Query Knowledge Enhancement in Information Storage and Retrieval System by using Term Correlation

[1]Dharmendra Sharma ,[2]Dr. Harish Nagar
[1]Research scholar,[2]Reasearch Supervisor
[1]Department of Computer Science and Engineering, Mewar University, Chittorgarh, India

***Abstract* - the main focus of this paper is the query knowledge enhancement scheme for information storage and retrieval system to improve the recall by using term correlation. In information storage and retrieval system instead of searching query term if we search query term with its related term then the recall value improved. Our hypothesis is that a meaning of a term generally decided by its related term for example the meaning of query containing the term "apple fruit" is well describe by its related term as "banana", "health", "vitamin", "eat" etc. Thus instead of searching the term "apple fruit" if we search "apple fruit" with its related term then recall value improved. In this work we have calculate the term correlation between all query terms and terms exist in document set. We rank the term by their correlation values. The combination of top five terms with query terms are used to retrieve the document. From the result we have seen that by using the query term with its related term the recall value is improved by 20 percent.**

***Index Terms* - query, recall, information storage and retrieval system, term, hyponymy, lexeme.**

I. INTRODUCTION

Amazing development of Internet and digital library has triggered a lot of research areas. Text categorization is one of them. Text categorization is a process that group text documents into one or more predefined categories based on their contents [1]. It has wide applications, such as email filtering, category classification for search engines and digital libraries. Associative text classification, a task that combines the capabilities of association rule mining and classification, is performed in a series of sequential subtasks. They are the preprocessing, the association rule generation, the pruning and the actual classification. Out of these, the first step, that is, 'Preprocessing', is the most important subtask of text classification. The importance of preprocessing is emphasized by the fact that the quantity of training data grows exponentially with the dimension of the input space. It has already been proven that the time spent on preprocessing can take from 50% up to 80% of the entire classification process [2], which clearly proves the importance of preprocessing in text classification process. 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 this work we have calculate the term correlation between all query terms and terms exist in document set. We rank the term by their correlation values. The combination of top five terms with query terms are used to retrieve the document

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 WORE

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 terms present in query are compared with the surface term 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], term frequency (TF) [20], augmented normalized term frequency [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 term with its related term to improve the recall value.

III. METHODOLOGY

In this section, we present the method of calculating the recall value by using query term with combination of related term. 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 Filteration

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. Term selection:

For the selection of term we evaluate the term correlation between query terms and term present in document set. We create the sorted list of terms exist in document set on the basis of their correlation values. The below table shows the term correlation values for query term "apple fruit".

Table I : Correlation matrix

Document term	Query term		
	"Apple"	"Fruit"	Average
Banana	0.9	0.87	0.45
Nutrition	0.87	0.83	0.435
Vitamin	0.78	0.78	0.39
Organic	0.77	0.68	0.385
Seed	0.65	0.62	0.325
Eat	0.45	0.56	0.225
Strawberry	0.34	0.35	0.17
Tree	0.23	0.21	0.115
health	0.2	0.01	0.1

We have taken the average of query terms and document terms to create the combination.

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 500 documents to test our system that include 340 documents related with apple fruit. This experiment system is implemented by R including text mining (tm) package. In this experiment, the query contains the term "apple fruit". We improve the query knowledge by selecting the term from term correlation matrix

V. EXPERIMENTAL RESULT

We have shown the result in the case of query containing the term "apple phone" in table I. There are five hundred documents according to query containing the term "apple phone". For the query containing the term "apple phone" only 244 documents are retrieved but when we search with term "Banana", "Nutrition", "Vitamin" etc the number of document retrieved by system improved as 340. Thus adding up of related term improves the recall value by 20%. We show the result in the case of query containing the term "apple fruit" in table I.

TABLE I: QUERY WITH TERM "apple fruit"

S.N.	Query term	Document retrieved	Recall %
1	Apple, Fruit	244	72
2	Apple, Fruit, Banana	260	76
3	Apple, Fruit, Banana, Nutrition	275	81
4	Apple, Fruit, Banana, Nutrition, Vitamin	280	82
5	Apple, Fruit, Banana, Nutrition, Vitamin, Organic	301	88
6	Apple, Fruit, Banana, Nutrition, Vitamin, Organic ,Seed	312	92

VI. CONCLUSION

In this paper, we have presented the automated process of query knowledge enhancement for information storage and retrieval system to improve the recall. In information storage and retrieval system instead of searching a query term if we search query term with its related term then the recall value improved. In our work we use query term as "apple fruit" and related term as banana, nutrition, vitamin, organic and seed. From the result we have seen that by using the query term with its related term the recall value is improved by 20 percent.

REFERENCES

[1]. K.Aas and A.Eikvil, "Text categorization: A survey", Technical report, Norwegian Computing Center, June, 1999.

[2]. Katharina, M. and Martin, S. the Mining Mart Approach to Knowledge Discovery in Databases Intelligent Technologies for Information Analysis,Springer, Pp. 47-65. (2004)

[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 TermNet: 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, "Termweighting 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.