www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 7 July 2015, Page No. 13264-13267

A Review on Annotation Process

***Sabna N S*[1], *Jayalekshmi S*[2]**

[1] M.Tech CSE, LBSITW, Poojappura,
Thiruvananthapuram, Kearala, University, India
sabnans1988@gmail.com

[2] Associate Professor, Dept. of CSE, LBSITW, Poojappura
Thiruvananthapuram, Kerala University,India
j.lekshmi.s@gmail.com

Abstract: ***Annotations are useful in effective information retrieval. Annotation can be applied to several fields like image, videos, documents, etc. Annotation helps to understand and retrieve the documents very easily. For doing annotation, firstly important attributes have to be identified. Making annotations on documents is a hardwork because people have to read the documents fully to think which sentences have to be annotated. If people are automatically given the sentences which are most suitable for annotations, it is more easier. Also annotations can be done to webpage-metadata also. In this paper, a survey about annotations is discussed and a brief description about proposed methodology is given.***

Keywords: Annotation, Metadata, CADS, Information Extraction

1. Introduction

Data mining is the process of automatically searching huge amounts of data to discover useful patterns. The main goal of the data mining process is to extract wanted information from the large sets of data’s and those datas were transformed into useful formats for further use. Information extraction is the process of extracting information from a set of documents. For extracting information, annotation, content extraction and other multimedia document processing techniques are used**.** Annotations are used to understand a particular document easily. It provides the users the most suitable information without including unwanted information. Also annotation process will helps to increase the efficiency of searching. It will give accurate search results. Text annotations serves a variety of functions. Some of the functions are:-educational applications, social reading, writing and text-centered collaboration.

Annotations are used to find writers opinion in a document. To finding such information, annotation is one of the powerful method. To make annotation in documents is a hardwork because people had to read the documents carefully to check which part to be annotated. Automatic recommendation of sentences helps to shaving off time. Inaddition to automatic annotation in documents, image annotation is also an area regarding this. Automatic image annotation is the process by which the system automatically assigns metadata as keywords to a digital image. Many methods are there to provide automatic annotation. Automatic annotation helps user to save time and also make the documents in structured format. Webpage metadata is the data related with website. Metadata for website contain the description of webpage. Those are expressed in the form of metatags. Metadata in the webpages is in machine understandable form. Inorder to understand the metadata, it is very hard to read the full metadata. For this purpose, some techniques can be used. Annotation is applied to the metadata to retrieve main attributes form the metadata. This helps the user to understand the basic details of the website from the metadata.

Annotation can also be applied to documents. By doing so main attributes are retrieved from it and save to the database for future searching. This helps to improve the searching process.

Advantages of automatic annotation process are:-

- Speed.
- Less recommendation of annotation compared to manual annotation.

Here in this paper, section 2 gives a brief history of annotation, section 3 gives an overview of annotation process, section 4 gives small description of proposed system and section 5 gives the conclusion.

2. Background

Increasing use of computer databases result in the evolution of data mining. As information increased, new processing techniques also emerged. Annotation is also a process related to information processing and extraction. Document annotation is one of the old form like writing on media. In the medieval era, scribes who copied manuscripts made annotation and then circulated with the community. As the popularity of printing press increased socially shared annotation declined and text annotation emerged.

3. Literature Survey

In traditional systems, they do not have the “attribute-value” annotation. Because, annotation which uses attribute-value pairs require more careful in their annotations. Users must have a good idea about when and how to use those annotations. Users are unwilling to perform the attribute-value annotation even if the system allows. Such difficulties arise in basic annotations, which is limited to simple keywords. Those types of annotation make the analysis and querying poor. Users are often limited to simple keywords ie, basic annotation fields such as “creation date” and “size of document”.

Some of the existing methods on annotation is discussed below.

3.1 Collaborative Annotation [1]-[2]

There are many systems that use collaborative annotation of object based on user created tags. Tags are user created entities. Lot of works are done on predicting tags for documents,webpages,videos,images etc.

3.1.1 Tag behavior in Flickr[1]: Flickr describes about photo annotation. Flickr and Zoomr allows users to share their photos to their community through online. Using these services users can manually do their annotations using tags. These tags describe the content of the photo as well as give additional information about the photo. This shows that if same photo is annotated by different user different description is also provided. In flickr, users can find many photos in same name from different users with variety of tags. This photo annotation provides the user’s personal view about that photo. Here, author analyze two concepts:-how users tag photos and what type of tags they provide. Also describes about four different tag recommendation strategies. There are lots of photos tagged by the users. Relationship between those photos can be derived by the global co-occurrence metric.

$$J(t_i, t_j) := \frac{|t_i \cap t_j|}{|t_i \cup t_j|} \qquad (1)$$

Tag co-occurence is the key to tag recommendation approach. The quality of relationship between two tags cannot be measured using the raw tag co-occurrence as these do not take the frequency of individual tags. Normalisation measures are used to normalize the co-occurence of two tags. After the tag co-occurence, tag aggregation and promotion is done.

3.1.2 LabelMe[2]: LabelMe is a database and an online annotation tool. It allows sharing of images and annotation. This online tool provides some drawing functionalities. This paper describes about annotation tool and dataset and also provide evalutaion of the quality of the labeling. The goal of this annotation tool is to provide a drawing interface that will works on many platforms. It provides high quality labeling.If an user wants to label an image, select an image. Labeling the object is done by clicking the control points along the object boundary. Finishing point is same as starting point. After completing, a pop up dialog button will appear asking for object name. This label is saved in the Labelme database and is displayed on the corresponding image. The label is then available for download and viewable for all users. This annotation tool is simple and easy to use. One of the important concern is quality control. Another issues solved here are complexity of the polygons provided by the users and about what to label. The user can label any objects that is presented in the image. When they are satisfied with the labelling, they can proceed to next image for labelling. When a users enters the page, previously entered labels will appear on the image. There is also a provision for renaming the label if there is any error.

3.2 Query Forms[3]-[4]

To access database, forms-based query interfaces are used. The design of form based interfaces is a key step in the process of database. But in that interfaces, it is capable of expressing very limited number of queries. The form expresses all the queries that an user may have. Form is a simple query interface used to access database. It doesn’t require any additional knowledge from the part of user for processing forms. The users no need to worry about the internal language used or other types of internal organization. This query forms will give an overview of the underlying datas. Most users will use these query forms to access database because of these simplicity. While creating form based interfaces,careful analysis is needed as any error will make the form ineffective. So creation of form must need full evaluation of the user preferences and user requirements. To design such forms, interface must develop a clear understanding about the datas.

A common problem of database systems is that it is very hard to give query for users who are uncomfortable with a formal query language. Inorder to solve this problem, form-based interfaces are used. Here it will automatically checks which questions are the most important for setting the query. One of the drawback of form based query interfaces is it is restrictive i.e, it doesn’t allow the user to express the query in another form.

3.3 Dataspaces[5]

It is a new abstraction for data management. It is a data co-existence approach. Its main goal is to provide functionality over all data sources. Two main scenarios are there:-Personal Information Management and Scientific Management.

1) Personal Information Management: The main goal of PIM is to offer easy access of all the information on a persons desktop. Recent desktop tools have some disadvantages like they are limited to keyword queries. Normally all desktops contain structured data like spreadsheets so the next step of PIM is to search the desktop more meaningful ways. For example, find the list of students who got A grade in database paper.

2) Scientific Data Management: This will focus on environmental observation and forecasting. A research group will monitor a coastal ecosystem through weather stations, and remote imagery. These calculations may require importing data and model outputs from other groups. The observations thus got from these calculations were the inputs to programs to generate data products. People can access these datas that had some basic file attributes like time period covered, geographic region, height or depth. This scenario explains many dataspace requirements like:-a dataspace-wide catalog, support for data lineage and creating collections and indexes.

3.4 Information Extraction[6]

Information Extraction systems are used to extract information that is relevant to a particular document. Here the system focus on different kinds of resources on the web:HTML tables and database behind forms. There are three types of systems:TextRunner system, webTables, and deep web. The TextRunner system focus on unstructured data. The Webtables system extracts relational tables from HTML. The deep-web mainly focus on backend database which are accessible via HTML forms.

3.5 CADS

Collaborative Adaptive Data Sharing Platform 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.

CADS system has two types of actors: producers and consumers. Producers upload data in the CADS system using interactive insertion form. Consumers search for relevant information using adaptive query forms. Here two modules are present: Insertion module and Query module. Insertion Phase: In this phase submission of the document is done. After the upload, CADS analyzes the text and creates an adaptive insertion form. This insertion form must contain probable attribute values to annotate the document. The user fills the form with suitable information and submits it to the database.

Query Phase: In this phase, the user work with adaptive query form. Here some default attributes are present and if any user wants to add more attributes that provision is also available. There is also a description attribute if user wants to describe about the document. In some cases, attributes recommendation is also helpful. For example, if a user specifies the attribute "Category" and other users who specified "Category" also specified "Type", then the adaptive form suggest to the user the attribute "Type". But if the attribute suggested by the user is similar to the already existing attribute, then the CADS will suggest a mapping between the two attributes. After completing the query form, it will submit to database. Finally the CADS system will find the most important pieces of data. Also CADS will return the whole document.

3.6 Document Annotation[7]

Normally organizations generate and share their descriptions of their services. Such datas contain structured information which buried under unstructured text. Information extraction algorithms are expensive to process these unstructured information, as sometimes these information does not focus on the targeted information. A novel approach is present to find out the datas which contain the information of interest. This information is used for querying the database. Here a joint utilization of the content and query workload is done. Collaborative Adaptive Data Sharing (CADS) platform helps to annotate the document automatically by identifying and suggesting attributes.

Collaborative Adaptive Data Sharing Platform (CADS), which is meant for annotating the document as they created by the owner inaddition to examining the full document later. The main goal of this paper is to suggest annotations for the document. While identifying and suggesting attributes, the attributes must have high querying value with respect to the query workload i.e., they must appear in many queries and also high content value i.e., they must be relevant to the document. These suggested attributes were used for annotating the document.

In this paper, users no need to examine the full document for suggesting and identifying the attributes for annotating the document. The CADS will automatically suggest the needed attributes by checking the frequency and relevancy of each attributes. Here attributes are generated by two processes in parallel i.e., by inspecting the content of the document and by inspecting the types of queries used. Normally attribute-value annotation is used. But it is limited to simple keywords. Querying from those annotations is very difficult as user have to type more queries for accurate suggestions. Users must know the detailed schema and all field types to use. Another issue here is it may sometimes have more attribute names for single attribute.

4. Proposed Methodology

The main aim is to develop an annotation on webpage-metadata based on attributes which appeared frequently on the metadata. Here main attribute values are considered inorder to understand the metadata. Also another processing is also done which is based on documents. Main attribute values from a document are extracted so that with that attribute values annotation can be done on documents. These attribute values are extracted from the documents automatically thereby reducing the effort of reading the full document to understand what the document really meant.

Then searching is also provided with these attribute values. In the document annotation, while uploading the document the main attribute values are found with the help of an algorithm. This will reduce the workload and also save time. The algorithm selects data from the text file and assigns it to a default cluster. Then high frequency values from this file are selected as attribute information. Check whether this attribute information is in the cluster. If so, assign this files to that cluster otherwise create a new cluster. This will improve the searching process than normal searching.

5. Conclusion

The annotation system helps to reduce the query workload by automatically suggesting attributes. Many data mining techniques are proposed. Here some of the annotation schemes used is discussed. Also a brief description about the proposed methodology is given above. This will improve the searching process and also reduce workload and time. With the help of this technique, searching and analysis of document and webpage –metadata will become efficient and fast. Here attribute values are found that have frequent occurrence. Using these attribute values annotation process can be improved.

References

[1] B. Sigurbjo¨rnsson and R. van Zwol, *"Flickr Tag RecommendationBased on Collective Knowledge,"* Proc. 17th International conference on World Wide Web (WWW '08).New York,ACM,2008 pp. 327-336,.

[2] B. Russell, A. Torralba, K. Murphy, and W. Freeman, *"LabelMe: A Database and Web-Based Tool for Image Annotation,"* Int'l J.Computer Vision, vol. 77, pp. 157-173, International Journal of Computer Vision.

[3] M. Jayapandian and H.V. Jagadish, *"Automated Creation of a Forms-Based Database Query Interface,"* Proc. VLDB Endowment, vol. 1, pp. 695-709, Aug. 2008.

[4] M. Jayapandian and H. Jagadish, *"Expressive Query Specification through Form Customization,"* Proc. 11th Int'l Conf. Extending Database Technology: Advances in Database Technology (EDBT '08), pp. 416-427, http://doi.acm.org/10.1145/1353343.1353395, 2008.

[5] M. Franklin, A. Halevy, and D. Maier, *"From Databases to Dataspaces: A New Abstraction for Information Management,"* SIGMOD Record, vol. 34, pp. 27-33, http://doi.acm.org/10.1145/1107499.1107502, Dec. 2005.

[6] M.J. Cafarella, J. Madhavan, and A. Halevy, *"Web-Scale Extraction of Structured Data,"* SIGMOD Record, vol. 37, pp. 55-61, http://doi.acm.org/10.1145/1519103.1519112, Mar. 2009.

[7] Eduardo J.Riuz, Vagelis Hristidis, Panagiotis G.Iperiotis, *"Facilitating Document Annotation using Content and Querying Value"* IEEE Transactions on knowledge and data engineering vol 26.pp.no.2 year 2014.