An improvement of information retrieval system using word location information

KwangYoung Kim, DuSeok Jin, YoonSoo Choi, JinSuk Kim, Jerry Seo, YoungKyon Suh
Korea Institute of Science and Technology Information
Group for Intelligent Information system
{kykim, dsjin, armian, jinsuk, jerryseo, yksuh}@kisti.re.kr

Abstract

Today it's difficult to effective searching the internet because it's a huge collection of documents. Currently most people search documents by just typing a few keywords. But this is very import keywords for searching documents.

This paper suggests that the word location information is an import element for searching relevant documents. This paper suggests that using word location information of a user query can exactly improve the result of information retrieval system.

This paper makes directly experiments on retrieval system with various weight methods of word location information.

1. Introduction

With the rapid expansion and popularity of the World Wide Web and the internet, it has made huge index systems like Google, Yahoo and so on. These index systems are very good. But most people have a feeling that it is hard to find relevant documents.

Generally most commercial information retrieval engines adopt Boolean query as a query type. A user query like Boolean type is useful to information retrieval engines [1].

Today most information retrieval systems based on a user query like Boolean type.

A user query is only information which a user wants to find documents. So a user query is the most important element to find relevant documents. Most information retrieval systems use to judge whether this document is a relevant match through a user query.

We must consider the best way to find the best match document through a user query. This paper suggests the word location information of a user query is a very import element to find relevant documents. The weight of word location information is a very import element to decide relevant documents from searching result.

This paper makes experiments how word location information affects the searching result. This paper will makes up the indexing files of word location information for this experiment and makes experiments with various weight methods of word location information.

2. Structure indexing file

Generally to find relevant documents have importantly used the document frequency of indexing word importantly because it affects search result pages. The document frequency of information retrieval system is generally used to judge whether this document is a relevant match or not. This paper has posting files[1] of the document frequency information.

Location information of indexing word importantly improved the precision of searching result with document frequency [2]. This paper additionally consists of indexing files with word location information and makes new experiments with these.

New indexing files consist of the posting and the word location information files like Figure 1. The indexing files of posting information have information of document frequency. The indexing files of word location information have real word location information. The indexing files of word location information specially consist of the page numbers and the word numbers like figure 1.

[1] index files

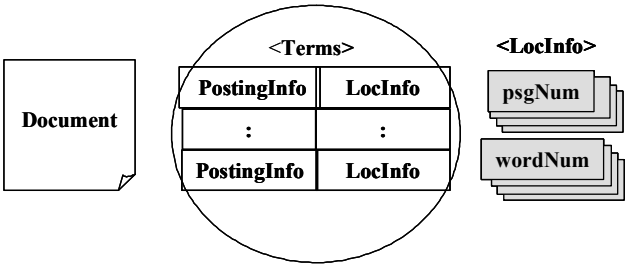

Fig 1 structure the indexing files

The page numbers have information of the phrase distinction. The word numbers importantly have real word position information list. Using the value of word number this paper decides whether two terms are close or not like Figure2. This paper easily decides whether two terms are close or not through comparing the distance of two terms.

A general vector weight means the vector space model of information retrieval system. The distance weight of word means that is difference between two terms. The closer terms are the more weights have.

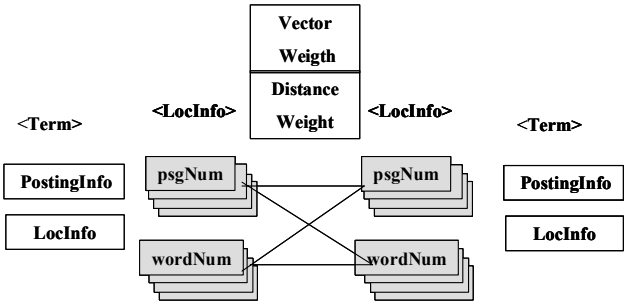

Fig 2 structure word location information

This paper evaluates a general vector weight including the distance weight of word. This paper makes experiments with various methods of word location weight.

This paper will decide whether various methods have good results or not through experiments.

3. The distance weight of word

This paper makes experiments with four methods for like table 1. Because of using the vector model is comparing only vector weight with distance weight of word in this paper.

Table 1 the distance weight of word

Expression I : $Similarity(doc_i) = Weight_{Vector}(doc_i) + \boldsymbol{\alpha}$

Expression II : $Similarity(doc_i) = Weight_{Vector}(doc_i) + \boldsymbol{\alpha/distance}$

Expression III : $Similarity(doc_i) = Weight_{Vector}(doc_i) + \boldsymbol{1.0/(l.0+\log(distance))}$

Expression IV : Expression3 : $Similarity(doc_i) = Weight_{Vector}(doc_i) + \boldsymbol{1.0/\sqrt{(distance)}}$

$Weight_{Vector}(doc_i) = \log(TF^2+1.0)*IDF^3/\log(STF^4+1.0)$
$Distance = Minimum(Term_{i+1} - Term_i)$

The $Weight_{Vector}$ calculates Vector weight of a $document_i$

The distance is chosen the minimum value between two terms.

Expression I means that this paper will have a constant alpha value. It does not matter whether two terms are close or not. This paper will just add alpha value without two terms distance.

For example, there are two documents like table 2.

Table 2 Example

A. "**Information system** is … "
B. "**Information** retrieval has … using the meta **system** …"

When a user finds documents in table 2 with the query of 'information system', two documents have the same weight from Expression I.

The main purpose of using word location information will try to assign more weight of a document A than a document B like table 2.

Expression II means that the closer word location is the more weight has according to a distance value.

Expression III has the same meaning of Expression II but the alpha value will be increased by a log value.

[2] Term Frequency
[3] Invert Document Frequency
[4] Sum Term Frequency

Expression IV has the same meaning of Expression II but the alpha value will be increased by a square root value.

This paper will try to get the highest average precision from the changing of the alpha value. The highest average precision is the best method for searching result.

4. Experiment

This paper makes experiments on KRISTSAL[5] system with Expressions of the tables 1. KRISTAL system is information retrieval management system developing by GIIS [6] in KISTI. (Korea Institute of Science and Technology Information)

This paper used the testing collection of HANTEC[7] (like TREC) version 2.0 for evaluation. The total document of HANTEC is 120,000 documents. The testing collection used 50 queries for evaluation like TREC. This paper used the relevant file of L2.rel to get average precision evaluated [3, 4].

This paper was evaluated average precision by using HANTEC Version 2.0 and using 50 queries.

The results of experiments are like table 3. The value of alpha actually used between 0 and 1.0.

Table 3 the result of the experiment

alpha	0	0.1	0.2	0.3	0.4	0.5	1
Ex1	0.1915	0.1988	0.1908	0.1838	0.1779	0.1738	0.1699
Ex 2	0.1915	0.2005	0.1906	0.1737	0.1642	0.1565	0.1349
Ex3	0.1915	0.2018	0.1935	0.1851	0.1698	0.1639	0.1431
Ex4	0.1915	**0.202**	0.1946	0.1907	0.1775	0.1671	0.1462

The average precision of Expression I rapidly was downs when alpha values was increased like figure 3.

The expression I maximum of average precision is 0.198% like table 3.

The word location information weight affected the improvement of information retrieval system. Using too much alpha value will decrease the improvement of system like table 3 and figure 3.

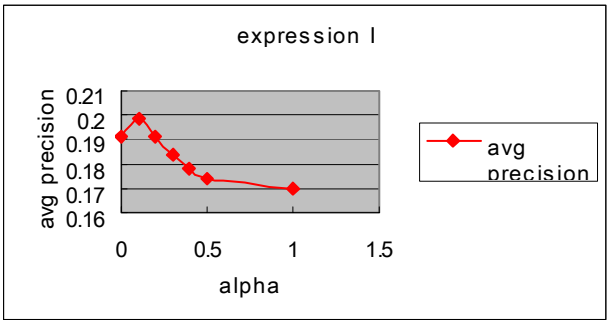

Fig 3 Expression I

Expression II was evaluated like figure 4. The maximum of average precision is 0.2005%. It was too decreased from more than the point of 0.1.

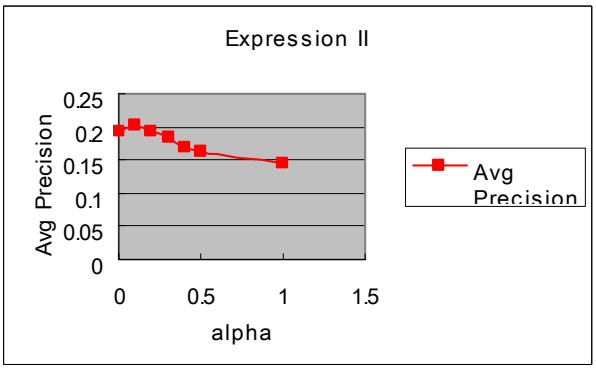

Fig 4 Expression II

Expression III was evaluated like figure 5. The maximum of average precision is 0.2018%. It was too decreased from more than the point of 0.1.

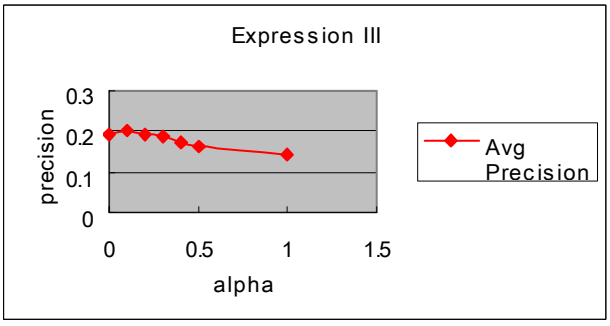

Fig 5 Expression III

[5] IRMS(Information Retrieval Management System)
[6] group for intelligent information system
[7] Hangul Test Collection

The expression IV was evaluated like figure 6. The maximum of average precision is 0.202%. It was too decreased from more than the point of 0.1.

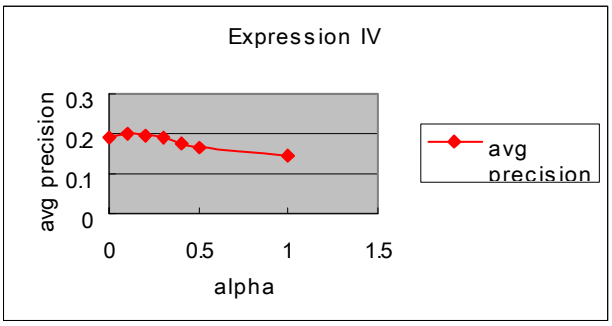

Fig 6 Expression IV

The best result of experiments is Expression IV like tables 2 and figure 6. The highest average precision is 0.202% like tables 2 and figure 6. Average precision 0.0105% was increased through experiments.

5. Summary

This paper knew that the word location information affected to find the relevant document match. The average precision 0.0105% was increased through the expression IV result of an experiment.

Using word location information if two terms very are close, this paper tries to move up to top-ranking documents like table 2. So the relevant documents were located in the top-ranking from the searching result. When we watched directly the result of the relevant document, the searching result was good.

This paper's main issue is that we don't try to find the top 10 relevant documents but try to move up to first top-ranking documents in the top 10 relevant documents.

Therefore, word location information makes a decision that 9th relevant document could move up to a first relevant document.

6. References

[1] Hyun-Young Lee, "**Intelligent Information Retrieval Using Interactive Query Processing Agent**", Journal of the Korea Computer Industry Education Society), 1229-9650, 4(12), pp.901-910, 2003

[2] Dae-Won Park, "**Phrase search using posting file in Korean Information Retrieval System"**, Proceedings of the Korean Information Science Society Conference, Proceedings of The 27th KISS Spring Conference), pp.384-386, 2002.

[3] Jun-Ho Lee, "**Developing the KRIST test collection for researches in information retrieval"** Journal of the Korean Society for information Management, 1013-0799, 12(2), pp.225-232, 1995

[4] Sung-Hyon Myaeng, Jun-Ho Lee "**Construction of a Balanced Test Collection for Evaluation of Information Retrieval System"** Journal of the Korean Society for information Management, 1013-0799, 16(2), pp.135-148,1999