 ISSN: 2454-4248
176 – 181

Non Overlapping Clustering based Meta Search Engine

Naresh Kumar[1], Sonali[2] and Shivani Gupta[2]
[1]Associate Professor, [2]B.Tech Students
Computer Science Department, MSIT, New Delhi, India
narsumsaini@gmail.com, sonali850099@gmail.com, shivani12011996@gmail.com

***Abstract*—**With each passing day, the volume of information on WWW is increasing enormously which makes very difficult for a user to search the necessary document. To manage and present available information in an effective manner, authors have proposed a Meta Search Engine based on clustering where terms are extracted from URL Tag, Title Tag and Meta Tag to cluster similar documents effectively. These parts of a webpage have been selected because they better define the features of a webpage. Final results are organized and presented in the form of clusters. The obtained results of implemented MSE have been compared with existing MSEs in terms of relevancy of results.

Keywords*-Meta Search Engine; Clustering; Webpage; Search Engine; Relevancy.*

*****

I. INTRODUCTION

Internet is overloaded with information and it is very difficult to search for the desired information that fulfills users' need [1]. Meta Search Engine (MSE) is a search tool that is designed to search information from WWW efficiently [2]. MSEs have become frequent and popular tools to search for any type of information [3]. The information retrieved by MSE may be useful but it may return large numbers of webpage (WP) which are difficult to approach by the users [4]. Due to large number of active users and huge size of WWW the WPs are updated very frequently. According to [5], 40% WPs of the entire web are changed daily. So the indexing of MSEs is also needed to be updated periodically which is a time and resource consuming process [6].

This paper has suggested and implemented a Non Overlapping Clustering based MSE (Noc based MSE) that uses the concept of clustering for grouping the results. The Noc based MSE uses results from multiple Search Engine(SE) that are Google and Bing and presents results to the user in the form of clusters. The Noc based MSE returns a list of labeled clusters containing URLs of the webpage. The main characteristics of Noc based MSE are removal of duplicate links from the search results and does not produce overlapping clusters. The files contain all the data that exists within respective WPs. WPs contain many fruitless terms or content which may reduce relevancy of results. WPs also contain tags information which is unproductive [9]. Also the WP is full of punctuation marks. So tags and punctuation marks must be removed from these WPs to obtain useful content. Tokenization is used to perform this. It insulates all the words, characters and numbers from a document and these insulated words, characters and numbers are known as tokens [13].

II. RELATED WORK

- Authors of [7] proposed a MSE that uses the theory of clustering and ranking so that user will get relevant results. The main modules of the system are user interface, relevancy calculator, cluster generator and webpage adjuster. This system considers top 10 results from different SEs (Google, Bing and Alta Vista) as most relevant and tested for 30 different queries. Their MSE shows results in the form of most relevant, partial relevant and least relevant clusters. Performance of this MSE is found to be bit improved than existing MSEs. Proposed MSE has loopholes in terms of time and space.

- In [8], large data sets are used along with applications of WEKA to generate sufficient clusters. Source code from different websites is used to collect data for processing. In-links and out-links of a website, URL length, title length and number of keywords are retrieved from collected source code. By using these parameters clusters are formed. Relevant results can be obtained if more data is retrieved from multiple websites. The proposed method produces results having > 60000 back-links, < 50 title length, > 3 title keyword, < 25 URL length and >200 in-links which are useful for optimization.

- Weblog extraction with fuzzy classification method that uses folksonomy and fuzzy clustering algorithm is proposed in [1]. Fuzzy clustering algorithm is an extension of traditional set theory. It is mainly designed for related relevant terms that are semantically related. Main modules of this system are

 ISSN: 2454-4248
176 – 181

interface, user query engine, MSE and aggregation of documents. Weblogs are used in this method. The results are shown using query OLED. If user enters the query OLED then system will return two folders to the user one having information of OLED and another having LED.

- Authors of [9] generate clusters by using k-means clustering. Here they extract URL Tag, Title Tag and Meta Tags. These attribute scan provide most of the information about a WP. They assume that by doing so there is no need to analyze the whole WP. After retrieval of these tags K-means clustering is used to organize the results. This method produces results with maximum inter-cluster distance and minimum intra-cluster distance.

- Intelligent Cluster Search Engine is proposed in [10] which uses meta directory tree for knowledge base and tree based search algorithm. This system also provides more relevant results because it uses different directories like Yahoo, ODP, Google etc. Semantic knowledge is not used for analysis of keywords. It uses meta directory trees which are maintained by humans and therefore they may not provide up-to-date information. Results show that it takes 507.54 ms average access time when it was tested for 20 different queries. It has low computation time because it uses meta directory tree.

- WISE is a search system proposed in [11] which uses content mining technique and hierarchical soft clustering for organization of their results. Concept and phrase are extracted for document processing. PoBOC soft clustering algorithm is used for organization of results into clusters. System focuses more on relevant documents only and discards lesser one. Here documents are represented semantically so that they can also be used for further analysis.

III. Problem Formulation

The following are the problems that exist in MSE and clustering techniques:

1) MSE uses different SEs to organize its result. But today, MSEs are unable to present the results to the user in an effective manner due to information overload on web [7].
2) Some online MSEs like yippy[12] produce clusters with labels. Labeling is done on the basis of highest frequency term that is contained in the documents of a cluster. But the label of the cluster may not satisfy the user query as there are many useless terms exist within a WP.
3) Some websites may take first position in MSEs by paying to that MSE but it is also possible that these websites are less relevant as compared to good ones that appear at the bottom of a SE.
4) In [9] semantic relation between the documents is not considered which may lead to the development of unrelated clusters.
5) The MSE proposed in [11] is that is allows a single document to be in more than one cluster. Therefore it gives a problem of overlapping clusters.

IV. Proposed Method

The Fig. 1 shows the Noc based MSE framework. It uses clustering technique to organize different WPs into clusters and present them to the user. Noc based MSE consists of the following modules:

1) Downloader
2) Content Extractor
3) Tag Extractor
4) Stop Word Remover
5) Stemmer
6) TF-IDF Calculator
7) Relevancy Calculator
8) Cluster Generator

 ISSN: 2454-4248
176 – 181

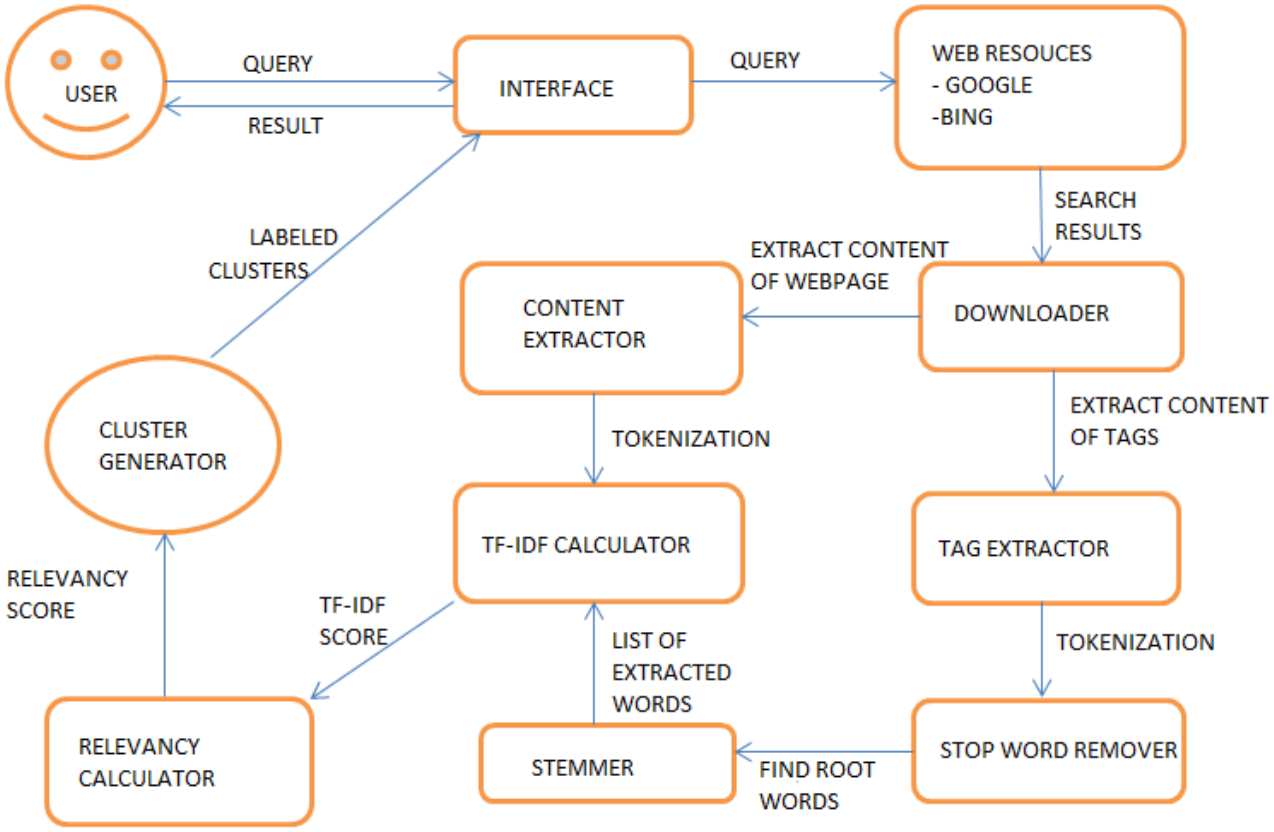

Figure1: Framework of proposed Noc based MSE

The following are the modules that are implemented in proposed framework:

1) Web Resources: It provides the list of SEs which are used for searching of information from WWW.

2) Downloader: This module downloads the search results.

3) Content Extractor: This module extracts the textual content of a WP and stores it in a file. This module performs tokenization. By using tokenization punctuation marks like comma “,”,full stop “.”,exclamation mark “!”, question mark “?”,semi-colon “;”,colon “:”,apostrophe “‘”,quotation marks “ “” ” , hyphen “-“ ,brackets “( )” or “[ ]” , slash “/” are removed from the document.

Eg: Sentence: Her son, John Jones Jr., was born on Dec. 6, 2008.
After removal of punctuation marks:
Her sons John Jr was born on Dec 6 2008

4) Tag extractor: From each WP, Title Tag, Meta Tag and URL Tag are retrieved and stored in a file. Title tag contains the title of the WP which describes the objective of that page. URL tag defines the URL of the site whereas meta tag is the fragment of the WP that provides information about the page content but meta tags themselves do not appear in the WP. All the useful information about the WP can be gathered by using these tags.

5) Stop Word Remover: This module is designed to abolish all the stop words and punctuation marks from the content that is obtained from URL tag, meta tag and title tag extractor. These tags may also contain stop words and punctuation marks which are not beneficial.

Eg: Sentence: John and Smith are good friends.
After removing stop words: John Smith good friends.
The following are the examples of stop words how, what, where, are, were, is, was and many more.

6) Stemmer: Stemming is a process of reducing crumple words to their word stem. The root word and stemmed word may or may not be same.

Eg: Before stemming: cares, cared, caring, careless, carefree, caretaker
After stemming: care for all the above words
In stemming, end part is truncated to form the correct or meaningful word. Sometimes it may not produce good results. Inspiring, inspirable will produce result inspir which has no meaning. Porter’s algorithm for stemming is used here. This algorithm works without lexicons and does not take into account the meaning of word therefore some words may be damaged [14].

 ISSN: 2454-4248
176 – 181

7) TF-IDF Calculator: This module calculates the significance of a term within a WP. The popularity of a word boosts as the appearance of word increases within a document. The relationship among number of WPs can be calculated by counting the number of times a term appears in a WP. Total occurrence of a term in a WP is known as term frequency. Many terms occur very frequently in a WP but they are of no use or may not provide the relation between different documents. So, to minimize their effect Inverse document frequency is used [9].

Tf(Term frequency) for a term in a WP is calculated using formula:

$$Tf_{ij} = \frac{TC_{ij}}{T_j} \qquad (1)$$

where i is term and j is document.

TCij is the frequency of the term i in the WPj and Tj is the count of all terms in the WPj.

Idf (Inverse document frequency.) is calculated using formula:

$$Idf_i = \log\left(\frac{D}{DT_i}\right) \qquad (2)$$

D is sum of j that is total number of WPs and DT is the number of WPs in which a term i occurs.
Weight to each term i in a document j is assigned using Tf-Idf

$$TfIdf_{ij} = Tf_{ij} \times Idf_i \qquad (3)$$

8) Relevancy Calculator: Relevancy score is used to determine the extent of match between user query and a webpage. Relevancy score for each WP is calculated using Relevancy Calculator which uses the output of TF-IDF Calculator for summing the TF-IDF values for each WP.

9) Cluster Generator: It is used to generate the required number of clusters that are entered by the user at the query time. Upper and lower relevancy scores are used to generate the clusters. According to ranges of clusters and relevancy score of WPs, WPs are assigned to these clusters. Further the clusters are sorted and returned to the user.

A. PSEUDOCODE:

Input: User query, number of required clusters (noc), number of results to be retrieved from each search engine.
Output: Labeled clusters with links of web pages in each generated cluster.

Step 1. [Downloading]
Download webpages of each search engine and remove duplicate links.

Step 2. [Webpage Content Extractor]
For each webpage
2.1 Extract text contents by performing tokenization.
End for

Step 3. [Tag content extractor]
For each webpage
3.1 Extract url tag, title tag and meta tag contents and store them in a file D.
End for

Step 4. [Stop word remover and stemmer]
For each term $term_i$ in D
4.1 Remove duplicate words
4.2 Remove Stop Words
4.3 Perform stemming using Porter's Stemming Algorithm
End for

Step 5. [TF-IDF calculator]
For all terms $term_i$ in D
For all webpages j
$TF\text{-}IDF_{ij} = TF_{ij} * IDF_i$
End for
End for

Step 6. [Relevancy Calculator]
For each webpage j
For all terms $terms_i$ in D
$TF\text{-}IDF_j = TF\text{-}IDF_{j+} TF\text{-}IDF_{ij}$
/*Relevancy Score of each webpage*/
End for
End for

Step 7. [Record lower and upper relevancy_score among all webpages]
min = min(relevancy_score)
max = max(relevancy_score)

Step 8.[Generation of required number of clusters]
a = (max-min)/noc
b = min

Step 9.[Deciding range of each cluster]

class clusters{
1ow
up
}cluster [noc]

Step 10.[Assignment of ranges to each cluster]
For k = 1 to noc
cluster[k].low = b
cluster[k].up = b+ a
b = cluster[k].up
End for

 ISSN: 2454-4248
176 – 181

Step 11.[Assignment of WPs to each cluster according to relevancy_score]

```
For all webpages j
    For each cluster[k] of noc
        if((relevancy_score(j) >= cluster[k].low) && (relevancy_score(j) <= cluster[k].up))
        setnoc[k] = j
        End for
End for
```

Step 12. Return the labeled clusters

B. EXPERIMENTAL RESULTS

To test the Noc based MSE, authors have used two search engines – Google and Bing. Noc based MSE has been implemented in Java using Netbeans IDE 8.0.2 (on Windows 10 platform). Oracle 11g database has been used to store the TF-IDF values and relevancy score of each downloaded WP. The Noc based MSE has been analyzed with 13 different queries taken from different domains as shown in Table I. In Table I, the first column specifies the domains of the queries, the second column specifies the queries related to a particular domain and the third column specifies the percentage of duplicate links removed from the results. Duplicate links are removed from the list at the time of retrieval of WPs. For each given query, top 10 links are retrieved from each SE. Terms from URL, Title Tag and Meta Tag are extracted from each WP and stored in a document. From this list of extracted terms, stop words and duplicate words are removed and stemming is performed. Then, relevancy score of each WP is calculated using the concept of TF-IDF and clusters are generated.

Table I. Queries taken from different domains and percentage of duplicate links removed for each query

Domains	**Queries**	**Percentage of Duplicate links Removed**
Technical	Java	20%
	Python	30%
	C	30%
	Antivirus	25%
Fruits	Orange	40%
	Peach	40%
	Papaya	50%
Mixed	Bottle	55%
	Katrina	25%
	Mouse	50%
	Delhi	45%
	Toy	50%
	Bike	35%

Table II shows the comparison between Noc based MSE and existing MSEs. The main feature of Noc based MSE is that it does not produce overlapping clusters therefore reducing poor relevancy.

Table II. Comparison between NOC based MSE and different MSEs

MSE	No. of SEs Used	Clustering	Relevancy of Results	Main Features
NOC based MSE	2 (Google and Bing)	Yes	High	Non-Overlapping Clusters
Meta Crawler [15]	3 (Google,Yahoo,Bing, and others Ask.com, About.com, MIVA)	No	Moderate	Searching of image, video, news, business, personal, telephone directory, audio
Web Crawler [16]	Uses WWW	No	Moderate	Provides full text search
ixQuick [17]	14-various	No	High	Searching in 17 languages
Apocalx [18]	3	No	Low	NA
Qksearch [19]	NA	Yes	NA	Blend search and split search
OpenText [20]	4 (Google,Yahoo, Bing, Ask, Wikipedia and Open Directory)	Yes	NA	For artificial intelligence
Gnome [21]	NA	Yes	NA	NA
iBoogie [22]	uses all the web and MSN	Yes	NA	Customizable search type tabs

V. CONCLUSION

This paper has introduced an Noc based MSE to present search results from Google and Bing using document clustering technique. Java JDK 1.8 is used to implement Noc based MSE. Noc based MSE has been tested for queries taken from different domains. The clustering results are purely based on text that exists within the tags of WP. Relevancy score produced by Noc based MSE is short because it uses tags of a WP therefore there is no need to analyze the whole WP. The only loophole of Noc based MSE is that it takes little more time to process the user query when compared to existing

MSEs but experimental results show that it produces better non – overlapping cluster and removes duplicate links from search results which is a major problem with other existing MSEs.

REFERENCES

[1] E. Portmann, "Weblog Extraction with Fuzzy Classification Methods," Second International Conference on the Applications of Digital Information and Web Technologies, pp.411-416, ISSN: 978-1-4244-4457-1, 2009.

[2] N. A. Kadir, A. M. Lokman and A. Ahmad, "Dominant User Context (DUC) Filtering Framework for Web Personalized Search," IEEE Symposium on Wireless Technology and Applications, pp.280-285, ISSN: 978-1-4673-2210-2, 2012.

[3] H. J. Li and J. K.Wang," Precise Image Retrieval on the Web with a Clustering and Results Optimization," International Conference on Wavelet Analysis and Pattern Recognition, Beijing, Vol. 1, pp.188-193, ISSN: 1-4244-1066-5, 2007.

[4] Peng Jiang, Chunxia Zhang, GuisuoGuo, ZhensdongNiu and DongpingGao, "A K-means Approach Based on Concept Hierarchical Tree for Search Results Clustering," Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009, pp.380-386, DOI: DOI 10.1109/FSKD.2009.658.

[5] Naresh Kumar and RajenderNath, "A Novel Parallel Domain Focused Crawler for Reduction in Load on the Network," International Journal Of Computational Engineering Research, Vol. 2, issue 7, pp.77-84, ISSN: 2250-3005(online), 2012.

[6] Y. Shen and D. Lee, "A Meta-search Method Reinforced by Cluster Descriptors," Web Information Systems Engineering, Second International Conference, Vol. 1, ISSN: 0-7695-1393-X, 2002, , pp.125-132, DOI:10.1109/WISE.2001.996473.

[7] N. Kumar and R. Nath, "A Meta Search Engine Approach for Organizing Web Search Results using Ranking and Clustering," International Journal of Computer, Vol. 10, issue 1, pp.1-7, ISSN: 2307-4531, 2013.

[8] M. Jindal and N. Kharb, "K-means Clustering Technique on Search Engine Dataset using Data Mining Tool," International Journal of Information and Computation Technology, Vol. 3, issue 6, pp.505-510, ISSN: 0974-2239, 2013.

[9] S. Poomagal and Dr. T. Hamsapriya," K-means for Search Results clustering using URL and Tag contents," International Conference on Process Automation, Control and Computing, pp.1-7, ISSN: 978-1-61284-764-1, 2011.

[10] C.W. Tsai, K. Huang, M. C. Chiang, and C. S. Yang, "A Fast Tree-Based Search rAlgorithm for Cluster Search Engine," Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics San Antonio, TX, USA, pp.1603-1608, ISSN: 978-1-4244-2794-9, 2009.

[11] R. Campos, G. Dias and C. Nunes, "WISE: Hierarchical Soft Clustering of Web Page Search Results based on Web Content Mining Techniques," Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, 2006, pp.301-304, ISSN: 0-7695-2747-7, DOI: 10.1109/WI.2006.201.

[12] The yippy website [Online]. Available: http://www.yippy.com/at 4:55PM (IST) on 8/5/2017.

[13] Vikram Singh and BalwinderSaini, "An Effective Tokenization Algorithm for Informaiton Retrieval Systems," Computer Science & Information Technology (CS & IT-CSCP), 2014, pp.109–119, DOI : 10.5121/csit.2014.4910.

[14] Atharva Joshi, Nidhin Thomas and MeghaDabhade, "Modified Porter Stemming Algorithm," International Journal of Computer Science and Information Technologies, Vol. 7 (1), pp.266-269, ISSN: 0975-9646, 2016.

[15] The metacrawler website [Online]. Available: http://www.metacrawler.com/at 4:56PM0(IST) on 8/5/2017.

[16] The webcrawler website [Online]. Available: http://www.webcrawler.com/ at 4:56PM (IST) on 8/5/2017.

[17] The ixquick website [Online]. Available: https://www.ixquick.com/at 4:58PM (IST) on 8/5/2017.

[18] The apocalx website [Online]. Available: http://www.apocalx.net/at 4:57PM (IST) on 8/5/2017.

[19] The qksearch website [Online]. Available: http://qksearch.com/at 4:58PM (IST) on 8/5/2017.

[20] The opentext website [Online]. Available: http://www.opentext.com/at 4:58PM (IST) on 8/5/2017.

[21] The Gnome website [Online]. Available: https://www.gnome.org/at 4:59PM (IST) on 8/5/2017.

[22] The iBoogie website [Online]. Available: http://www.iboogie.com/at 4:57PM (IST) on 8/5/2017.