Query based Text Document Clustering using its Hypernymy Relation

S.Vijayalakshmi
Assistant professor
Sethu Institute of Technology
Department of Applied Sciences,
Kariapatty

Dr.D.Manimegalai
Professor& Head
Department of Information Technology
National Engineering College
Kovilpatty

ABSTRACT
Clustering of text can be organized in an unsupervised manner. In this paper, Text document clustering is done based on query and its semantic relation. The method utilizes hypernymy to identify its relation. It was detected by using the Word Net. It act as background knowledge of the Query and provides its synonymic terms. This paper proposed the new term-document matrix called Query based document vector model, which is constructed using query with two terms and its hypernymy. The results show that our new measure Cluster Accuracy is significantly better to evaluate the quality of cluster and better results are obtained.

General Terms
Text mining and Information Retrieval, Partitioning Method

Keywords
Clustering, Noun, Word net, Query based document vector model, Hypernymy, Accuracy.

1. INTRODUCTION
Information retrieval, Information Extraction and Text Mining [1] play an important role, due to the growth of the enormous amount of text document. Document clustering is a knowledge discovery technique which categorizes the document set into meaningful groups. TDC is used for efficient organization and summarization of a large volume of documents to quickly obtain the desired information. Since the past few years, the usage of additional knowledge resources like Word Net and Wikipedia, as external knowledge base has increased. Most of these techniques map the documents to Wikipedia, before applying traditional [2] document clustering approaches (Tan et al., 2006; Kaufman and Rousseeuw, 1999). (Huang et al., 2009) (Huang et al., 2008) uses Wikipedia based concept-based representation and active learning. Enriching text representation with additional Wikipedia features can lead to more accurate clustering texts. (Banerjee et al., 2007) (Hu et al., 2009). This paper addresses two issues:

The first issue of concern is to use Interactive query for document clustering. Noun based interactive query leads to effective improvements and enriches result in system execution.

Another issue is to reduce the information gap between conceptually similar terms. It is achieved by using hypernymy. The term with its hypernymy reduces the information gap between terms, which are conceptually same but do not match physically.

Finally, we proposed Query with its hypernymy based term-document matrix for document representation, which uses K-Means clustering.

2. RELATED WORK
There are two ways to search a document collection organized in taxonomy [3]. First method is Top-down Search (Manu Kochady, 2006), begin at the root of the taxonomy and search for a specific cluster by comparing the query with cluster representatives at lower levels. One risk with this method is the possibility of making a cluster matching error that leads to a wrong path in taxonomy. Alternate method is bottom-up search. Queries are compared with most specific clusters at the lowest level. The likelihood of finding irrelevant document is lower when we compare queries against a set of specific cluster. Query expansion [4] and related issues have been long interest in IR research. The literature on how searchers formulate and reformulate their queries [13] to improve precision and/or recall is extensive. Some of the major problems [14] with information retrieval in search engines are general keyword searching as the default. Clustering [5][6][7][8] is an important unsupervised classification technique used in identifying some inherent structure present in a set of objects. The purpose of cluster analysis [5] is to classify objects into subsets that have some meaning in the context of a particular problem. When striving to name a concept with great precision, it helps to understand synonyms, hypernyms and hyponyms. The root "nym" comes from the Greek onoma, a name. We use the term nym to identify many classes of words. In this module, User will look at two specific nyms: hypernyms, and hyponyms. Users are familiar with synonyms: words that mean the same thing. It is also wise to scan the results of your first search for strong keywords. Synonyms (or any keyword) act as either hyponyms or hypernyms with the contents stored in the database. It improves effectiveness [9][10] of the clustering.

3. PREPARING DOCUMENT VECTOR MODEL
In this phase, we prepare a document vector model for document clustering. Instead of using bag-of words based approach, our scheme uses the index of the refined Query (N) in noun form (N = 1 and 2), their weighted importance in document .

3.1 Identifying hypernymy
According to the result provided by Hai-Tao Zheng, et.al, 2009, [11] Hypernymy gives better semantical relationship than Hyponymy, Meronymy and Holonymy. Hypernymy is used to broaden the search. So we used only hypernymy with respect to query. User query will be passed to collect hypernymy list from

Word Net. User will select required terms. These hypernymy terms act as Input query (Fig 1) and maintain the result separately in the database. It will be added to vector model as represented in(Fig 2).

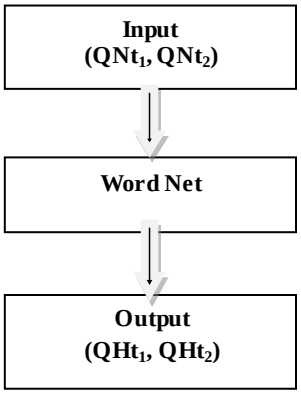

Fig 1: Process of identifying Hypernymy

3.2 Using 20 News group

The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. It was originally collected by Ken Lang. The 20 newsgroups collection has become a popular data set for experiments in text applications and machine learning techniques, such as text classification and text clustering. Here we have been chosen randomly from all 20 groups from the collection.

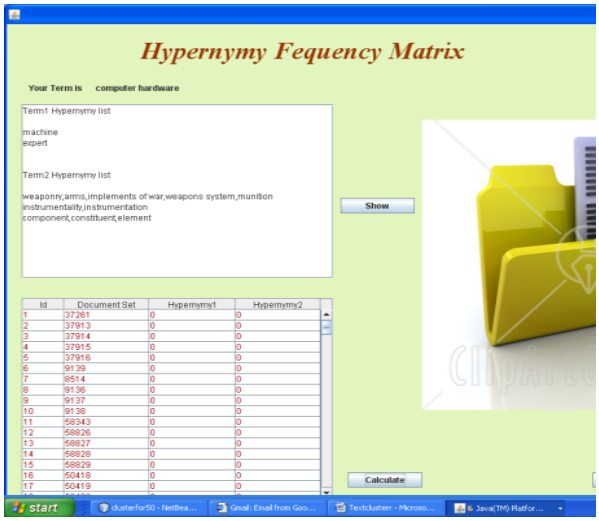

Fig 2: Identifying Hypernymy using Word Net

3.3. Query based Document Vector Model

Creating query-based document representation by mapping terms and phrases within documents available in the corpus (20 News group) and calculating a similarity measure that evaluates the semantic relatedness between terms in the document.

Step1: Documents and queries are represented as vectors.

$$d_j = (t_{1,j}, t_{2,j}, \ldots, t_{n,j}) \quad \ldots\ldots\ldots E1$$

$$q = (QN_{t1}, QN_{t2}) \quad \ldots\ldots\ldots E2$$

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

Step 2: Calculating Every identified hypernymy

Extract hypernymy based on User query from the Word Net .If the hypernyms occur in the document, its value in the vector is non-zero.

$$d_j = (t_{1,j}, t_{2,j}, \ldots, t_{n,j}) \quad \ldots\ldots\ldots E1$$

$$q = (QH_{t1}, QH_{t2}) \quad \ldots\ldots\ldots E3$$

Step 3: To evaluate the effectiveness,

a. Find the hypernymy frequency to calculate the weight of the query.

b. Construct term document matrix with four attributes. First two attribute can be calculated weight based on user query (QN_{t1}, QN_{t2}) and the remaining two attributes based on the hypernymy term (QH_{t1}, QH_{t2}) with respect to the user query.

c. calculate the term frequency and Inverse Term Document Frequency

$$TF_{ij} = f_{ij} / max\{f_{ij}\} \quad \ldots\ldots\ldots E4$$

f_{ij} - Frequency of noun terms or hypernymy term in document j

$$IDF = \log(N/n_j) + 1 \quad \ldots\ldots\ldots E5$$

IDF - determining which documents are most relevant to a query high weight in TF–IDF is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms. The TF-IDF value for a term will be greater than zero if and only if the ratio inside the IDF's log function is greater than 1. Depending on whether a 1 is added to the denominator, a term in all documents will have either a zero or negative IDF, and if 1 is added to the denominator a term that occurs in all but one document will have an IDF equal to zero.

d. To find the $Wn_{1,j}$, $Wn_{2,j}$, Wh_{1j}, Wh_{2j}, are the statistical measure used to evaluate how important a word is to a document in a collection.

$$W_{n1,j} = TF_{n1,j} \times IDF_{n1,j} \quad \ldots\ldots\ldots E6$$

$$W_{n2,j} = TF_{n2,j} \times IDF_{n2,j} \quad \ldots\ldots\ldots E7$$

$$W_{h1,j} = TF_{h1,j} \times IDF_{h1,j} \quad \ldots\ldots\ldots E8$$

$$W_{h2,j} = TF_{h2,j} \times IDF_{h2,j} \quad \ldots\ldots\ldots E9$$

$$\begin{bmatrix} & QN_{t1,j} & QN_{t2,j} & QH_{t1,j} & QH_{t2,j} \\ D_1 & W_{n1,1} & W_{n2,1} & W_{h1,1} & W_{h2,1} \\ D_2 & W_{n1,2} & W_{n2,2} & W_{h1,2} & W_{h2,2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ D_j & W_{n1,j} & W_{n2,j} & W_{h1,j} & W_{h2,j} \end{bmatrix}$$

Fig 3: Proposed Document representation using Query terms in noun form and it hypernymy

The above matrix (Fig 3) was implemented in Net beans and MySQL and gives weight matrix. The above mentioned Query based vector model uses K-Means algorithm.

4. DOCUMENT CLUSTERING USING K-MEANS

K-means clustering [11][12][1] is one of the simplest and most popular unsupervised learning algorithms to solve the clustering problem. K-means clustering [13] generates a specific number of disjoint, flat clusters. That is, the K-means function partitions the observations extracted from the data into k mutually exclusive clusters, and returns a vector of indices, indicating to which of the k clusters each feature set has been assigned.

This method is more efficient than hierarchical clustering [14], especially for large data sets and high-dimensional data sets.

The basic algorithm for the k-means method is as follows:

1. Specify the number of clusters k and then randomly select k observations to initially represent the k cluster centers. Each observation is assigned to the cluster corresponding to the closest of these randomly selected objects to form k clusters.

2. The multivariate means (or ``centroids") of the clusters are calculated, and each observation is reassigned (based on the new means) to the cluster whose mean is closest to it to form k new clusters.

3. Repeat step 2, until the algorithm stops when the means of the clusters are constant between two consecutive iteration.

In the traditional k-means approach, ``closeness" to the cluster centers is defined in terms of squared Euclidean distance.

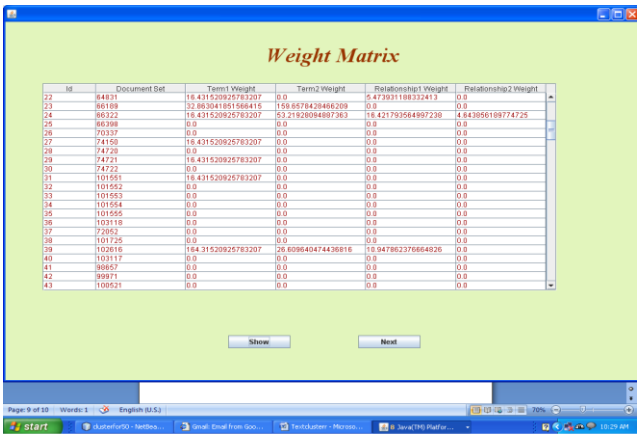

Fig 4: Implementation of Weight Matrix Using Net Beans & MSQL

Evaluation Metrics: Cluster quality is evaluated by using Cluster Accuracy. It is a standard evaluation metrics and we have used to evaluate the accuracy of cluster using the following:

Cluster Accuracy= Tot. No. of docs clustered correctly /Tot. No. of docs…….E10

Experiment Input:

Total number of documents: 200

All retrieved documents are compared with root level and produced two clusters. So both clusters contain documents relevant to the query. The accuracy of the cluster is evaluated using E10 and it is listed in Table-1. Documents in cluster accuracy vary based on User Query. Since, K-Means with our proposed document representation has a higher accuracy with cluster I except the query Universal problem (Figure 5) and based on the results obtained using the 20-Newsgroup dataset (Table 1), it is clear that the proposed scheme performs better.

Total number of clusters: 2

5. CONCLUSIONS

In this paper, we introduce a new Query based term document matrix with its hypernymy to enrich the term Document representation. This approach improves the Query based similarity in the cluster, which results in an improvement in accuracy of clusters. In the proposed method constructing Term document matrix is based on Query terms. According to the results obtained, Cluster-I is comparatively better than the cluster-II

.

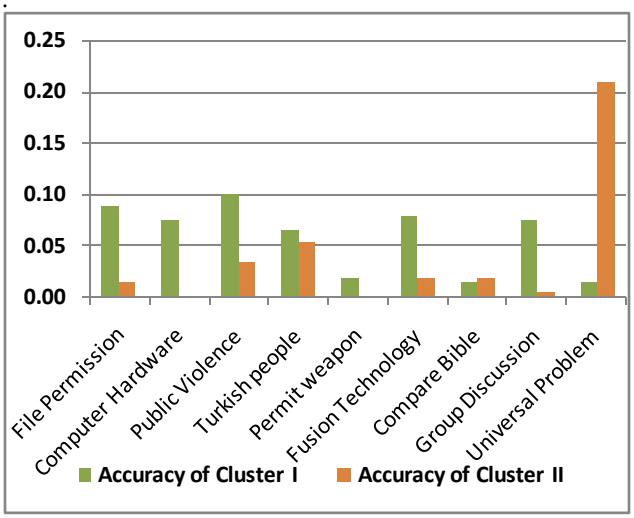

Fig 5: Comparison of Accuracy between Clusters I & II

5.1. Further Enhancement

User must know exactly what is he required. User must give effective keyword. Formulation of queries is difficult for many Web searchers. Preprocessing work must be done in both Query and Dataset every time. It is a time consuming process. Before sending the weight matrix for clustering, sparse region must be identified and eliminated to improve the cluster quality. Further work may be carried out in this direction

Table 1. Comparison of Clustering Accuracy between Cluster I & Cluster II

K = 2 (Total No. of Doc-200)				
keywords	**No. of doc. clustered in Cluster- I**	**Cluster- I Accuracy**	**No. of doc. clustered in Cluster-II**	**Cluster- I Accuracy**
File Permission	18	0.09	3	0.02
Computer Hardware	15	0.08	0	0.00
Public Violence	20	0.10	7	0.04
Turkish people	13	0.07	11	0.06
Permit weapon	4	0.02	0	0.00
Fusion Technology	16	0.08	4	0.02
Compare Bible	3	0.02	4	0.02
Group Discussion	15	0.08	1	0.01
Universal Problem	3	0.02	42	0.21

6. REFERENCES

[1] Congnan Luo, Yanjun Li, Soon M.Chung, titled "Text document clustering based on neighbors", Journal on Data & Knowledge Engineering, Volume 68, Issue 11, November 2009, Pages 1271-1288.

[2] Kaufman. L and Rousseauw. P, 1991, Finding Groups in data: An introduction to cluster analysis, 1999, John Wiley & Sons.

[3] Manu Kochady, Textmining Application Programming, Akash Press, Delhi- 20, 2007.

[4] Makrechi. M, "Query-relevant document representation for text clustering", Digital Information management (ICDIM), 2010, Fifth International Conference on Digital Object Identifier: 10.1109/ICDIM.2010.5664205 pages: 132-138.

[5] M.R.Anderberg, Cluster Analysis for Application, Academic Press, New York, 1973.

[6] A.K.Jain and R.C. Dubes, Algorithm for clustering data, Prentice Hall, Englewood Cliffs NJ, 1998.

[7] S.Guha, R.Rastogi and K.Shim, CURE: An efficient clustering for large databases in Proceedings of the ACM SIGMOID, International Conference on Management of Data, 1998, pages 73-84.

[8] H. Frigui, R. Krishnapuram, A robust competitive clustering algorithm with application in computer vision, IEEE Trans. Patt. Anal. Machine Intelligence. 21 (1) (1999) pages 450–465.

[9] Fazli Can, Ismail Sengor Altingovde, Engin Demir,"Efficiency and effectiveness of query processing in cluster based retrieval", Journal on Information Systems, ScienceDirect, Volume 29, Issue 8, December 2004, Pages 697-717.

[10] Anastasios Tombros,Robert Villa, C.J.Van Rijsbergen,"The effectiveness of query specific hierarchical clustering in information retrieval", Journal on Information processing & Management, Volume 38, Issue 4, July 2002, pages 559-582.

[11] Hai-Tao Zheng, Bo-Yeong Kang, Hang-Gee Kim, "Exploiting noun phrases and semantic relationships for text document clustering", Journal on Information Sciences 179 (2009), ScienceDirect, page 2249-2262.

[12] Yanjun Li, Soon M.Chung, John D. Holt, "Text Document Clustering based on frequent word meaning sequences", Journal on Data & Knowledge Engineering, Volume 64, Issue 1, January 2008, Pages 381-404

[13] Meedeniya. D.A and Perera. A.S, "Evaluation of Partition-Based Text Clustering Techniques to Categorize Indic Language Documents", IEEE International Advanced Computing Conference, 2009(IACC 2009), Digital Object Identifier: 10.1109/IADCC.2009.4809239, Pages: 1497-1500.

[14] Jinxin Gaoa, David B. Hitchcock James Stein," shrinkage to improve k-means cluster analysis", Journal on Computational Statistics and Data Analysis (2010) pages. 2113 - 2127.