ADAPTIVE FOCUSED CRAWLING STRATEGY FOR MAXIMISING THE RELEVANCE

Sagar Pise[1], Abhijit Kulkarni[2], Mangesh Udawant [3], Ganesh Shinde[4]

Guided by : **Prof. S. R. Durugkar**

[1,2,3,4]SND College of Engineering & Research Centre, Yeola, (Nashik) India
[1]Pise.sagar1@gmail.com

Abstract

Identify the request of interest of user on world wide web, is challenging for, crawlers or spiders or boats so to effectively manage present the interest of user with the help of most relevant information retrieval we proposed system "Adaptive Focused Crawling Strategy for Maximising the Relevance" focused crawler focuses on the similar web technology to analysed semantic results and documents. with the help of lingo clustering algorithm ,In which system we represent the lingo as search result clustering algorithm which is based on the good quality of clusters . In the clustering the different kind of data are combining by considering the similarity among this data. The clustering search engine uses lingo algorithm to cluster the document using snippet. The conclusion with respective this document is made in final section.

Key Words: *Focused Crawler, Lingo Algorithm, Snippet, Clustering.*

---***---

1. INTRODUCTION

In today's era of world the use of internet has increasing rapidly .the internet has collection of huge amount of information, the size of data is large .through the search engine user can find out there queries or related information. search engine can retrieve the user queries from the internet ,for that queries enormous amount of search results are retrieve ,and this should be time consuming task .user sometime tired to finding the relevant data from searched queries. Also all resulting webpages are not visited by client, like if user mainly visited the first 5 to 10 pages out of 100 or more pages which is provided by search engine .in that results sometime other pages are unrelated to his expected results. for the possible solution to above problem is overcome by clustering search technique .clustering is the technique in which grouping the different relevant data and create a cluster of relevant data, in short clustering technique can grouping the meaning full clusters Crawler are a tools for combining the web content locally ,

In this , focused crawler has been introduced for to fulfill the need of users it can maintained web portals or document collections locally, usually the requirements of user are need for high quality and up-to-date relevant result while reducing or minimizing the time space and bandwidth. focused crawler try to fetching number of pages related to that user queries as they can, while keeping the number of unrelated pages downloaded to a minimum. crawlers are given a seed pages as there inputs .identify the outgoing links appearing in the seed pages and visit only those links which are gives relevant result. Crawler can continuously visiting the web pages until a required no of results has been retrieved of local resources. Crawlers used by general search engine and accessing the lot of web pages those are unrelated to topics .and focused crawler combining the content of accessed web pages and links of the web for more visiting pages with higher possibility of related to a given topics.

In the data mining process main task is to fetch the only required or important information from web server. In this data is extracted from web server and manages clusters according to different pattern like using snippet the clusters are build. We can apply lingo algorithm to implement clusters by identifying synonyms in snippet that has improve the quality of cluster. Lingo algorithm is done using synonymity with inclusion of extended worldnet with increase quality of cluster labels and these documents. Identifying groups of related data queries helps user to search efficient data over internet

2. LITERATURE SURVEY

With the development of internet in world it is very important that search engine provides engine provides the efficient result. So that, there time is saved on searching required information. The result of search query is may be scattered so it is quite time consuming to get result. So it is very important that crawler is provide better result that only focuses on the only that content that are relevant of user query. Also data are scattered overall so it is very difficult for user to get exact result.

The focused crawling uses semantic web technology to analyses the semantic of link and web contents. Crawler fetches web contents based on ontological concepts with organizing contents of web, categorized it [1]. Clustering can be form by using semantic similarity between web contents. In this the snippet are used to making quality of clusters. It uses lingo algorithm to make efficient clustering of web documents in less time [2]. It uses top down clustering techniques to build clusters using lingo algorithm. The top down clustering means the first build bigger cluster and does fine grained clustering on these clusters [3]. The goal of search results clustering should be good to attract user more quickly, this focuses on supplying users with meaningful and unambiguous clustering labels [4]. It consists of the effective clusters designing for data retrieval in data mining. Data are effectively retrieved from last database using clustering techniques [5].

3. PROPOSED SYSTEM

The proposed method is focuses on the web search result clustering. It consist of the crawling techniques that only focuses on the organizing and categorizing web document or filtering irrelevant webpage's with regards to the users query. The lingo algorithm is used in order make clusters and assigns meaningful labels to that clusters. The lingo algorithm uses semantic similarity calculation to produce clusters. The snippet is used to in order to make quality of clusters and meaningful labels to that clusters.

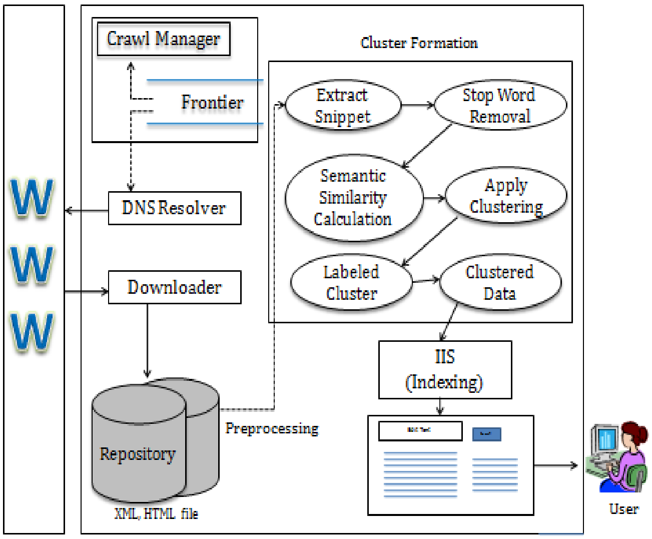

Fig -1: Architecture

The formation of the cluster of web result is consists of different phase:

1. Term extraction:
   In the key that user want to search are extracted.
2. Stop word removal:
   In this the searches for entire document for stop word and consider it for removal.
3. Similarity calculation:
   In this the check the similarity threshold of information and similar words are recognize with unique seed webpage.
4. Applying clustering algorithm:
   In this the clusters are builds and cluster labels are decided.

In this proposed method, the clusters name are directly assigned by using the pre-processing on that clustered data using term extraction and Stop word removal. The semantic similarity threshold is used to search the similarity of the user search key and similar word are recognised with sole seed key word which recognised after key extraction and stop word removal. The right and left phrases of the query are discover and combined into set of complete phrases. Only those phrases that are exceed the frequency threshold that are chosen and added into the corresponding clusters. The term or phrases document matrix are builds that contains term that not mark as a stop word and then matching that term with key and added into clusters are form with a provisional or restrictive parameter. In the cluster label identification, the lingo algorithm identifies the abstract concepts and then phrase matching and label pruning is done. The SVD decomposition of the resultant cluster word is used to match the given phrase in phrase matrix. The algorithm assigns the snippet to the clusters labels with sample contents.

4. ALGORITHM USED

In proposed system the lingo algorithm are used to made the cluster. The lingo algorithm consists of the following steps in order to make cluster.

Step 1: Preprocessing
In this the segmentation are perform on input snippet. These searches whole documentation for recognizing stop word and spot it for removal.

Step 2: Similarity calculation
In the step check the similarity among key terms with search keywords with a similarity threshold and similar words are recognized with single root keyword.

Step 3: Phrase extraction:
In this all document are collected and find out the complete phrase from document and it check it the threshold value. If this threshold value is above the given keyword then only it only those are added into the corresponding cluster.

Step 4: Cluster label Assignment:
In This decomposition technique are used. Here resulting cluster word are used to match given phrase keyword with the phrase keyword in matrix and assign the label to cluster.

Step 5: Cluster content discovery:
In this the content into cluster are added according to label assign. Cluster is form with the provisional or restrictive parameter

5. ADVANTAGES

1. In Proposed system the cluster is form by taking account of snippet rather than the complete documents.
2. Also it decreases the time complexity of searching the keyword.
3. The system also avoids duplicate copies WebPages.

6. CONCLUSIONS

The proposed system mainly focuses on deliver the exact relevant data to the user that he wants in a short span of time. The proposed system find out the web document and make the cluster. Cluster is design by taking view of the snippet except the entire documents. The cluster is form within a less time. This clustering reduced the time required for user to searches relevant information from the web. The existing system uses whole keywords to search the information that takes more time. The system provides the user cluster that provides user simplicity to get exact information. In this system can be added pre-indexing of WebPages in a local cache which later on decreases the time to search the keyword. Also the system eliminates the duplicate webpages to improved the redundancy of clustering.

ACKNOWLEDGEMENT

We greatly thankful and gratitude to respected guide Prof. S. R. Durugkar in Department of Computer Engineering of SND college of Engineering, Yeola (Nashik), for his support for giving important technical advice and encourages us and motivation.

REFERENCES

[1]. Hai Dong, Farookh Khadeer Hussain, Elizabeth Change “A Surver in Web Technology-Inspired Focused Crawler” IEEE 2008.

[2]. M. Manikantan , S. Duraisamy “Efficient Clustering of Web Search Results Using Enhanced Lingo Algorithm” Research Journal of Applied Sciences, Engineering and Technology 9(5): 359-364, 2015.

[3]. Poonam C. Falat, Prof. S. S. Sikchi “Lingo An approach For Clustering” International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 3, May – 2012.

[4]. P Ajitha, Dr. G. Gunasekaran “Effective Feature Extraction for Document Clustering to Enhanced the Search engine using XML” Journal of Theoretical and Applied Information Technology, 10th October 2014, vol.68 No.1

[5]. Anoop Jain, Aruna Bajpai, Manish Kumar Rohila “ Efficient Clustering Technique for Information Retrieval in Data Mining”, International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Volume 2, Issue 6, June 2012.

[6]. Ricardo, B.Y., H. Carlos and M. Marcelo, 2007. Improving search engines by query clustering. J. Am. Soc. Inform. Sci. Technol., 58(12): 1793-1804.

[7]. Beeferman, D. and A. Berger, 2000. Agglomerative clustering of a search engine query log. Proceeding of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston, pp: 407-416.

[8]. Hongwei, Y., 2010. A document clustering algorithm for web search engine retrieval system. Proceeding of the International Conference on e-Education, e-Business, e-Management and e-Learning, pp: 383-386.

[9]. Minky, J. and K. Nisha, 2013. K-means clustering technique on search engine dataset using data mining tool. Int. J. Inform. Comput. Technol., 3(6): 505-510.

[10]. Wang, J. and Z.Z. OuYang, 2010. The research of K-means clustering algorithm based on association rules. Proceeding of the International Conference on Challenges in Environmental Science and Computer Engineering, pp: 285-286.

BIOGRAPHIES

Mr. Sagar Pise: Student of BE in Computer Engineering in SND Collage of Engineering and research Center, Babulgaon Yeola, Dist. Nashik, india Email:Pise.sagar1@gmail.com

Mr. Abhijit Kulkarni: Student of BE in Computer Engineering in SND Collage of Engineering and research Center, Babulgaon Yeola, Dist. Nashik, india email:Abhijitkulkarni1993@gmail.com

Mr. Mangesh Udawant: Student of BE in Computer Engineering in SND Collage of Engineering and research Center, Babulgaon Yeola, Dist. Nashik, india Email:Mangeshudawant13793@gmail.com

Mr.Ganesh Shinde: Student of BE in Computer Engineering in SND Collage of Engineering and research Center, Babulgaon Yeola, Dist. Nashik, india Email: ganeshraj.shinde@gmail.com