Retrieval of Documents by using Annotation Techniques

Ms Vaishali R. Khobragade[1], Mr V.D Thombare [2]
[1, 2] Department of Computer Engineering
[1, 2] SKN Sinhgad Institute of Technology & Science ,Lonavala Pune

Abstract- *Annotation is a process of adding the information into the Document which is useful in effective information retrieval. Annotation can be applied to several fields like image, videos, documents, etc. It 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 hard work because people have to read the documents fully to think which Sentences have to be annotated.*

In this paper, CADS System is used to generate the structured attribute by identifying the documents which contain the information of interest and this information in future useful for querying the database.

Here, we use LableMe which is an annotation tool for image annotation.

The major contribution of this paper, here, we provide a facility of annotating the images that may present in the documents which helps user to retrieve data that is useful and faster than existing system. Techniques used in this paper will give the better results compared to the methods which only relay on the content of the document or only on the query workload.

Keywords- Document Annotation, CADS system, Image Annotation, LableMe annotation tool.

I. 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 data 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 serve 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 hard work because people had to read the documents carefully to check which part to be annotated. Automatic recommendation of Sentences helps to shaving off time. In addition 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.

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.

For Annotation process, we address CADS (Collaborative Adaptive Data Sharing platform), is an “annotate-the document” infrastructure which promote fielded data explanation to direct the annotation system our CADS system uses Query Workload, along with examining the content of the document. The goal of CADS is to lower the cost of annotated document which can be useful for the user given queries.

In this approach, the user generates a document what the user want to annotate and uploads it to the database repository. Then, CADS investigate the each content of the document and create an adaptive insertion form. This adaptive insertion form contains the information need (query workload), the best attribute names of the text document and possible attribute values given the document text.

II. LITERATURE REVIEW

[1] Eduardo J. Ruiz, Vagelis Hristidis, and Panagiotis G. Ipeirotis : ”Facilitating Document Annotation Using Content and Querying Value.” This paper[1] presents the algorithms that identify structured attributes that are likely to appear within the document, by jointly utilizing the content of the text and the query workload.

[2] CADS: This paper [2] proposed CADS system, which is used as a Collaborative Adaptive Data Sharing platform, and is a data sharing platform where the integration and annotation take place at the time of data insertion i.e. production and querying i.e. consumption actions. A main goal of CADS is to influence the information demand for creation of adaptive insertion and query forms.

[3] Proximity Ranking: The Recent studies show that the term proximity is highly correlated with relevancy of document, and proximity aware ranking increases the top results precision significantly. And, there are only few studies which increase proximity-aware searching query efficiency using techniques of early-termination [3], [4]. The techniques which are discussed in [3], [4] generate an additional inverted index for each term pair, which results in a large space. [4] studied only the problem for queries with two keywords.

[4] LableMe : B. Russell, A. Torralba, K. Murphy, and W. Freeman : propose a paper “LabelMe: A Database and Web-Based Tool for Image Annotation”. A tag prediction for images is proposed in this paper[5]. It proposes web-based tool for easy image annotation and instant sharing of annotations. It detects the objects and finds similarity with existing dataset. It helps for image search in web.

[5] A tag prediction for images is proposed in this paper[5]. It proposes web-based tool for easy image annotation and instant sharing of annotations.

[6] Instant Search: The integration of proximity information in instant fuzzy search for achieving the better complexities is explained in Many recent studies focused on the instant search. The studies in [6] proposed query and indexing techniques to support the instant search. Li et al. [6] studied the instant search on relational data which is modeled as a graph.

III. IMPLEMENTATION DETAILS

In the proposed system, we design methodology for annotating documents as well as images which will be able to reduce searching time, space requirement along with providing multiple search facility. Proposed system will provide a facility of image annotation, which is an effective and useful concept to find user’s interested information in less time as compared with the existing system.

Here, two techniques of annotations are implemented.

A. CADS System:

[Collaborative Adaptive Data Sharing Platform] CADS 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.

B. LableMe annotation tool.

Label Me 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 evaluation 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.

C. System Overview

The fig. 1 shows the Architectural view of the proposed system. The description of the system is as follows:

As shown in the Architecture for document annotation, user uploads the documents in the document uploader which already contains the annotated document and the. new document to annotate, where CADS system retrieves the documents and calculates the content value, query value for the attributes and also recommend the attribute with highest value and generate the new insertion form. This insertion form contains the best attribute for the new unstructured document. the annotated documents generated by CADS are stored in database. User also uploads the images in image uploader .these images are annoted by using online annotation tool LableMe.these annotated imges are stored into the database. When user search for a particular file ,it may be a file name, queary or content or it can be the image name too, the algorithm check the similarity of user's queries with the annotated documents and annotated images. The proposed algorithm also check the frequently asked queries by maintaining history of users queries and will return the most useful user's interested document or image.

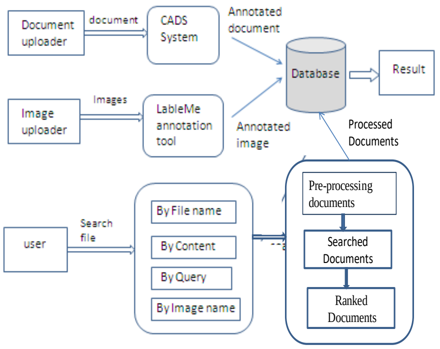


Fig.1: System Architecture of proposed system.

D. Mathematical Model

Let S be the system which contains inputs, functions, and outputs.

S = {I, Im,F, O} where

1. I = {I1, I2, I3. . . In}

Where, `I' is the set of documents that user wants to upload in text, pdf, word format and there can be multiple files uploaded on server by multiple users or dataset of documents.

2). Im = { Im1, Im2,…….. In}

Where, Im is the set of images that user wants to upload jpg,etc...

3) F = {F1, F2}
Here, two functions are defined which forms the system where

F1 = Identification, separation of attribute values from attribute names and their insertion in csv file.

F2 = Instant search with proximity ranking

4) O = {O1, O2, O3. . . O}

Where, `O' is the set of outputs which contain:

O= Set of resulted documents, images.

Ranking function:

Ranking will use following function to rank the resultant documents:

For each document d,

$$W = \sum_{1}^{n} i \qquad (1)$$

Where,

1} W = Weightage of query keywords in Documents.

2} i = weightage of each word in the Document

=1/total no. of words in the document

3) n = total no. of query keywords **.**

E. Algorithm:

The proposed system implements following Algorithm[3]. It is used for searching and ranking relevant documents:

Inputs: Documents in dataset D,

Query entered by user Q.

Output: Ranked relevant documents list.

Let n be the total no. of documents in dataset.

I. When user enters a valid query,

1. for i =1 to n

2. Read document content

3. Compare query keyword with content of document

4. If (70% word match found)
Display the document

5. Else
Ignore and Go to next document.
II. Ranking function:

Finally, the valid segmentations are ranked using eqution(1).

IV. EXPERIMENT

Data set:
For our experiments, we use an online shopping site "EBAY" for collecting documents. It consists thousands of electronic product reviews. The data set contains different kinds of products like Mobiles, cameras, video games, television, audio sets, and alarm clocks.

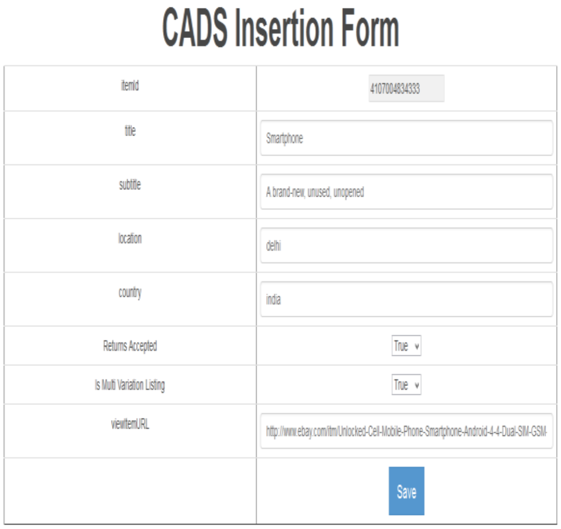
Annotations:

We generated annotations for the data sets, which we use as training and test data.

To annotate the EBAY reviews, we used the EBAY specifications page for each product. The page contains structured data for a product in the form of "attribute name, value." Given that we are only interested in annotations that come from the document text (i.e., the product's review), we removed annotations that are not mentioned in any sentence in the review text.

Fig. 2 presents the adaptive insertion form for that document. The system adds the suggested attributes to a set of default attributes like: "Document Type," "Date," and "Location," which are the basic metadata that the user always provides, as defined by a domain expert.

Here,CADS system present an adaptive technique for automatically generating data input forms, for annotating unstructured textual documents, such that the utilization of the inserted data is maximized, given the user information needs.



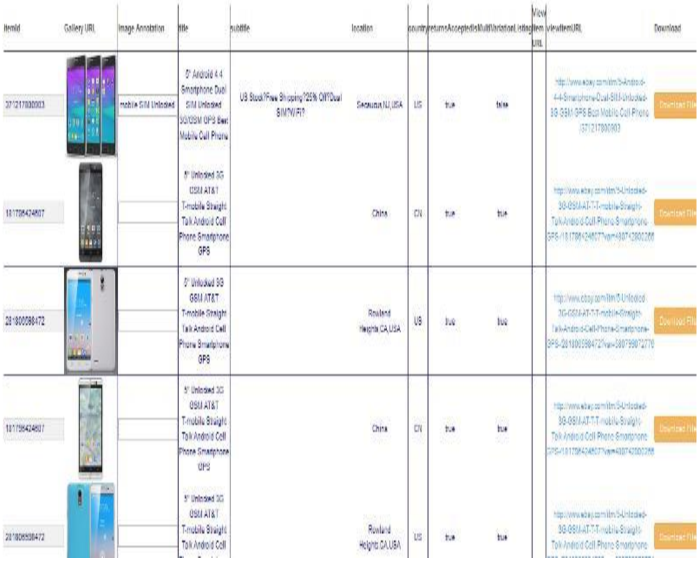
CADS Insertion Form

itemId	4107004834333
title	Smartphone
subtitle	A brand-new, unused, unopened
location	delhi
country	india
Returns Accepted	True
Is Multi Variation Listing	True
viewItemURL	http://www.ebay.com/itm/Unlocked-Cell-Mobile-Phone-Smartphone-Android-4-4-Dual-SIM-GSM-
	Save

Fig 2. Adaptive insertion form

V. RESULT

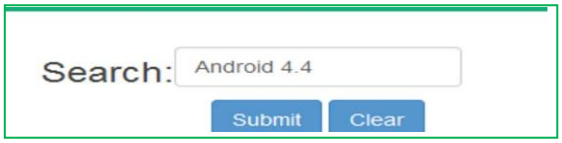
This system shows the result, which automatically generates the annotations for documents and images which uploads by the user with the help of CADS system and LableMe annotation tool.



itemId	Gallery URL	Image Annotation	title	subtitle	location	country	returnsAccepted	isMultiVariationListing	View Item URL	viewItemURL	Download
371217800903		mobile SIM Unlocked	5" Android 4.4 Smartphone Dual SIM Unlocked 3G/GSM GPS Best Mobile Cell Phone	US Stock?Free Shipping?25% Off?Dual SIM?WIFI?	Secaucus,NJ,USA	US	true	false		http://www.ebay.com/itm/5-Android-4-4-Smartphone-Dual-SIM-Unlocked-3G-GSM-GPS-Best-Mobile-Cell-Phone-/371217800903	Download File
181786424607			5" Unlocked 3G GSM AT&T T-mobile Straight Talk Android Cell Phone Smartphone GPS		China	CN	true	true		http://www.ebay.com/itm/5-Unlocked-3G-GSM-AT-T-T-mobile-Straight-Talk-Android-Cell-Phone-Smartphone-GPS-/181786424607?var=480742800266	Download File
281800598472			5" Unlocked 3G GSM AT&T T-mobile Straight Talk Android Cell Phone Smartphone GPS		Rowland Heights,CA,USA	US	true	true		http://www.ebay.com/itm/5-Unlocked-3G-GSM-AT-T-T-mobile-Straight-Talk-Android-Cell-Phone-Smartphone-GPS-/281800598472?var=580799072770	Download File
181786424607			5" Unlocked 3G GSM AT&T T-mobile Straight Talk Android Cell Phone Smartphone GPS		China	CN	true	true		http://www.ebay.com/itm/5-Unlocked-3G-GSM-AT-T-T-mobile-Straight-Talk-Android-Cell-Phone-Smartphone-GPS-/181786424607?var=480742800266	Download File
281800598472			5" Unlocked 3G GSM AT&T T-mobile Straight Talk Android Cell		Rowland Heights,CA,USA	US	true	true		http://www.ebay.com/itm/5-Unlocked-3G-GSM-AT-T-T-mobile-Straight-Talk-Android-Cell-Phone-Smartphone	Download File

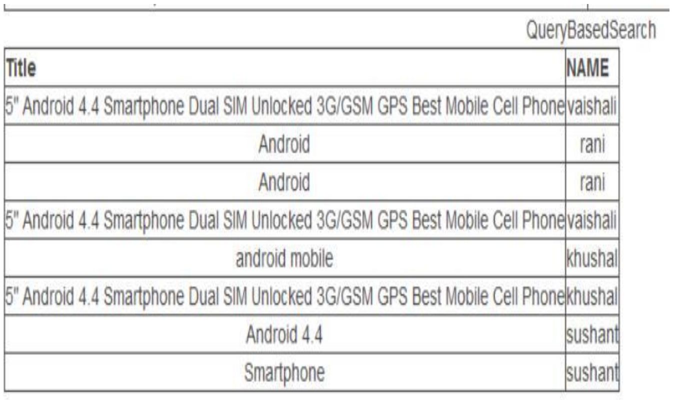
Fig 3. Annotation of documents and images.

When the end user enters the queryas shown in fig.4 query based search, he will get the ranked documents .User will receive only documents of his interest within less time as compared to the existing system, as shown in the fig 5. and In fig 6.



Search: Android 4.4

Fig 4. Query based search



QueryBasedSearch

Title	NAME
5" Android 4.4 Smartphone Dual SIM Unlocked 3G/GSM GPS Best Mobile Cell Phone	vaishali
Android	rani
Android	rani
5" Android 4.4 Smartphone Dual SIM Unlocked 3G/GSM GPS Best Mobile Cell Phone	vaishali
android mobile	khushal
5" Android 4.4 Smartphone Dual SIM Unlocked 3G/GSM GPS Best Mobile Cell Phone	khushal
Android 4.4	sushant
Smartphone	sushant

Fig 5. Output for Query based search.

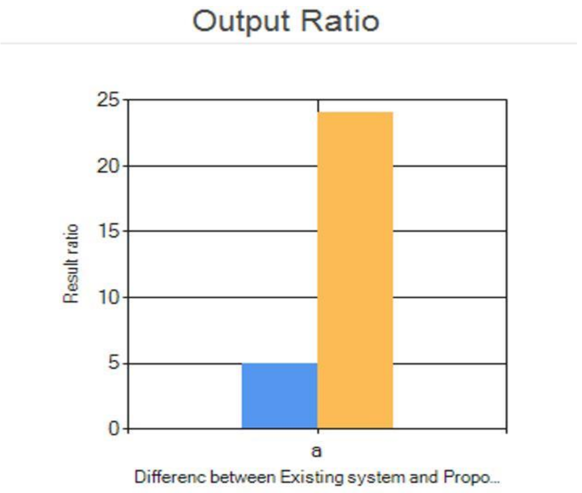


Fig 6. Output ratio.

V. CONCLUSION

This paper proposes a new approach for efficient document retrieval including multiple annotation techniques as well as searching & ranking techniques. The system tries to satisfy querying needs of user efficiently. This system gives different ways for searching: the values of Content and Query, by image name, and file name . Using these techniques, we can increase chances of documents visibility up to maximum percent. Also use Query expansion algorithm to remove stream word and stop word. It helps to reduce space requirement and searching time. The proposed system is able to provide fast and accurate data retrieval along with multiple search facility.

FUTURE SCOPE

Presently, the system generates annotation suggestions based on only one given paper, but in future additional information such as other relevant papers and citation information can be used to improve the annotation result. Annotation technique can become more powerful tool for retrieval of document as well as images, but it needs more work for suggesting appropriate attributes. Query based searching can be said as the future in information retrieval.

Document clustering can also be used in future.

ACKNOWLEDGMENT

This is to acknowledge and thank all the individuals who played defining role in shaping this paper. Without their constant Coordination, guidance and reviewing, this task could not be completed alone.

I avail this opportunity to express my deep sense of gratitude and whole hearted thanks to my guide Prof. V.D. Thombre for giving his valuable guidance, inspiration and encouragement to embark this paper.

REFERENCES

[1] [1]] Eduardo J. Ruiz, Vagelis Hristidis, and Panagiotis G. Ipeirotis :”Facilitating Document Annotation Using Content and Querying Value.”[2014]ieee paper.

[2] V. Hristidis, E. Ruiz, " CADS: A Collaborative Adaptive Data Sharing Platform", SCIS, International University, Florida, 2009

[3] H. Yan, J. Wen, S. Shi, F. Zhang, T. Suel,, ―Efficient term proximity search with the term-pair indexes,"CIKM, 2010, pp. 1229-1238.

[4] H. Bast, , A. Chitea, F.Suchanek,Weber, ―Ester : efficient search on text,entities, and relations," SIGIR, 2007”

[5] B. Russell, A. Torralba, K. Murphy, and W. Freeman : propose a paper “LabelMe: A Database and Web-Based Tool for Image Annotation”

[6] Md. Abu Nisar Masud, Md. Munasir Mamun, ―A General Approach to Natural Language Generation‖ In Proceeding of IEEE, INMIC, 2003.