An innovative Term Weighting method for Indexing in Elasticsearch Document Retrieval

Gayatri S. Kapadia[1], Rustom D. Morena[2]

[1]Assistant Professor, Sarvajanik College of Engineering & Technology, Surat, Gujarat, India
[2]Professor, Department of Computer Science, Veer Narmad South Gujarat University, Surat, Gujarat, India
(E-mail: [1]gayatriskapadia@gmail.com, [2]rdmorena@rediffmail.com)

***Abstract*—**In document retrieval, term weighting and indexing are areas of research focus since long. For document retrieval, most of the traditional IR systems use Term Frequency (TF) x Inverted Document Frequency (IDF) term weighting method. However, to retrieve most relevant documents from a collection of document sets for given query, we propose a new term weighting for indexing based on TF, IDF and Relevance Frequency (RF). We have implemented this in open source search engine Elasticsearch. Our experimental results using standard data sets show that TF x IDF x RF improves the relevant document retrieval as compared to conventional term weighting methods.

Keywords*—Relevance Frequency (RF), TF x IDF, Elasticsearch, Indexing, Document Retrieval.*

I. INTRODUCTION

The Information Retrieval (IR) has very broad meaning. IR is the process of finding the text documents that satisfies the user's query. It is a science of retrieving relevant information which fulfils user's demand. The size of electronic information grows rapidly, and the information retrieval system is required to retrieve the needful data from this huge collection. The user query is generally unstructured and it contains a number of words for retrieving information relevant to it. The document retrieval is the part of information retrieval which can be defined as matching user query with available set of text records.

One of the major problems is to find the relevant documents for the given user's query. In IR, the main challenge is deriving the relevance of a document as accurately as possible from the set of documents to a given query. Further challenge is to determine a relative measure between documents and the query so that retrieved documents can be accessible to the user. The IR system returns a ranked list (of retrieved documents) which is an assortment of the values of relevant measure in descending order. In short, the IR System's main issue is to find the relevant documents and rank them. According to Zhai et al. [1] finding the important keywords and assigning proper weight to term are two major issues in Information Retrieval. The other issues associated with the retrieval of the documents for the given query are incomplete query and/or specification and vocabulary mismatch.

Our focus is therefore to address this issue and to improve the accuracy and efficiency in searching the terms in textual information, particularly how close the terms appear in a document and provides a good degree of relevance with accuracy and efficiency.

Most of the systems use term weight that include Term Frequency (TF), Inverted Document Frequency (IDF). Richard McCreadie et. al. [2] mentions that term weight measure would help the real-time news vertical search in an efficient manner. Term Weight is an important predictor which is used to estimate the importance of the keywords in a text while retrieving the documents. The term weight (TF x IDF) for each term in each document through the inverse proportion of the frequency of the term in the set of documents is calculated. If the term weight of a particular term is high that means it has a strong correlation with the document and thus, if that term was to appear in a query means, the document could be of interest to the user. This entire process is known as indexing.

For document retrieval, most of the traditional IR systems use Term Frequency (TF) x Inverted Document Frequency (IDF) term weighting method or other methods based on it. However, to retrieve most relevant documents from a collection of document sets for given query, this paper presents a new term weighting method for indexing based on TF, IDF and Relevance Frequency (RF).

The paper is organized as follows:

Section II Discusses term weight parameters for Indexing. Section III contains the study of literature and finds the gap in existing systems. Section IV explains the methodology used for our work. Section V demonstrates the experimental results and comparison with existing work. Section 6 concludes the paper.

II. DOCUMENT RETRIEVAL AND INDEXING:

Simple diagram showing the process of document retrieval is shown in figure 2.1.

In order to understand Indexing, the important terms called parameters for Term Weight are briefed here:

A. Term weight parameters for indexing:

- Term Frequency:

  Term Frequency refers to a number of occurrences of a word in the given document.

  TF(t) = (Number of times term appears in a document) / (Total number of terms in the document).

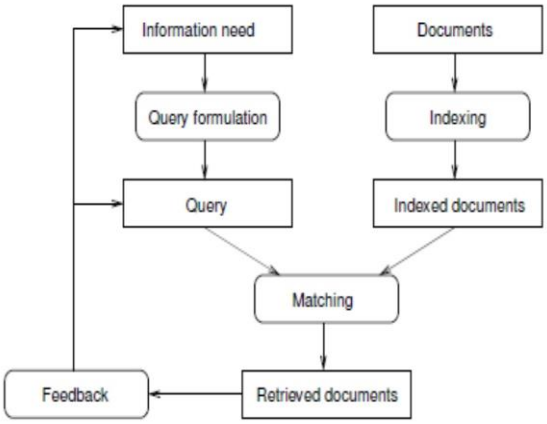

Fig. 2.1 {The process of Document Retrieval}

- *Inverted Document Frequency:*

  The Inverted Document Frequency is a statistical weight used for measuring the importance of a term in a text document collection. The document frequency (DF) of a term is defined by the number of documents in which the term appears.

  IDF(t) = log (Total number of documents / Number of documents with term t in it)

- *Relevance Frequency:*

  The relevance frequency, is the proportion of relevant documents in which a term occurs divided by the proportion of non-relevant items in which the term occurs.

  RF = log (1 + (Number of Relevant Documents/ Number of Non-Relevant Documents))

  There are two other parameters of an IR System, Precision which is used to retrieved only relevant information from given documents set and another is Recall which is used to retrieved documents relevant to given query.

B. Information Retrieval Models:

The commercial search engine provides a new set of challenges to Information Retrieval Researchers. To achieve consistent and effective retrieval results on large collection, the current state-of-the-art academic retrieval models are not robust enough. An interesting characteristic of the retrieval models used in information retrieval is that they demonstrate statistical properties of text rather than the linguistic structure. It means the ranking algorithms are more focused on the counts of word occurrences rather than the word's part of speech, i.e., the word is a noun or an adjective. The linguistic features are incorporated by more advanced models but those are of secondary importance. The main objective of IR model is to achieve effective retrieval results from a huge document sets. There are three classical models of Information Retrieval: Boolean, Vector Space, and Probabilistic. We have used Vector Space Model for our method.

C. Search Engine:

Elasticsearch is a search engine based on Lucene Technology which is an open source and real time search engine. In Elasticsearch, Index is a lightweight data organization system. Elasticsearch provides Index API and/or Indices API which not only adds but also updates JSON document on a specific index for making it easily searchable. Indexing shards, indexing recovery, clustering, and indexing replica are main characteristics of Elasticsearch. We have chosen Elasticsearch for indexing text data. Indexing in Elasticsearch includes creating a dictionary and building indexes.

III. RELATED WORK

In modern Information Retrieval System for document retrieval, a Term Weight is a significant characteristic as per Chirs Buckley [3]. Terms are composition of words, phrases which are used to identify the contents of a text. Gerard Salton et. al. [4] invented term weight which is an important indicator associated with every term.

Term Frequency (TF), Inverted Document Frequency (IDF) and Document Length (DL) are three major components which affect the importance of term in text. [5] TF is frequently visited where as IDF and DL not. To study DL in detail, Amit Singhal et. al. has developed tools which isolate document length and applied document length normalization functions for document ranking. [6] However, what we observed is that, this document length isolation is not very effective because chances of ignoring short length documents are higher. Jiaul H. Paik [19] proposed a novel approach on Term Frequency (TF). Two components were considered, where one component was based on Term Frequency for long documents and other was based on short documents. For these components, it used query length normalization. Chandra Shekhar Jangid et. al. suggested that the conventional method TF x IDF gives highest precision value for news corpus [20].

A global term weighting scheme which can handle any length terms, called N-grams, proposed by Masumi Shirakawa et. al. [7]. This N-grams is based on the information distance and compares the weight of words and phrases. Saptaditya Maiti et. al. [8] proposed a term weighting measure, in which the factor Document Weight (DW) is introduced along with classic term weighting measure TF x IDF i.e., they proposed TF x IDF X DW. DW is Document Length Normalization Component which includes the overall content of the document, its length. DW is measured using Shannon's entropy as mentioned by C. E. Shannon [10] and Karmeshu [9].

K. Sparck Jones et. al. first proposed a statistical measure of determining the term significance. [11] This statistical method attempts to summarize the specific term or general term in particular document by interpreting how many documents are relevant to a given term.

From the paper of Xin Fu et. al. [12] and Man Lan, Chew-Lim Tan et. al. [13] we can say that the IDF factor is not sufficient to improve the term weight or importance of the term. TF x IDF works well when there is no class information,

i.e., text classification. Hee-soo Kim, et. al. [14] and Youngjoong Ko et. al. [15] tried to use class information for calculating term weight measures like Relevance Frequency and Delta TF x IDF. However, it is extremely difficult to represent a collection of documents using class information. For text categorization, relevance frequency improves term discriminating power. [18]

K.L. KWOK [16] proposed a measure called Inverse Collection Term Frequency (ICTF), which has been correlated with IDF factor for ad-hoc test documents retrieval. In other words, it is an alternative of IDF which counts the occurrences of an individual term rather than absent or present terms of documents in IDF. Wu, H. C. at. el. [17] identified another major difference between ICTF and IDF was that ICTF calculates new terms addition based on per-occurrence, whereas IDF calculates new terms addition based on per-document.

There is a method named Proportioning Documents Word Embedding (PDWE) proposed by Jiahua Du et. al. [21]. Here, the proportion of documents analyze the categorical parts of documents. But it is designed without tuning and domain knowledge. Moreover, it is not implemented on different size of the corpus, types, and the datasets are unbalanced or unstructured.

Madhu Kumari et. al. [22] proposed Synonym-Based Term Weighting Scheme (SBT) to make the keyword extraction more efficacious by conquering the limitations of TF x IDF weighting measure. The limitations are – while calculating TF x IDF, synonyms are not considered and counting the different but similar words. This SBT measure helps better keywords' extraction from document for indexing, document classification and document clustering. Khoo Khyou Bun et. al. [23] proposed TF x PDF (Term Frequency x Proportional Document Frequency) algorithm for recognizing the terms which are relevant to main topics. This algorithm assigns the heavy term weight to terms which frequently appear in many newswire documents. Researchers have proved that this algorithm works fine with news archive, but it requires sentence vector clustering prior to calculation of term weight.

The Relevance is an important concept of Information Retrieval (IR) as per Fabio Crestani et. al. [24]. There have been many efforts towards defining the term "Relevance", and no unique precise definition is available. However, we can define relevance as a relationship between document and user's query.

The Relevance Frequency was proposed by Youngjoong Ko [25] and M. Kan et. al. [26] for using the ratio of term occurrences of the negative class and the positive class to calculate term weights. But here, the authors did not discuss how they make the test documents' representation.

Man LAN et. al. [27] proposed rf measure for improving the discrimination power of terms. This measure is significant and consistent for both supervised and unsupervised terms in text classification. The rf measure integrated with other parameters will be helpful into various text mining operations such as text summarization, information retrieval etc.

Various parameters as mentioned in Literature Review have both advantages and disadvantages. most retrieval methods presume a "bag of terms" representations for both queries and documents. Term Frequency (TF) is intended to retrieve higher score to the document which has more query term occurrences, while Inverse Document Frequency (IDF) penalizes words which are more popular in the whole document collection. Document Length Normalization (DL) avoids long documents which may contain more chances for matching query terms. Calculating and Normalizing Document Weight (DW) and/or Document Length requires more time than actual indexing task. An Inverse Collection Term Frequency (ICTF) is better compare to IDF because whenever new term is added into an existing document, it counts this term individually, whereas in case of IDF new term is counted per document. The calculation of Proportional Document Frequency (PDF) requires sentence vector clustering before actual term weight calculation, which is time consuming. The calculation of Proportioning Documents Word Embedding (PDWE), does not consider data types, different sizes plus the documents are not formatted. So, the structuring of documents need to be done which takes more time compare to indexing the documents. In Synonym Based Term (SBT) weighting, the synonyms are not anticipated means all the similar kind of words are counted.

There are two major issues with term weight: first which documents are to be considered? and second how to distinguish the important terms from the less important ones? Content representation and term weight are respective answers to these questions.

There are various term weighting measures for indexing, and their limitations as mentioned above. To overcome those limitations, we have developed a new term weight measure which can provide the accurate results for the given query.

IV. METHODOLOGY

We have retrieved relevant documents as per given query from TREC dataset. Before applying indexing, we analyzed: Original Data, Required Information, Original File formats (which is in Standard Generalized Markup Language (SGML) format) of both query and document (shown in Figure 4.1 and 4.2 respectively). Thereafter, preprocessing was done and term weighting parameters were applied.

The preprocessing steps included extraction of data i.e., data transformation (as shown in Figure 4.3 and 4.4), tokenizing, removing stop words, building inverted index.

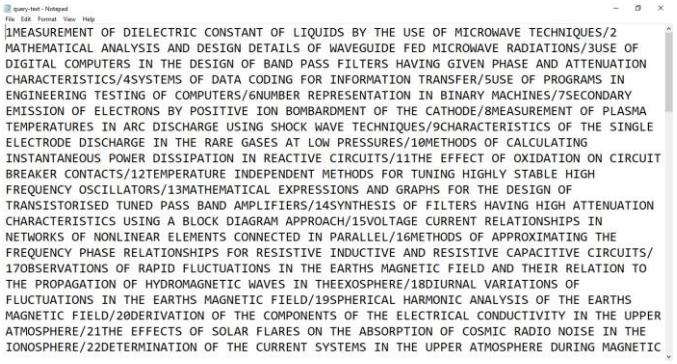

Fig. 4.1 {Query file in SGML format}

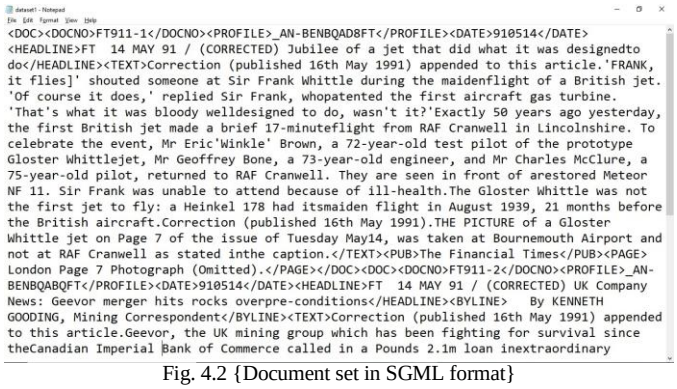

Fig. 4.2 {Document set in SGML format}

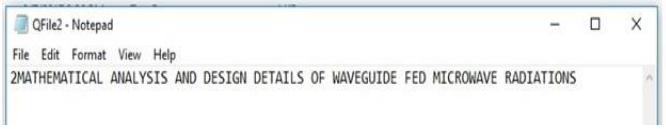

Fig. 4.3 {Query file format after splitting}

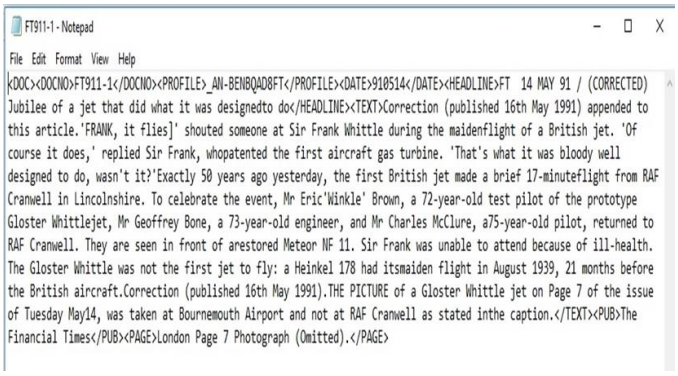

Fig. 4.4 {Document file format after splitting}

The steps of work flow model including preprocessing are as below:

Step 1. Analysis of query set and document set.

Step 2. Split the query set and document set into text files.

Step 3. Convert the text documents into JSON files.

Step 4. Calculate Term Frequency (TF) for each word in query file and document file.

Step 5. Calculate Inverted Document Frequency (IDF) for each word in both query file and document file.

Step 6. Calculate Relevance Frequency (RF) for each word in query file and document file.

Step 7. Find the term weight of each word in both query and document file.

Step 8. Evaluate term measures Precision and Recall Scores.

Step 9. Apply Sorting on Score.

V. EXPERIMENTAL RESULT

There are numerous types of datasets available electronically, and selecting an authentic one is important. We used TREC datasets. The U.S. NIST (National Institute of Standards and Technology (NIST)) has sprinted a very large IR test collection series since 1992. Within this context, there have been a wide range of tracks over a range of different test collection. The TREC Ad Hoc track test collections are the best which were used during the first eight TREC evaluations between 1992 and 1999.

We have compared our method TF x IDF x RF with other researchers in following table 5.1.

TABLE 5.1 COMPARISON OF EXPERIMENTAL RESULTS

Model	Precision	Recall	Accuracy
TF x IDF [11]	85.7%	58.20%	81%
TF x IDF x DP (Discrimination Power) [28]	90%	64%	87%
TF x IDF x RF	87%	57.82%	82.3%

The results in above table 5.1 shows that the accuracy of our method is better than the accuracy of method that uses TF x IDF [11] parameters.

The accuracy of our method is less than the accuracy of method that uses TF x IDF x DP [28]. However, we consider using RF parameter better than using DP because DP needs well defined terms of query which puts restriction in query forming. For all users, query formation with well-defined terms is difficult task.

VI. CONCLUSION

Most of researchers worked on term weight parameters which includes only TF, some of them worked on term weight parameters which includes both TF and IDF. But none of them have used TF, IDF and RF together.

Elasticsearch is an efficient, popular and open-source search engine, where TF x IDF are considered for indexing. We used TF x IDF x RF for indexing in order to retrieve the relevant documents in Elasticsearch for finding relevant documents with better accuracy.

Comparing our results with the results obtained by other researchers, it is found that the accuracy of our method is better than the accuracy of method that uses TF x IDF parameters. Whereas, the accuracy of our method is less than the accuracy of method that uses TF x IDF x DP. However, we consider using RF parameter better than using DP because DP needs well-defined terms of query which puts restriction in query forming. For all users, query formation with well-defined terms is difficult task.

Therefore, it can be concluded that indexing with TF x IDF x RF is better term weight for efficiently retrieving relevant documents.

REFERENCES

[1] Zhai, C. X. and J. Lafferty, "A study of smoothing methods for language models applied to ad hoc Information Retrieval". In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 334-342, September 9-12 2001.

[2] Richard McCreadie, Craig Macdonald, and Iadh Ounis, "News Vertical Search: When and What to Display to Users", SIGIR'13, July 28–August 1, 2013. Dublin, Ireland. Copyright 2013 ACM 978-1-4503-2034-4/13/07.

[3] Chris Buckley, "The importance of proper weighting methods". In M. Bates, editor. In Proceeding HLT '93 Proceedings of the workshop on Human Language Technology, pp. 349-352, 1993.

[4] Gerard Salton, Chris Buckley, "Term-weighting approaches in automatic text retrieval". Information Processing and Management, 24(5): 513-523, 1988.

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

[6] Amit Singhal, Chris Buckley, and Mandar Mitra, "Pivoted Document Length Normalization." In Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval pp. 21-29, August 1996.

[7] Masumi Shirakawa, Takahiro Hara, Shojiro Nishio, "N-gram IDF: A Global Term Weighting Scheme Based on Information Distance", WWW 2015, May 18–22, 2015, Florence, Italy. ACM 978-1-4503-3469-3/15/05. http://dx.doi.org/10.1145/2736277.2741628. pp. 960-970.

[8] Saptaditya Maiti, Deba P. Mandal, Pabitra Mitra, "Tackling Content Spamming with a Term Weighting Scheme", J-MOCR-AND '11 Beijing, China, ACM 978-1-4503-0685-0/11/09, pp. 1-5, 2011.

[9] Karmeshu, editor. "Entropy Measures, Maximum Entropy Principle, and Emerging Applications". Springer-Verlag, Berlin, 2003. Doi: 10.1007/978-3-540-36212-8

[10] C. E. Shannon. "A mathematical theory of communication". Bell System Technical Journal, Vol. 27. pp. 379–423, 1948.

[11] K. Sparck Jones, "A Statistical Interpretation of Term Specificity and its Application to Retrieval", Journal of Documentation Vo. 28, No. 1, pp. 111-121, 1972.

[12] Xin Fu and Miao Chen, "Exploring the Stability of IDF Term Weighting", H. Li et al. (Eds.): AIRS 2008, LNCS 4993, pp. 10–21, 2008. Springer-Verlag Berlin Heidelberg 2008

[13] Man Lan, Chew-Lim Tan, Hwee-Boon Low, Sam-Yuan Sung, "A Comprehensive Comparative Study on Term Weighting Schemes for Text Categorization with Support Vector Machines" WWW 2005, Chiba, Japan. ACM 1-59593-051-5/05/0005. pp. 1032-1033, May 10.14, 2005.

[14] Hee-soo Kim, Ikkyu Choi, Minkoo Kim, "Refining Term Weights of Documents Using Term Dependencies" SIGIR'04, Sheffield, South Yorkshire, UK. ACM 1-58113-881-4/04/0007, July 25–29, 2004, pp. 502-503.

[15] Youngjoong Ko, "A Study of Term Weighting Schemes Using Class Information for Text Classification". SIGIR'12, Portland, Oregon, USA. ACM 978-1-4503-1472-5/12/08. August 12–16, 2012, pp. 1029-1030.

[16] KWOK, K. L. 1995. A network approach to probabilistic information retrieval. ACM Trans. Inf. Syst. 13, 3, 324–353.

[17] Wu, H. C., Luk, R. W. P., Wong, K. F., and Kwok, K. L. 2008. Interpreting TF-IDF term weights as making relevance decisions. ACM Trans. Inform. Syst. 26, 3, Article 13 (June 2008), 37 pages. DOI = 10.1145/1361684.1361686 http://doi.acm.org/10.1145/1361684.1361686

[18] Man Lan, Sam-Yuan Sung, Hwee-Boon Low, Chew-Lim Tan, "A Comprehensive Comparative Study on Term Weighting Schemes for Text Categorization", In the Proceedings of International Joint Conference on Neural Network (IJCNN '05). pp. 1-6, July 2005. Montreal, Canada.

[19] Jiaul H. Paik, "A Novel TF-IDF Weighting Scheme for Effective Ranking". SIGIR'13, pp. 343-352, July 28–August 1, 2013. Dublin, Ireland. ACM 978-1-4503-2034-4/13/07

[20] Chandra Shekhar Jangid, Santosh K Vishwakarma, Kamaljit I Lakhtaria, "Ad-hoc Retrieval on FIRE DataSet with TF-IDF and Probabilistic Models", International Journal of Computer Applications (0975 – 8887), Volume 93 – No. 18, pp. 22-25, May 2014.

[21] Jiahua Du, Jing He, "Proportioning Documents over Categories based on Word Embeddings", ACSW '17 Geelong, Australia, ACM, ISBN 978-1-4503-4768-6/17/01 DOI: http://dx.doi.org/10.1145/3014812.3014819, pp. 1-5 2017.

[22] Madhu Kumari, Akshat Jain and Ankit Bhatia, "Synonyms Based Term Weighting Scheme: An Extension to TF.IDF". Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016) doi: 10.1016/j.procs.2016.06.093

[23] Khoo Khyou Bun, Mitsuru Ishizuka, "Topic Extraction from News Archive Using TF*PDF Algorithm". In Proceedings of the 3rd international conference on Web Information Systems Engineering, IEEE Computer Society Washington, DC, USA, pp. 73-82, Dec. 2002. Print ISBN: 0-7695-1766-8.

[24] Fabio Crestani, Mounia Lalmas, Cornelis J. Van Rijsbergen, Iain Campbell, "Is This Document Relevant?. . Probably": A Survey of Probabilistic Models in Information Retrieval ACM Computing Surveys, Vol. 30, No. 4, pp. 532-534 December 1998.

[25] Youngjoong Ko, "A Study of Term Weighting Schemes Using Class Information for Text Classification", SIGIR'12, August 12–16, 2012, Portland, Oregon, USA. ACM 978-1-4503-1472-5/12/08. pp. 1029-1030.

[26] M. Kan, C.-L. Tan and H.-B. Low, "Proposing a new term weighting scheme for text categorization". In AAAI 2006, pp. 763-768, 2008.

[27] Man LAN, Chew-Lim TAN, Hwee-Boon LOW, "Proposing a new term weighting scheme for Text Categorization". AAAI'06 Proceedings of the 21st National Conference on Artificial Intelligence – Volume 1 Pages 763-768, July 16-20, 2006.

[28] Sa-Kwang Song a,b, Sung Hyon Myaeng b, "A novel term weighting scheme based on discrimination power obtained from past retrieval results". In Proceedings of Information Processing and Management 48. pp. 919-930, 2012.