검색 시스템에서의 재현정확도 향상방법** to the best of my knowledge and belief.

I declare that all statements made herein of my knowledge are true, and that all statements made on information and belief are believed to be true, and that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code.

Signed: 

_____

Name: Yoon Hee Choi

Date: 5 August 2019

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.