INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

AN EFFICIENT PROCESSING OF WEBPAGE METADATA AND DOCUMENTS USING ANNOTATION

Sabna N.S[1], Jayaleshmi S[2]

[1]M.Tech Scholar, Dept of CSE, LBSITW, Poojappura, Thiruvananthapuram
sabnans1988@gmail.com
[2]Associate Professor, Dept of CSE, LBSITW, Poojappura, Thiruvananthapuram
j.lekshmi.s@gmail.com

Abstract: *Annotations are useful in effective information retrieval. Annotation can be applied to several fields like image, videos, text, etc. Webpage metadata is the data related with website. Metadata for web pages contain description of the pages contents as well as keywords linked to the content. These are usually expressed in the form of metatags.. From the metadata, an user can't understand the full metadata. So here an adaptive insertion form was created which will display the most probable attributes while a user submits an url. The form will displayed only the main attribute values. Also in the adaptive insertion form, a brief description about the website is shown. These details were saved in the database. The user can use these annotations for searching purposes. The searching results will based on a ranking format. When a user submits a query regarding the annotated website, the details will be displayed including the url, description, and the main attributes. This annotation technique provides the processing of metadata in an efficient manner rather than manual understanding of the metadata. In addition to these, in these thesis documents are also processed using annotation.*

Keywords: *Annotation, Metadata, CADS, Information Extraction, PCP2P.*

1. INTRODUCTION

Data mining refers to huge collection of data. It is the process of discovering interesting patterns from large amounts of data. There are many sources for these data. They include databases, data ware houses, the web, other information repositories etc. We usually mine data when there is too much data and too little information. Also we mine data if there is a need to extract useful information from the data and to interpret the data. There are lot of tools for processing the data. Document annotation may be as old as writing on media. It becomes a prominent activity around 1000 AD in Talmudic commentaries and arabic rhetorics treaties. In the medieval era, scribes who copied manuscripts often made marginal annotations that then circulated with the community. As the popularity of printing press and individual copies of text increases, socially shared annotations declined and text annotation become more popular. Annotation is emerged as a different stream in data mining. It is defined as the task of adding metadata information in the document which is useful in information extraction. Information extraction (IE) is the process of automatically extracting structured information from unstructured documents. Recent activities in multimedia document processing like automatic annotation and content extraction could be seen as IE. Webpage-metadata is the data related with website. Annotation process is applied on that metadata for retrieving important attributes present on it. This helps user to know about the wanted attributes without going through the full metadata. Those metadata is not easily understandable for all users. This process is done by using the link of corresponding website. The annotation system also process documents based on the content present on the documents. Annotation in documents helps for efficient searching. The main attributes from the documents can be retrieved using annotation which is used for future searching purposes. Our aim is to create annotation in webpage and also in documents. For efficient searching results, here an algorithm is used for processing documents. The algorithm is Probabilistic Clustering for Peer to Peer (PCP2P). This helps for clustering the documents based on the content present in it. The comparison is done only for relevant clusters hence time is saved. Here annotation in both webpage and documents is done by suggesting relevant attributes. This paper is organized as follows. Section 2 includes the literature survey which is a summary of works related to this topic that were referred for the completion of this work. Section 3 describes the architecture of the proposed system and design details. Section 4 provides the proposed scheme and section 4 reports some test results.

2. LITERATURE SURVEY

Here several existing properties from previous studies in the field of annotation are described.

In 2005 M. Franklin, A. Halevy, D. Maier proposes data spaces and their support systems as a new agenda for data management. A Database Management System (DBMS) is a generic repository for the storage and querying of structured data. A DBMS offers a suite of interrelated services and guarantees that enables developers to focus on the specific challenges of their applications, rather than on the recurring challenges involved in managing and accessing large amounts of data consistently and efficiently [1].

In 2009 M.J. Cafarella, J. Madhavan, A. Halevy, [2] introduced three recent extraction systems that can be operated on the entire Web. The Text Runner system focuses on raw

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

natural language text, the Web Tables system focuses on HTML-embedded tables, and the deep-web surfacing system focuses on hidden databases. Here they describe the different characteristics of data produced by the three extractors. Also discussed a series of unique data applications that are enabled by aggregating extracted Web information.

In 2008 B. Sigurbjornsson and R.Van Zwol present and evaluate tag recommendation strategies to support the user in the photo annotation task by recommending a set of tags that can be added to the photo. In this paper they also investigated how users can be assisted in the tagging phase. Here the results are presented by means of a tag characterisation focussing on how users tags photos and what information is contained in the tagging [3].

In 2008 B. Russell, A. Torralba, K. Murphy, and W. Freeman [4] build a large collection of images with ground truth labels to be used for object detection and recognition research. Here they developed a web-based tool that allows easy image annotation and instant sharing of such annotations. LabelMe, is a database and an online annotation tool that allows the sharing of images and annotations. The online tool provides functionalities such as drawing polygons, querying images, and browsing the database. Here they described about the annotation tool and dataset and provide an evaluation of the quality of the labeling. Also presented a set of extensions and applications of the data set. Then a comparison is done between the LabelMe dataset against other existing datasets commonly used for object detection and recognition.

In 2009 Vagelis Hristidis, Panagiotis G. Ipeirotis presented a new platform called Collaborative Adaptive Data Sharing Platform (CADS) [5] which uses query workload to annotate the data at insertion-time. The main advantage of CADS is that it learns with time the most useful attributes and uses this knowledge to guide the data insertion and querying. The CADS system will found out the important pieces of data.

In 2014 Eduardo J. Ruiz, Vagelis Hristidis and , Panagiotis G. Ipeirotis [6] presented a novel alternative approach that facilitates the generation of the structured metadata by identifying documents that are likely to contain the useful information and this information is going to be more useful for querying the database. The main contribution of this paper is that here they present algorithms that identify structured attributes that are likely to appear within the document. This is done by jointly utilizing the content of the text and the query workload.

3. ARCHITECTURE

The Figure1 and Figure2 show the overall architecture of the proposed system. The system first get the metadata from the host address that copied by the user. The metadata is obtained online by the system. The system read the metadata from the host address using the domain name and access view source. Then the system will read the full contents and only access the metadata from it. Those metadata is not understandable to a normal user. After obtaining metadata, then the next step is to find out the most probable attributes from the metadata. For identifying those attributes, analysis of metadata have to be done. That is, first stop word method is used. Then stemmer method is used to filter the data. Stemmer algorithm is used to filter the words from the metadata. Also remove the unwanted characters and digits from the metadata using corresponding algorithm. User can also view the metadata description from the metadata. After finding the most probable attributes, next step is to store it in the database for future searching. These attributes are stored as annotated values for the metadata. Another provision is also there to search with those attributes. Searching is used to search the attributes which are present in the metadata. Searching is done by finding the index of each attributes. Frequency of each searched terms are also shown. Hence we can understand the user preferences by checking the index. Documents are also processed here. The system automatically finds out the most frequently used attributes from the document. For this process, CNET dataset is used. First upload the folder and then select a file for annotation to be done. After selecting the file, pre-processing is done. Then stemming algorithm is carried out to remove unwanted characters. Here an algorithm is also used for clustering called Probabilistic Clustering for P2P(PCP2P). This reduces the number of required comparisons by an order of magnitude. This approach also helps to reduce the network traffic to reduce the number of required comparisons between documents and clusters. Instead of considering all clusters for comparison with each document, only a few most relevant ones are taken into account. Searching is done effectively with the help of this algorithm.

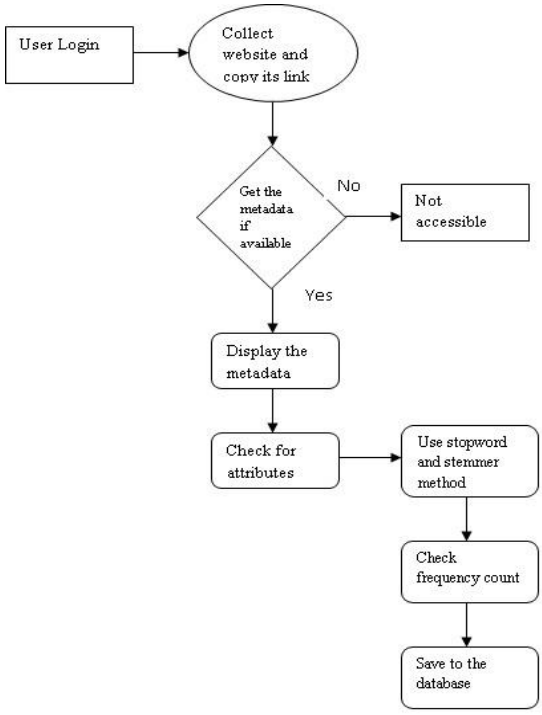

Figure 1: Architecture for metadata processing

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

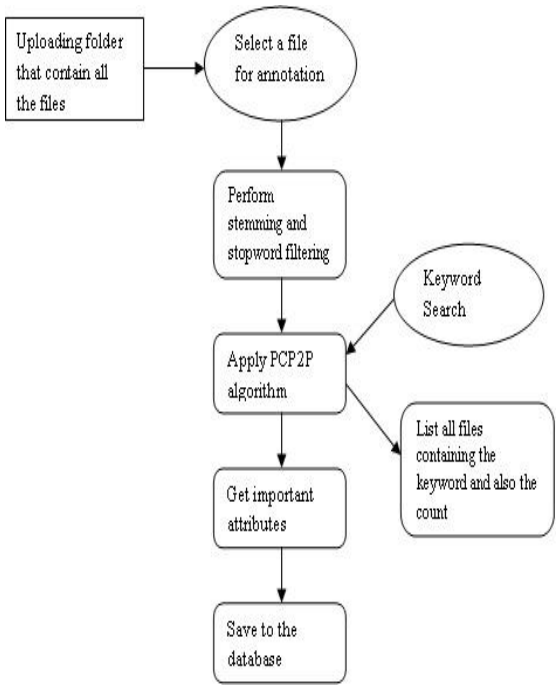

Figure 2: Architecture for document processing.

While searching for an attribute presented in the document the results are displayed that contain all the attributes from all the files. Also frequency of the particular attribute is also displayed.

4. PROPOSED SCHEME

The proposed scheme uses PCP2P [7] algorithm for the efficient searching process. In this annotation system, first the user login to the system and chooses webpage processing or document processing. The overall system is working can be explained as different modules.

In metadata processing, webpage metadata is the main focus. Webpage-metadata is the data related with website. Those metadata is not easily understandable to normal users. So we can apply annotation process here. Annotation is the process of extracting frequent terms from a document or any other files.

In this module, a website address is entered and gets metadata of that corresponding website. Then annotation process is carried out. For this pre-processing methods are done i.e, first stop words are removed from the metadata and then apply stemmer algorithms. Then the next step is to find the frequently appeared words in the metadata. Those words are displayed in the attributes field which are the relevant terms from the metadata. From those attributes, we get the important attributes present in the metadata. The attributes are retrieved with the help of system. Those attributes are saved for future searching purposes. While searching for a word, system will display all files that contain the particular word.

In this module a brief description is also shown from the metadata which will help users to understand the details of the website. From the metadata, normal users can't understand what is the content described in the particular metadata. This module helps to understand the metadata without reading the full meta data. Thus annotation helps in processing the website-metadata. The attributes from the metadata is saved as the annotated values for the particular metadata. While searching, these attributes are displayed.

For searching purposes, normal searching method is used here. The attributes retrieved from metadata is saved in to the database for future searching purposes. Those attributes are frequently used by users. Hence if we search those attributes, website information is displayed including number of times that particular keyword appears and about the description of particular website. For processing documents, a dataset is needed. Here dataset from CNET is used for processing. The dataset contains user reviews from several products from CNET. Those reviews were sometimes a large file. Hence user can't understand what is described on that document. Therefore annotation helps to find out most probable attributes from the document. Here an algorithm called PCP2P is worked out for efficient searching. This algorithm relies on a Distributed Hash Table (DHT) infrastructure. DHT offers efficient hash table capabilities in a P2P environment by arranging peers in the network. By applying this algorithm, searching will be fast and we get correct results. Select the documents for processing and upload the folder and also select a text file. Assign the file to a default cluster. Then process the file. High frequency values from this file are selected as attribute information. Then check whether this attribute information are in the cluster or not. If a cluster is there then assigns this files into that cluster. Otherwise create a new cluster and store into that cluster. Whether the searched information is found then get attribute information and check which cluster is assigned to this file. Then select that particular cluster.

Keyword searching is possible with the help of clustering algorithm. While searching for a particular keyword, clustering algorithm do searching in all clusters for the particular keyword instead of searching all files. This will helps to save time and results in efficient searching. Folder uploading is another provision in this system. For processing documents, CNET review dataset is used. The dataset contains many reviews of users about various products. Here dataset folder is uploaded directly for processing the documents. Another focus is towards searching process. For searching purposes, clustering algorithm is used. The algorithm is Probabilistic Clustering for peer to peer. This helps for e_cient searching as it reduces the time for comparing all documents. Instead of searching in all documents they will compare with particular clusters only. Searching results will return the files that contain the keyword and also the count. The results will base on a ranking format. After searching, users can also download the corresponding file. When searching is done, results will display as a list. If we want to view the particular file, we can download the file.

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

4.1 PCP2P Algorithm

In PCP2P, a peer undertakes up to three different roles. First, it serves as document holder, i.e., it takes care of clustering its documents. Second, it participates in the underlying DHT by holding part of the distributed index, and routing DHT lookup messages. Third, a peer may become a cluster holder, i.e., maintain the centroid and document assignments for one cluster. PCP2P consists of two parallel activities, 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.

5. RESULTS

The new scheme has several advantages over the existing systems. The system is tested with 200 documents. From testing, precision and recall was calculated. Thus accuracy was checked. precision and recall is calculated for two modules, i.e, for webpage metadata processing and also for document processing. For these two modules accuracy is also found out. The system tests for more than 200 documents and also for more than 100 websites. Datasets is taken from CNET dataset where more than 100 reviews are found. Those were the reviews of customer who uses electronic equipment from CNET. Those reviews are taken for processing. These were in (.txt) format. The graph was plotted based on searching a keyword.

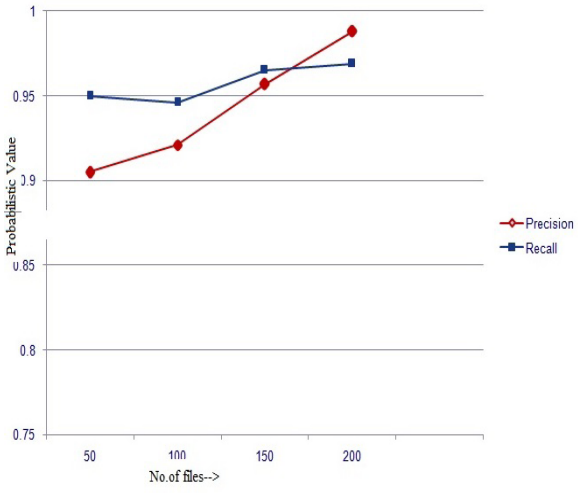

Figure 3: Precision and Recall graph while searching in documents

The graph no. 3 shows the precision and recall values for document processing.

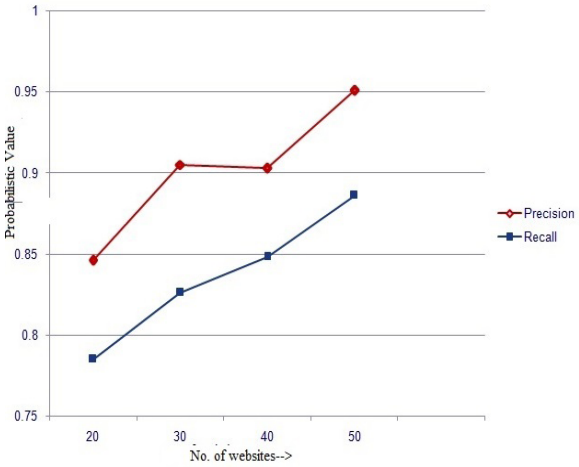

Figure 4: Precision and Recall graph for metadata processing

From the graph, we can infer that while searching in documents, recall is higher than precision and accuracy is more than metadata processing. The accuracy is higher for document processing is because of the presence of clustering algorithm. PCP2P clustering algorithm is used which is effective for searching. So accuracy is more than metadata processing. In metadata processing, normal searching techniques are used. And also a comparison graph is drawn based on the two algorithms.

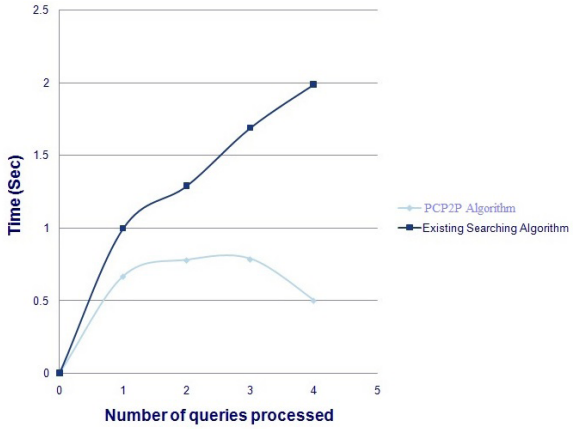

Figure 5: Comparison with normal searching algorithm and PCP2P algorithm

6. CONCLUSION AND FUTURE WORK

Here annotation is done for webpage-metadata and for documents. Here many website is taken for annotation process. And also CNET reviews for various products are taken as dataset for document processing. For document processing, PCP2P algorithm is used which is effective for searching purposes. Precision and recall graph is drawn based on the values retrieved by processing the datas. There is an inverse

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

relationship between precision and recall. In this thesis, annotation is done by suggesting relevant attributes. From the graph, it is seen that recall is higher than precision and it is inversely related. Also accuracy for document processing is more than metadata processing which is using PCP2P clustering algorithm. A comparison graph is also drawn based on the searching process used in metadata processing and document processing.

As a future enhancement, another algorithm can be used for metadata processing and also natural language processing can be employed in user review files to find out the nature of reviews.

References

[1] 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.

[2] M. J. Cafarella, J. Madhavan, and A. Halevy, "Web-Scale Extraction of Structured Data," SIGMOD Record, vol. 37, pp. 55-61, Mar 2009.

[3] B. Sigurbjornsson and R. van Zwol, "Flickr Tag Recommendation Based on Collective Knowledge," 17th Intl Conf.(WWW 08), pp.327-336,http://doi.acm.org/10.1145/1367497.1367542, 2008.

[4] B. Russell, A. Torralba, K. Murphy, and W. Freeman, "LabelMe: A Database and Web-Based Tool for Image Annotation," Intl J. Computer Vision,vol. 77, pp. 157-173, http://dx.doi.org/10.1007/s11263-007-0090-8, 2008, doi: 10.1007/s11263-007-0090-8.

[5] Vagelis Hristidis and Panagiotis G. Ipeirotis "CADS: A Collaborative Adaptive Data Sharing Platform"-VLDB '09, ACM.org.

[6] Eduardo J. Ruiz, Vagelis Hristidis and, Panagiotis G. Ipeirotis "Facilitating Document Annotation using Content and Querying Value" - IEEE transactions on knowledge and data engineering, 2014.

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