Automated Text Abstraction from Documents and Webpages Metadata using Probabilistic Clusteringalgorithms

Silja Joy[1] , Guided By Asst Prof. Nisha J R[2],
Department of Computer Engineering,
Marian Engineering College,
Trivandrum, Kerala, India
[1]*silu.joy@gmail.com*
[2]*nisha.jr@gmail.com*

Abstract - Annotations are comments, notes, explanations, tagsor other types of external remarks. Annotation can be added to a text document or few portions of document or to a webpage. Annotation helps effective information retrievals. Webpage metadata is the data related with website, it is machine understandable information about web resources or other tags.Collaborative annotations are based on user created tags to annotate new objects. These tags are related user created labels for entities and allows user to organize and index the contents. Tagging is the act of adding keywords to objects. There have been significant amount of work to be performed in coming up with the tags for text documents or other resources like webpages, images and videos. Automated Annotation System (AAS) which uses algorithms like K-Means and Distributed Hash Table (DHT) to automatically create the attribute or annotation from documents or metadata of webpages. This proposed annotation technique provides the processing of metadata and/or text to efficiently come up with annotations rather than manually understanding the metadata or analyzing the text.

Keywords *-Annotation, Metadata, AAS, Information Extraction, K-Means, DHT*

*****

I.INTRODUCTION

Summarized output on searching particular document is a prime requirement nowadays. To arrive at summarized search output, documents / data has to be maintained in smart way during processing. Annotation technique is the leadingindustry standard best featured technique to manage documents and get an effective search result. Efforts to keep such decent maintenance of such annotate documents user has spent significant amount of efforts. A scenario is cumbersome, complicated and tedious where there are large amount of fields data to be entered at the time of uploading document. Such difficulties finally tend to very basic annotations, if at all, that there are often limited to simple keywords. Such simple annotation makes the analysis and querying of the data cumbersome. This motivated us to work on Automated Annotation System (AAS), which is a framework that facilitates creating automated annotation [9] from text or webpages. The aim is to create annotation [5] in webpage and documents. For faster and quick searching of results from documents/webpages, there are algorithms used for processing the data [1]. The algorithms are K-Means and Distributed Hash Table. This helps for clustering the documents based on the content present in it. The comparison is done against only for relevant clusters applicable hence time is saved. Here annotation in both webpage and documents is done by creating a summarized view. The contribution of our system is the direct use of checking the content of document/webpage. AAS provides cost effective and good solution to help efficient search results. The goal of AAS is to support a process that creates nicely annotated documents/webpages that can be immediately useful for summarization of end users.

II.LITERATURE SURVEY

In this section, similar studies in the areas of annotation processing are used to conduct literature survey are listed;

*P. Heymann, D. Ramage, and H. Garcia-Molina*published a paper *"Social Tag Prediction"*. [7]

This paper talks about predicting tags for social media content. Concepts outlined in this paper are used as a general guideline in our proposed system this lacks the processing logic for document and webpage metadata annotation.

Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W.-C. Lee and C.L. Giles came up with a paper on*"Real-Time Automatic Tag Recommendation"*. [3]

The paper outlines the method for automatic tag recommendation based on algorithms, a similar concept and logic in followed in our proposed system.

M. Miah, G. Das, V. Hristidis, and H. Mannila describes in their paper on*"Standing out in a Crowd: Selecting Attributes for Maximum Visibility"*.[5]

This research explains the idea of extracting algorithm using an Integer Programming formulation of the issue in hand. Thought the process takes huge amount of duration for processing a small amount of workload but comes out with an optimal and solution closes to actual.

III.ARCHITECTURE

Architecture of the proposed system is outline in the below figure. The host address will provide the metadata. By collecting the website address or the web link the metadata is accessed from the website. Metadata may not be readable to human being. By using the domain name and access view source system can read the metadata. For some cases metadata is not accessible. If accessible, apply *Stopword* algorithm for filtering the unwanted words. It removes special characters and digits present in metadata. Then apply stemming to get the corresponding attributes. Applying the probabilistic clustering algorithm to find out the probable attribute. The most probable attributes is saved to the database which will help to efficient search in future. This attributes will explain the underlying information in the metadata. These attributes are stored as annotated values for the metadata. Another provision is also there to search with those attributes. Searching for the attributes which are present in the metadata. Index of each attribute is used for searching. Frequency of each term is calculated. Hence we can understand the user preferences by checking the index. Processing of documents can be done here. The system automatically finds out the most frequently used attributes from the document. A sample dataset is used for this process. A folder has to upload first and then select a file for annotation processing. After selection, applying pre-processing. Then stemming algorithm is carried out to remove special characters and unwanted words. The next step is clustering for that an algorithm is used called Probabilistic Clustering for P2P (PCP2P). This approach will reduce the number of required comparisons by an order of magnitude. This technique helps to reduce the network traffic by reducing the number of required comparisons between documents and their respective clusters. Instead of identifying all clusters for comparison processing with each document, only a few most relevant ones are taken into account. Searching is effectively donewith the help of this algorithm. While searching for an attribute presented in the document the results are displayed thatcontain all the attributes from all the files. Also frequency of the particular attribute is also displayed

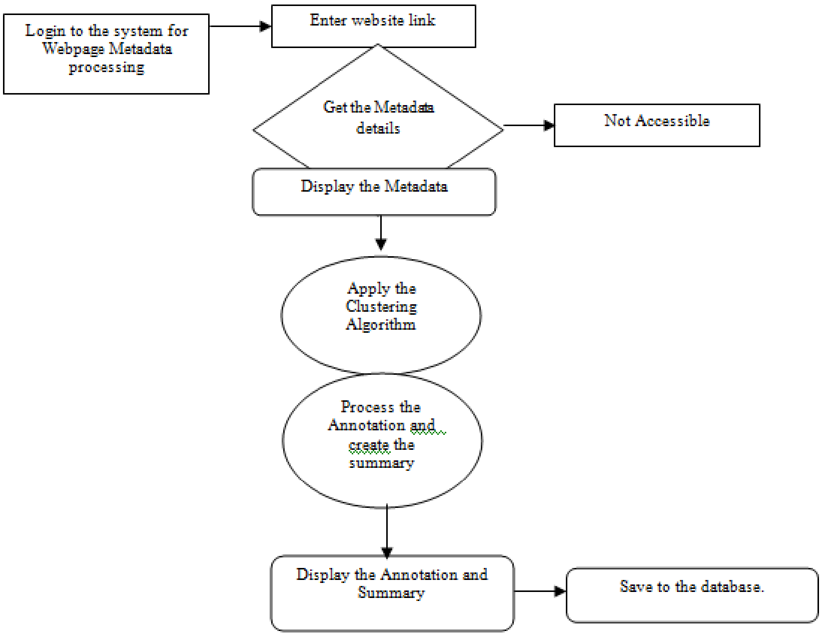
Login to the system for Webpage Metadata processing
Enter website link
Get the Metadata details
Not Accessible
Display the Metadata
Apply the Clustering Algorithm
Process the Annotation and create the summary
Display the Annotation and Summary
Save to the database.

Figure 1: Architecture for metadata processing

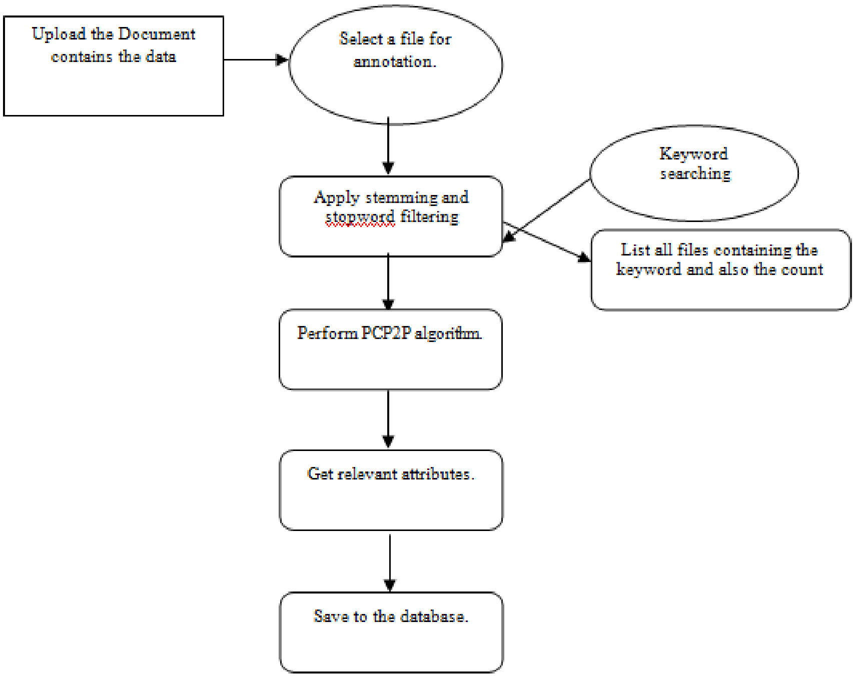
Upload the Document contains the data
Select a file for annotation.
Keyword searching
Apply stemming and stopword filtering
List all files containing the keyword and also the count
Perform PCP2P algorithm
Get relevant attributes.
Save to the database.

Figure 2: Architecture for document processing

IV.PROPOSED SYSTEM

The proposed system achieves effective searching within the text or webpage using K-Means and Distributed Hash Table (DTH) algorithms. In the Automated Annotation System (AAS), user has an option to choose between Webpage or Text annotation. The processing flow of the system can be explained in different modules below,

A. *Metadata processing Module:*

In this module the system uses webpage metadata as input. As mentioned earlier, webpage metadata is contains information related to the website under consideration. Typically metadata of a webpage is not easily understandable to humans. So the proposed system, takes metadata of webpage as input, process it, and create annotation about the webpage after summarizing the data extracted from metadata. Website address is keyed in to get the metadata of the corresponding website. During the annotation process, first stop words are eliminated from metadata and then apply stemmer algorithms. Repeated words in the metadata are identified. A brief description about the webpage is created, which helps users to get a summarized view of the webpage.

B. *Document processing Module:*

Identify the document which needs to be processed and upload to the document processing session of the system. Highly repeated words and contents from the document are identified as attribute information. Clustering of these information are analyzed, and assigned to existing cluster if not new cluster is created. Cluster algorithm helps in keyword searching faster by searching the keyword in all clusters instead of searching in all files. This makes the search much faster and efficient. For processing multiple documents, folder uploading is another feature available the system. By combining the output of clustered search words, the system creates a summary of the content present within the document.

C. *K-Means Algorithm*

K-Means clustering tries to partition n objects into k clusters in this way each object belongs to the cluster with the nearest mean. This method produces exactly k different clusters of greatest possible distinction. K-Means is an efficient method. However, it requires to specify the number of clusters, in advance as the final results are sensitive to initialization and often terminates at a local optimum. Unfortunately there is no standard method to find the optimal number of clusters in each scenario. A practical approach is to compare the outcomes of multiple runs with different k and choose the best one based on a predefined criterion. In general, a large k probably decreases the error but increases the risk of over fitting.

D. *Distributed Hash Table Algorithm*

DTH consists of two parallel task, cluster indexing and document assignment. Cluster indexing is performed by the cluster holders. The second activity, document assignment, consists of two steps, pre-selection and full comparison. In the pre-selection step, the peer holding d retrieves selected cluster summaries from the DHT index, to identify the most relevant clusters. Pre-selection already filters out most of the clusters. In the full comparison step, the peer computes similarity score estimates for d using the retrieved cluster summaries. Clusters with low similarity estimates are filtered out, and the document is sent to the few remaining cluster holders for full similarity computation. Finally, d is assigned to the cluster with the highest similarity. This two-stage filtering algorithm reduces the number of full comparisons.

V.RESULTS

The proposed system shows several benefits over existing systems. The system is validated with more than 100 webpages and more than 250 documents.

During testing, precision and recall are calculated to accurately test the system against webpages and documents. Both metadata processing and document processing are tested extensively to match the final annotation provides a meaningful information about the document or webpage used as an input.

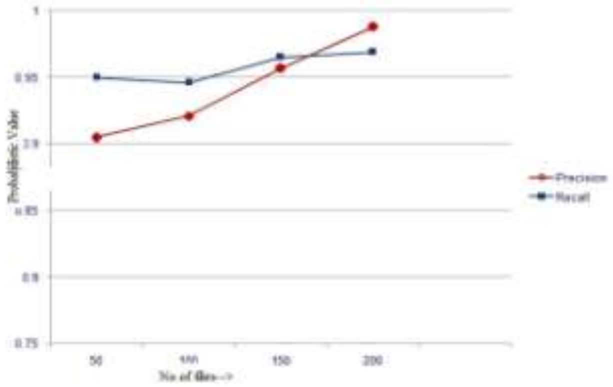

Figure 3: Precision and Recall graph while searching in documents

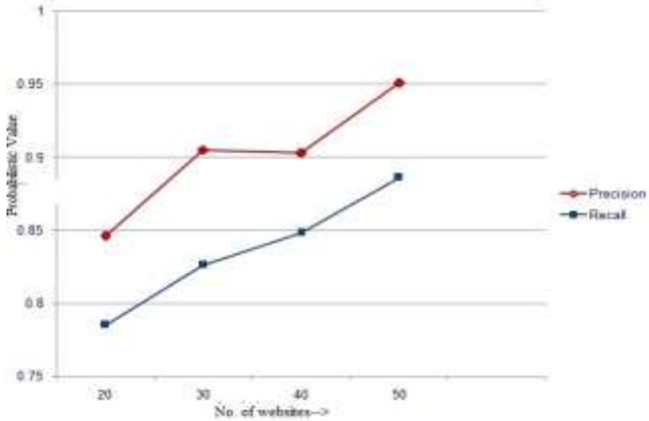

Figure 4: Precision and Recall graph for metadata processing

From the above two output charts, conclusion can be arrived that for document processing the accuracy is higher compared to metadata processing. This variation in the processing result is due to the fact that document processing uses clustering algorithms while metadata processing normal searching techniques. This directly results in increased accuracy for document processing as a result of efficient searching techniques produced by clustering algorithms.

VI.CONCLUSION AND FUTURE WORK

In this proposed system, automated annotation is done for webpages and documents using the system. For document processing, clustering algorithms are used to increase the efficiency of the searching. A summarized view the contents from documents and webpages are created by the system. The results of the system are analyzed for precision and recall by plotting charts against both metadata processing and document processing. There is an inverse relationship between the results of precision and recall. Conclusion can be arrived at by analyzing the results that document processing stands outs in accuracy compared to metadata processing, due the usage of clustering algorithms for efficient searching for document processing.

Future enhancements can be done in the direction of including image and multimedia to the annotation process along with annotation of webpages and documents. This will complete the 360 view of the annotation processing without limiting the medium of annotation processing.

VII.REFERENCES

[1] Eduardo J. Ruiz, Vagelis Hristidis, and Panagiotis G. Ipeirotis,“Facilitating Document AnnotationUsing Content and Querying Value”, IEEE TRANSACTIONS, VOL. 26, NO. 2, FEBRUARY 2014.

[2] M. Franklin, A. Halevy, and D. Maier, “From Databases to Data spaces: A New Abstraction for Information Management,”SIGMOD Rec, vol. 34, pp. http://doi.acm.org/10.1145/1107499.1107502, Dec. 2005.

[3] Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W.-C. Lee, and C.L. Giles, “Real-Time Automatic Tag Recommendation,” Proc. 31st Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’08), pp. 515-522, http://doi.acm.org/10.1145/ 1390334.1390423, 2008.

[4] M. J. Cafarella, J. Madhavan, and A. Halevy, “WebScale Extraction of Structured Data,” SIGMOD Record, vol. 37, pp. 55-61, Mar 2009.

[5] M. Miah, G. Das, V. Hristidis, and H. Mannila, “Standing out in a Crowd: Selecting Attributes for Maximum Visibility,” Proc. Int’l Conf. Data Eng. (ICDE), 2008.

[6] B. Sigurbjornsson and R. van Zwol, “Flickr Tag Recommendation Based on Collective Knowledge,” 17th Intl Conf.(WWW 08), pp.327336,http://doi.acm.org/10.1145/1367497.1367542 , 2008.

[7] P. Heymann, D. Ramage, and H. Garcia-Molina, “Social Tag Prediction,” Proc. 31st Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’08), pp. 531-538, http://doi.acm.org/10.1145/1390334.1390425, 2008.

[8] Vagelis Hristidis and Panagiotis G. Ipeirotis “CADS: A Collaborative Adaptive Data Sharing Platform”VLDB ’09, ACM.org.

[9] S.R. Jeffery, M.J. Franklin, and A.Y. Halevy, “Pay-as-You-Go User Feedback for Dataspace Systems,” Proc. ACM SIGMOD Int’l Conf. Management Data, 2008.

[10] Odysseas Papapetrou, Wolf Siberski, and Norbert Fuhr “Decentralized Probabilistic Text Clustering,” IEEE transactions on knowledge and data engineering, vol 24 NO.10, year 2012.