Automatic Naming of Domain Specific Clusters for Efficient Searching

Anshika Nagpal
Deptt of CS,
Meerut Institute of Engineering and Technology,
Meerut

Mukesh Rawat
Deptt of CS,
Meerut Institute of Engineering and Technology,
Meerut

ABSTRACT

This paper proposes a new and efficient methodology for clustering of html documents. The topic wise categorization of documents into different clusters makes searching easier and efficient. This technique can be utilized by search engines to provide relevant results to the user according to query and also utilized by online journal domains that are maintaining large set of documents. This paper suggests a good word matching and naming of automatic generated clusters , so, the time consume for finding the appropriate cluster for a document will be reduced. This paper shows the use of an efficient technique for finding the similarity between the documents and assigns them a proper cluster. The proper clustering of documents will be further utilized by multidocument summarization system, which produces a summary for the documents related to each other.

Keywords

Keywords are clustering, similarity, clusters etc.

1. INTRODUCTION

[10]As day to day numbers of websites are increasing at a tremendous rate. So a mechanism required so that retrieving of relevant documents against a search query by a search engine becomes faster and efficient. The paper suggested a new and efficient technique for proper clustering of web documents by keyword matching of the web documents, so that the indexing of the search results against a search query takes less time and also shows relevant results. By setting the title for a cluster helps in finding the proper cluster for a web document and also helps in finding the relevant results for a search query by matching the query terms with the title of the clusters.

2. EXISTING CLUSTER TECHNIQUE

k-Means Clustering Algorithm[11]

(1) Choose *k* cluster centres to coincide with *k* points inside the hyper volume containing the pattern set.

(2) Assign each pattern to the closest cluster center.

(3) Recomputed the cluster centers using the current cluster memberships.

(4) If a convergence criterion is not met, go to step 2.

Typical convergence criteria are: no (or minimal) reassignment of patterns to new cluster centers, or minimal decrease in squared error.

The clustering techniques can be categorized as:

Agglomerative vs. divisive:

This aspect relates to algorithmic structure and operation. An agglomerative approach begins with each pattern in a distinct (singleton) cluster, and successively merges clusters together until a stopping criterion is satisfied. A divisive method begins with all patterns in a single cluster and performs splitting until a stopping criterion is met.

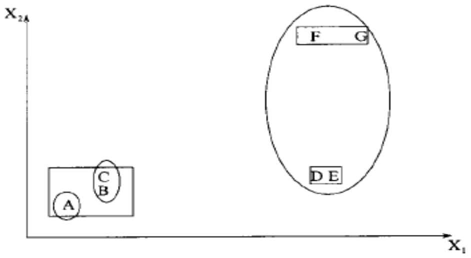


Figure 1 : The k-means algorithm is sensitive to the initial partition.

2.1 Vector Space model: T

[8]The standard way of quantifying the similarity between two documents d1 and d2 is to compute the cosine similarity of their vector representations

= $\overrightarrow{V(d1)}$ and $\overrightarrow{V(d2)}$

COSINE SIMILARITY

$$\mathrm{sim}(d\overrightarrow{1, d2)} = (V(\overrightarrow{d1)} \cdot V(\overrightarrow{d2))})/(|V\ (\overrightarrow{d1|)}(|V\ \overrightarrow{(d2|)}$$

where the numerator represents the dot product (also known as the inner product) of the vectors $\overrightarrow{V(d1)}$ and $\overrightarrow{V\ (d2)}$, while the denominator is the product of their Euclidean lengths. The Euclidean length d is defined to b e √ ∑M i=1 ~~V2i(d)~~

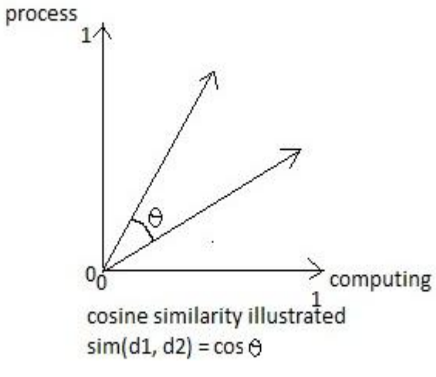


Figure 2 : Vector Space Model

2.2 Automatic Naming of Domain Specific Clusters for Efficient Searching

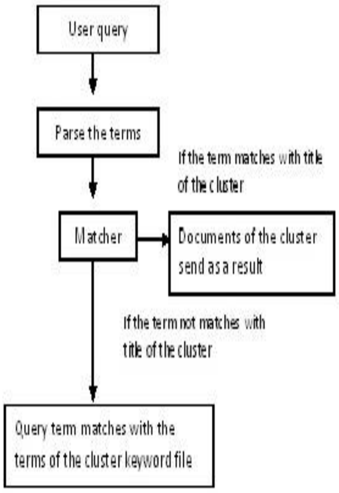

User query
Parse the terms
If the term matches with title of the cluster
Matcher
Documents of the cluster send as a result
If the term not matches with title of the cluster
Query term matches with the terms of the cluster keyword file

Figure 3 : flowchart for finding the proper cluster of documents against a search query supplied by the user

2.3 Algorithm for the Figure

```
Fetch_Result( )
begin
clusters[ ]  // array of generated clusters
terms[ ] ← Parse(query);
for each terms[i] in array terms[ ]
     For each cluster[i] in array clusters[ ]
        flag← Match(terms[i], title(clusters[i]))
        if(flag)
          begin
             return docs(cluster[i]);
          end
        else
           begin
              flag1←1
           end
       end for
         if(flag1)
           begin
              for each clusters[i] in clusters[ ]
               flag2←Match(terms[i], cluster_keyword_file(clusters[i]))
                if(flag2)
                     begin
                        return docs(clusters[i]);
                     end
                 else
                      begin
                          flag3←1
                      end
                 end for
              if(flag3)
begin
no result
end
end for
```

The work has been divided into two sections: html document processing and automatic generation of clusters and assign name for a cluster. In first section web documents which are in the form of html documents are parsed by removing tags from the html document and converting into a text file, then from this text file stop words, cue words and most frequently used words such as the, is are, they etc are removed.

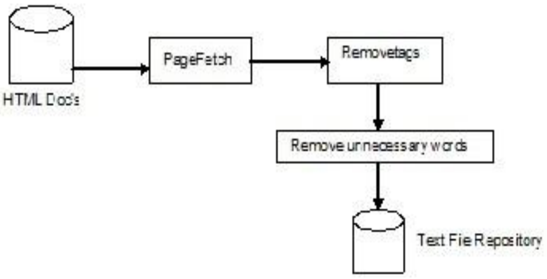


Figure 4: General Architecture for Html document processing

The automatic generation of cluster is done with the help of maintaining cluster keyword file that contains the keywords presented in the documents of a cluster.

The cluster keyword file maintains the keywords of the documents appearing in the cluster , for deciding the appropriate cluster for a new document. The similarity is measured between the keywords of the document and the terms of the cluster keyword file of each of the cluster and if the similarity measure is equal to or greater than a decided threshold value , the new web document is assigned to that particular cluster.

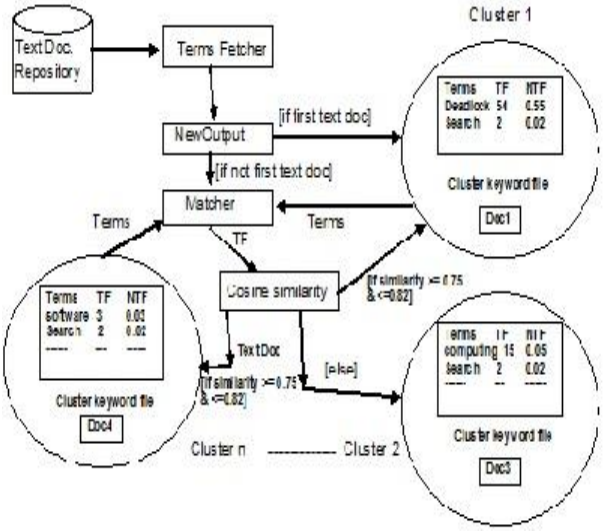
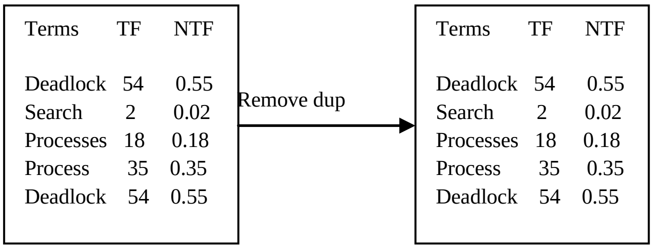


Figure 5: Domain Specific keyword based automatic web document clustering

3. DUPLICITY REMOVAL

Remove the duplicate terms from the cluster keyword file and also checks the existence of clusters when first document comes and if the no new cluster exists then create a new cluster and assign the document to the new cluster.



Terms	TF	NTF
Deadlock	54	0.55
Search	2	0.02
Processes	18	0.18
Process	35	0.35
Deadlock	54	0.55

Terms	TF	NTF
Deadlock	54	0.55
Search	2	0.02
Processes	18	0.18
Process	35	0.35
Deadlock	54	0.55

Figure 6 : New Output

3.1 Matcher

Extracts the terms of the terms of the cluster keyword file one by one and send to the module “similarity measure” for similarity measuring between the documents.

3.2 Similarity Measure

If more than 40% of the terms of the new document matches with the terms of a specific cluster keyword file then the document assign to that particular cluster.

3.3 Selecting the Title for a Cluster

The decision for assigning the title for a given cluster is suggested by calculating the frequency of the cluster keyword term in the documents of the cluster, the term which has the highest summation of frequency among the documents of the cluster is treated as the title of the cluster. The summation of the frequency of the term of the cluster keyword file is calculated as –

$\sum$ j=0 m $\sum$ i=0 k TF-------- a

Where m is the number of documents in a specific cluster and k is the number of cluster keyword terms.

The fig. below shows the calculation of the summation of the frequency of the term of the cluster keyword file-

Terms/ Doc id	Doc 1	Doc 2	Doc 3
Process	63	18	19
Computing	16	4	0
Wikipedia	11	0	1
Encyclopedia	2	0	0

Figure 6: frequency of the terms

The table given above shows the terms occurring in the documents of a particular cluster, the term “process” has the highest summation of frequency in the different documents, so, the title for a given cluster is suggested as “process”. The title of the given cluster is saved in a file named “title”. The terms of the query is matched with the title of the cluster if it matches the documents of the cluster is send as the result of the search query.

4. CONCLUSION AND FUTURE WORK

Currently this technique is working on few domains specific; by parallel processing it can be implemented on many domains simultaneously. It can be implemented on hierarchal based clustering, firstly main topic classification then sub topic classification. It can improve its efficiency.

5. REFERENCES

[1] McCallum, Andrew Kachites. ”MALLET: A Machine Learning for Language Toolkit.” http://mallet.cs.umass.edu. 2002.

[2] .S. Chakrabarti, B. Dom. R. Agrawal, P.Raghavan. Using taxonomy, discriminants and signatures for navigating in text databases, VLDB Conference, 1997.

[3] B. Liu, L. Zhang. A Survey of Opinion Mining and Sentiment Analysis. Book Chapter in Mining Text Data, Ed. C. Aggarwal, C. Zhai, Springer, 2011.

[4] M. Sahami, S. Dumais, D. Heckerman, E. Horvitz. A Bayesian approach to filtering junk e-mail. AAAI Workshop on Learning for Text Categorization. Tech. Rep. WS-98-05, AAAI Press. http://robotics.stanford.edu/users/sahami/papers.html.

[5] A. Y. Ng, M. I. Jordan. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. NIPS. pp. 841- 848, 2001.

[6] J. R. Quinlan, Induction of Decision Trees, Machine Learning, 1(1), pp 81–106, 1986.

[7] A. McCallum, K. Nigam. A Comparison of Event Models for Naïve Bayes Text Classification. AAAI Workshop on Learning for Text Categorization, 1998.

[8] C. Cortes, V. Vapnik. Support-vector networks. Machine Learning, 20: pp. 273– 297, 1995.

[9] Fabrizio Silvestri, Raffaele Perego and Salvatore Orlando. “Assigning Document Identifiers to Enhance Compressibility of Web Search Engines Indexes” In the proceedings of SAC, 2004.

[10] Van Rijsbergen C.J. “Information Retrieval” Butterworth 1979

[11] Oren Zamir and Oren Etzioni. “Web Document Clustering: A feasibility demonstration” In the pr oceedings of SIGIR, 1998.

[12] Jain and R. Dubes. “Algorithms for Clustering Data.” Prentice Hall, 1988

[13] Sanjiv K. Bhatia. “Adaptive K Means Clustering” American Association for. Artificial Intelligence, 2004

[14] Bhatia, S.K. and Deougan , J.S. 1998. “Conceptual Clustering in Information and Cybernetics.