© 2015 IJSRSET | Volume 1 | Issue 3 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Recall Improvement in Information Storage and Retrieval System by Enhancing Query Knowledge

Dharmendra Sharma, Dr. Suresh Jain
Department of Computer Science and Engineering, Mewar University, Chittorgarh, Rajasthan, India

ABSTRACT

The main focus of this paper is the query knowledge improvement scheme for information storage and retrieval system to improve the recall. In information storage and retrieval system instead of searching query lexeme if we search query lexeme with its related lexeme then the recall value improved. Our hypothesis is that a meaning of a lexeme generally decided by its related lexeme for example the meaning of query containing the lexeme "apple phone" is well describe by its related lexeme as "audio", "video", "iphone", " ipad" etc. Thus instead of searching the lexeme "apple phone" if we search "apple phone" with its related lexeme like "audio", "video", "ipad", "iphone" then recall value improved. From the result we have seen that by using the query lexeme with its related lexeme the recall value is improved by 28 percent.

Keywords: query, recall, information storage and retrieval system, lexeme, hyponymy.

I. INTRODUCTION

With the exponential growth of the World Wide Web (WWW), it has become the most popular place to gather information. However the size of the WWW makes it difficult for people to locate relevant documents [1][2]. About 85% of all Web users use search engines of some kind for this purpose [3][4]. However, existing search engines often do not return relevant document. Many Web users have been dissatisfied with using search engines. The main reasons for dissatisfaction are the inability to find relevant document. As number of information resources available on the Web have been increasing rapidly. Especially, there are a lot of fragmental data which are created by each person's device or which are created by amount of sophisticated sensors for science curiosities. Briefly speaking, we are not only retrieving but also creating these data every day which lead to mount of multi dimensional data on web .It is called "Big Data Era". Unfortunately, the available Big Data is not properly used by search engine to retrieve the relevant document. In the paper we use the hypothesis that the meaning of a lexeme generally decided by its related lexeme for example the meaning of "apple phone" is well describe by its related lexeme like audio, video, ipad, iphone etc. thus instead of searching the lexeme "apple phone" if we search "apple phone" with its related lexeme "ipad", "iphone", "audio" "video" then recall value improved

The remaining sections of this paper consist of four different parts. In related work section we focus on the work done in the performance improvement of the information storage and retrieval system. The methodology section describes the process of implementing the system. The result and discussion section focus our findings and significant. In conclusion we conclude our work with limitation and future work.

II. RELATED WORK

An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy.

In information retrieval process a query play a key role in retrieving of document because the lexemes present in query are compared with the surface lexeme of documents. The similarity measures between queries and information resources relevant to users' documents needs have been

studied for a long time. The most popular and basic method is vector space model [5]. Feature reduction techniques of vector space model have been used for developing traditional vector space models such as latent semantic indexing [6] and the mathematical model of meaning [7][8]. These techniques are applied to information resources, characterized by elements in a flat domain. However, it is to be noted that when the elements have a tree structure, all the elements are not orthogonal to each other. A few studies have used computational measures of feature relationships [9] in an orthogonal vector space. The mathematical model of meaning realizes a context-based dynamic semantic computation. However, it has to prepare a space for the semantic commutations before There have been studies defining similarity metrics for hierarchical structures such as WorldNet [10]. Rada et al. [11] have proposed a "conceptual distance" that indicates the similarity between concepts of semantic nets by using path lengths. Some studies [12][13] have extended and used the conceptual distance for information retrieval. Resnik [14] has proposed an alternative similarity measure based on the concept of information content. Ganesan et al. [15] have presented new similarity measures in order to produce more intuitive similarity scores based on traditional measures. On the other viewpoints, the reference [16] is surveyed. This survey [16] shows common architecture and general functionality as OBIE from various ontology-based information extraction researches. It consists of "information extraction module", "ontology generator", "ontology editor", "semantic lexicon" and some preprocessors. Their researchers are working for both of various researches of OBIE system implementation and research focusing on each module. Vargas-Vera et al. [17] proposed Semantic Annotation Tool for extraction of knowledge structures from web pages through the use of simple user-defined knowledge extraction patterns. KIM [18] provides a mature and semantically enabled infrastructure for scalable and customizable information extraction. IDocument [19]. The reference [20] describes survey about the weighting methods such as binary [20], lexemefrequency (TF) [20], augmented normalized lexemefrequency [20][21], log [21], inverse document frequency (IDF) [20], probabilistic inverse [20][21], document length normalization [22].

However, our method differs in the purpose from other methods. The purpose of our method is to improve the recall value by enhancing the query knowledge. We use the query lexeme with its related lexeme to improve the recall value.

.

III. METHODOLOGY

In this section, we present the method of calculating the recall value by using query lexeme and its related lexeme. In Information storage and retrieval system the recall value calculated as the ratio of number of relevant document retrieved for the query and total number of relevant documents exists in document set. For the input documents set we use Google search engine to aggregate the documents set. Our method consists of four different steps.

A. Data Filtration

Before creating the document term index we filter out the text documents set by eliminating the stop words and other meaningless terms. We use following steps to filter the data

1. Convert text into lower case.
2. Remove stop word, number and punctuation character from text.
3. Calculate the term frequency for each documents.

B. Document Term Index creation:

For effectively retrieving relevant documents by information retrieval strategies, the documents are typically transformed into a suitable representation. Each retrieval strategy incorporates a specific model for its document representation. We use vector space model to create the document term index. In vector space model Documents and queries are represented as vectors.

$$D_j = (w_{1j}, w_{2j}, \ldots . w_{tj})$$
$$Q = (w_{1q}, w_{2q}, \ldots . w_{nq})$$

Each dimension corresponds to a separate term. If a term occurs in the document, its value in the vector is non-zero.

C. Lexeme selection:

In linguistics, a hyponym is a word or phrase whose semantic field is included within that of another word, its hyperonym or hypernym. In simpler terms, a hyponym shares a type-of relationship with its hypernym. For example, pigeon, crow, eagle and seagull are all hyponyms of bird which, in turn, is a hyponym of animal. For the selection of related lexeme of query lexeme we use the dictionary based approach. From the dictionary we have seen that the query containing the

lexeme “apple phone” is well described by the lexeme “audio”,video”, “ipad” and “iphone”.

D. Query Evaluation:

To calculate the similarity between query and document any type methods are acceptable such as dot product, distance, norm, cosine similarity, etc. In this paper, we use dot product. The dot product of two vectors A and B are defined as

$$\mathbf{A} \cdot \mathbf{B} = \sum_{i=1}^{n} A_i B_i = A_1 B_1 + A_2 B_2 + \cdots + A_n B_n$$

IV. EXPERIMENTAL SYSTEM SETTING

We use Google Web API for aggregation of web documents. We aggregate 64 documents to test our system.

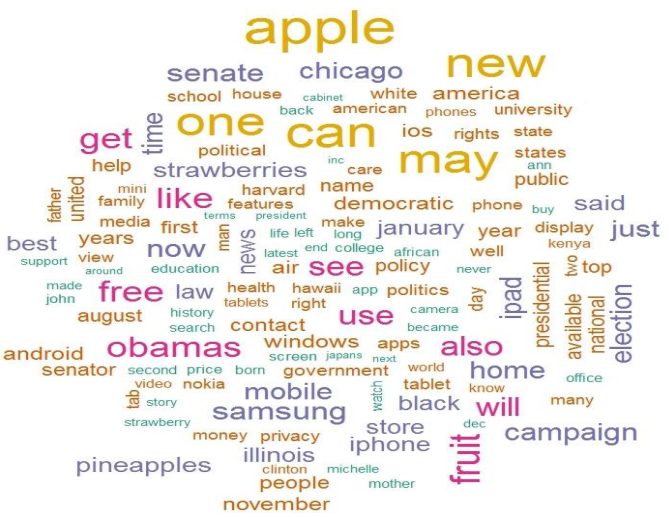

Figure 1: Lexeme Cloud

This experiment system is implemented by R including text mining (tm) package.

In this experiment, the query contains the lexeme “apple phone”. We improve the query knowledge by adding the lexeme “audio”, “video”, “ipad” and “iphone” with query lexeme “apple phone”. Fig. 1 shows the lexeme cloud of used documents set. paper.

V. EXPERIMENTAL RESULT

We have shown the result in the case of query containing the lexeme “apple phone” in table 1. There are sixty four documents according to query containing the lexeme “apple phone”. For the query containing the lexeme “apple phone” only 38 documents are retrieved but when we search with lexeme “audio”, “video”, “ipad” and “iphone” then 56 documents were retrieved. The additional lexeme “audio”, “video”, “ipad” and “iphone” improve the recall by 28%.

TABLE I: QUERY WITH LEXEME “apple phone”

S.N.	Query term	Document retrieved	Recall %
1	Apple, phone	38	59
2	Apple, phone, audio	40	62
3	Apple, phone, audio, video	44	68
4	Apple, phone, audio, video, ipad	48	75
5	Apple, phone, audio, video, ipad Iphone	56	87

Next, we show the result in the case of query containing the lexeme “apple fruit” in table II.

TABLE II: QUERY WITH LEXEME “apple fruit”

S.N.	Query term	Document retrieved	Recall %
1	Apple, fruit	48	75
2	Apple, fruit, eat	52	81
3	Apple, fruit, eat, vitamin	53	82
4	Apple, fruit, eat, vitamin, health	55	85
5	Apple, fruit, eat, vitamin, health, organic	59	92

From the table II it is clear that for the query containing the lexeme “apple fruit” only 48 documents are retrieved but when we search with lexeme “eat”, “vitamin”, “health” and “organic” then 59 documents were retrieved. The additional lexeme “eat”, “vitamin”, “health” and “organic” improve the recall by 17%.

VI. CONCLUSION

In this paper, we presented the query knowledge improvement scheme for information storage and retrieval system to improve the recall. In information storage and retrieval system instead of searching a query lexeme if we search query lexeme with its related lexeme then the recall value improved. A meaning of a lexeme generally decided by its related lexeme. We use small example for simplification and detailed checking of operations for our method. In our work we use query lexeme as "apple phone" and related lexeme as audio, video, ipad and iphone. From the result we have seen that by using the query lexeme with its related lexeme the recall value is improved by 28 percent. In one another experiment we use the query lexeme as "apple fruit" with its related lexeme "vitamin", "health", "eat" and "organic". From the result it is clear that in information storage and retrieval system the recall is depend upon the query lexeme and its related lexeme.

VII. REFERENCES

[1] O. King and M. Kobayashi, "Information Retrieval and Ranking on the Web: Benchmarking studies II", 1999.

[2] I. Melve, "Web Caching Architecture," *DESIRE Web* caching team, 2001.

[3] G.E. Dupret and M. Kobayashi "Information Retrieval and Ranking on the Web: Benchmarking studies I," *IBM TRL* Research Report, 1999.

[4]] M. Kobayashi and K. Takeda, "Information Retrieval on the Web," *IBM Research*, 2000.

[5] G. Salton, A. Wong, C. S. Yang, "A vector space model for automatic indexing," *Magazine Communications of the ACM CACM Homepage archive*, vol.18(11), pp. 613-620, Nov. 1975.

[6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, vol. 41(6), pp.391-407, 1990.

[7] T. Kitagawa, Y. Kiyoki. A mathematical model of meaning and its application to multidatabase systems. In RIDE-IMS '93: Proceedings of the 3rd International Workshop on Research Issues in Data Engineering: Interoperability in Multidatabase Systems, pp. 130-135, 1993.

[8] Y. Kiyoki, T. Kitagawa, T. Hayama. A metadatabase system for semantic image search by a mathematical model of meaning. *SIGMOD Rec.*, vol. 23(4), pp.34-41, 1994.

[9] K. Takano, Y. Kiyoki, "A superordinate and subordinate relationship computation method and its application to aerospace engineering information," *In ACST'07: Proceedings of the third conference on IASTED International Conference*, pp. 510-516, Anaheim, CA, USA, 2007.

[10] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller. "Introduction to LexemeNet: An on-line lexical database," *Journal of Lexicography*, vol.3(4), pp.235-244, January 1990.

[11] R. Rada, H. Mili, E. Bicknell, M. Blettner, "Development and application of a metric on semantic nets," *IEEE Transactions on Systems, Man and Cybernetics*, vol.19(1), pp. 17-30, Jan/Feb 1989.

[12] Y. Kim, J. Kim, "A model of knowledge based information retrieval with hierarchical concept graph," *Journal of Documentation*, vol.46(2), pp.113-136, 1990.

[13] Y. Li, K. Bontcheva, "Hierarchical, perceptron-like learning for ontology-based information extraction," *In Proceedings of the 16th international conference on World Wide Web (WWW '07)*, ACM, New York, NY, USA, pp. 777-786, 2007.

[14] C. Hwang, "Incompletely and imprecisely speaking: Using dynamic ontologies for representing and retrieving information," In *Proceedings of the 6th international workshop on ontology-based information extraction system.* Kaiserslautern, Germany, 1999.

[15] B. Yildiz, S. Miksch "ontoX - A Method for Ontology-Driven Information Extraction," *Lecture Notes in Computer Science*. 4707, pp. 660-673, 2007.

[16] A. Todirascu, L. Romary, D. Bekhouche, "Vulcain — An Ontology- Based Information Extraction System," *Lecture Notes in Computer Science*. 2553, pp. 64-75, 2002.

[17] M. Vargas-Vera, E. Motta, J. Domingu, S. Shum, M. Lanzoni, "Knowledge extraction by using an ontology-based annotation tool," *In Proceedings of the workshop on knowledge markup and semantic annotation*, ACM, New York, NY, USA, 2001.

[18] B. Popov, A. Kiryakov, D. Ognyanoff, D. Monov, A. Kirilov, KIM – a semantic platform for information extraction and retrieval. *Natural Language Engineering*, vol. 10(3-4), (September 2004), pp. 375-392,2004.

[19] B. Adrian, J. Hees, L. Elst, A. Dengel, iDocument: Using Ontologies for Extracting and Annotating Information from Unstructured Text. *Lecture Notes in Computer Science*. 5803, pp.249-256, 2009.

[20] T. G. Kolda, D. P. O'Leary, "A semidiscrete matrix decomposition for latent semantic indexing information retrieval", Journal ACM Transactions on Information Systems (TOIS) TOIS Homepage archive vol.16(4), pp. 322-346, Oct. 1998.

[21] G.Salton, C. Buckley, "Lexemeweighting approaches in automatic text retrieval," *Inf. Process. Manage*. 24, pp. 513–523, 1988.

[22] D. Harman, "Ranking algorithms. In Information Retrieval: Data Structures and Algorithms," *W. B. Frakes and R. Baeza-Yates, Eds. Prentice Hall, Englewood Cliffs*, NJ, pp.363–392, 1992.