Facilitating Document Annotation Using Content & Querying Value

Akshay Shingote
KBTCOE,Nashik (India)

Nikhil Vispute
KBTCOE,Nashik (India)

Priyanka Dhikale
KBTCOE,Nashik (India)

ABSTRACT **:** Collections of huge, large textual data contains significant amount of structured information, which remains hidden in unstructured text. Relevant information is always difficult to find in these documents.

In this paper we proposed an alternative approach that facilitates the generation of the structured metadata by identifying documents that are likely to contain information of interest and this information is going to be useful for querying the database. Here people will likely to assign metadata related to documents which they upload which will easily help the users in retrieving the documents.

Keywords *- Annotation, CADS, Information Extraction*

I. INTRODUCTION

Many systems do not have the basic "attribute-value" annotation that would make a querying feasible. Annotations that use "attribute-value" pairs require users to be more principled in their annotation efforts. Users need to have good idea in using and applying the annotations or attributes.

Even if the system allows users to annotate the data with such attribute-value pairs, the users are often unwilling to perform the task. Such difficulties results in very basic annotations that is often limited to simple keywords. Such simple annotations make the analysis and querying of the data cumbersome. Users are often limited to plain keyword searches, or have access to very basic annotation fields, such as "creation date" and "size of document" .

In this paper, we propose CADS (Collaborative Adaptive Data Sharing) platform which is an "annotate-as-you-create" infrastructure that facilitates fielded data annotation. A key contribution of our system is the direct use of the query workload to direct the annotation process, in addition to examining the content of the document. Our aim is to prioritize the annotation of documents towards generating attribute names and attribute values for attributes that will often used by querying users and these attribute values will provide best possible results to the user wherein users will have to deal only with relevant results.

II. RELATED WORK

There are several systems that favor the collaborative annotation of objects and use previous annotations or tags to annotate new objects [1]. There have been significant amounts of work in predicting the tags for documents or other resources.

We argue that our approach is different as compared to an traditional approach. But by assigning annotations to documents will help in improving faster efficiency in searching.

III. IMPLEMENTATION

3.1 Proposed Information Extraction Algorithm

Information Extraction algorithm is the algorithm we use to extract contents of text file. Following fig shows how information extraction takes place.

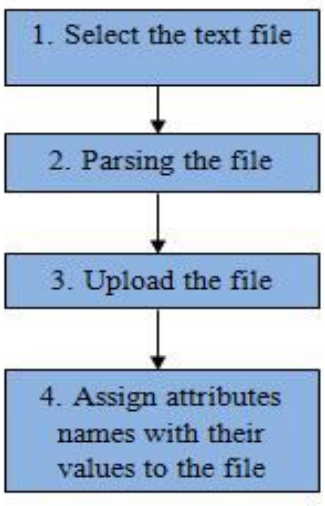

Fig 3.1 Information Extraction Algorithm

Our goal is to suggest annotations for a document.

1) Select a text file

2) Parse the text file. Ignore stopwords from it and count frequency of high querying keywords which will be important for content based search. Maintain frequency count of these keywords appearing in only single document.

3) Upload the file on to the server

4) Then fill all the annotations which are relevant to the document which can be useful for query-based searching.

Example : year=2012,location='Nashik' , author ='Bill Gates' etc.

3.2 QV,CV Computation and Combining

Algorithm:

1) Enter the queries for retrieving the document

Example: location='Nashik' and year=2012

2) Split the queries and pass it to database for retrieving

3) Check all related results and show the related results to user.

4) For much efficient and accurate results,users should try to enter maximum queries they can.

IV. RESULTS

Following are the experimental results of our implemented system.

Fig 5.1 An Publisher has chosen the file "imp.txt" for uploading. Here after uploading the document, document is parsed where stop words are ignored and high keywords are stored in database

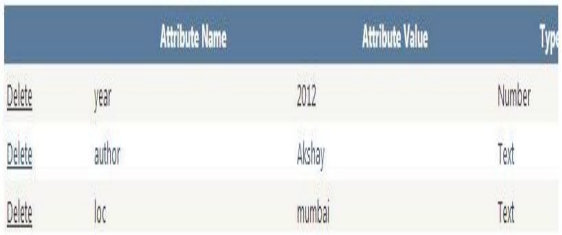

Fig 5.2 After uploading the file by an Publisher, Publisher assigns an attribute name "Domain" and its respected values and type "Data Mining" and "Text"

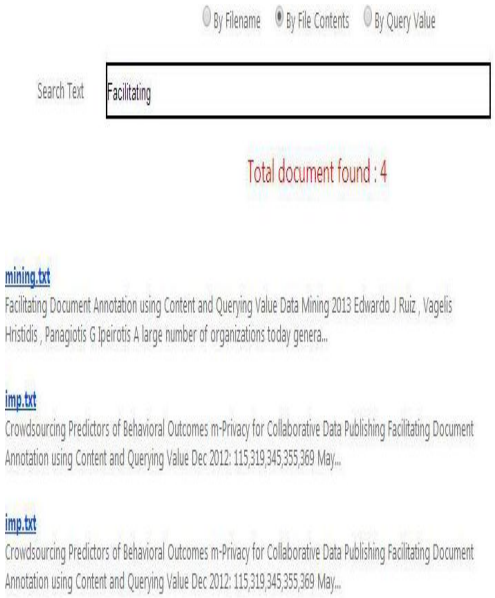

Fig 5.3 Content-based search example performed by end-user. The reviewer enters the word "Facilitating" and all documents containing that word is shown

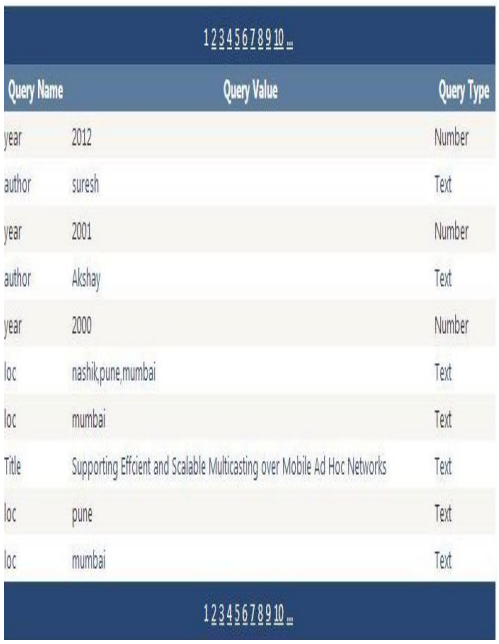

Fig 5.4 Full List of attributes/annotations,their respected values and types can be seen by end-user where he can use these values for query-based searching

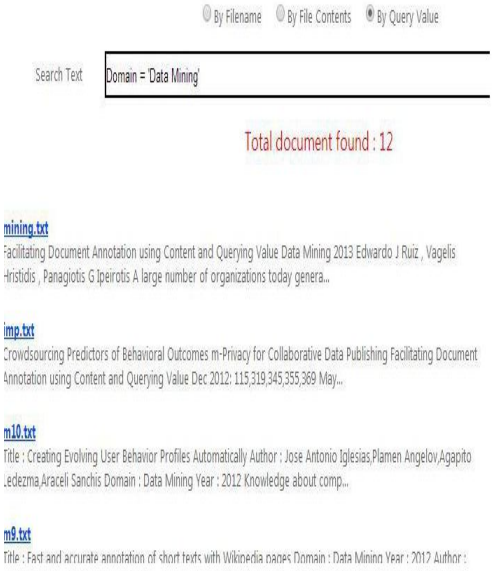

Fig 5.5 Example of query-based search where user enters attribute "Domain" and its respected value "Data Mining" and the related files are displayed

Fig 5.6 Another example of query-based search. Here user enters two queries "Domain='Data Mining' and author='Akshay' which gives much accurate and distinct results.

We have implemented this application using ASP.NET using C# for front end designing and SQL Server 2005 as back end. We have used ADO.NET technology of Microsoft to link front end and back end in our application.

V. CONCLUSION

We presented two ways to combine these two pieces of evidence, content value and querying value. The main advantages of our application is mainly that when users perform query based search, they could get minimum and distinct results where it could be easy for retrieval. By using these techniques, workload of application can reduce by large amount. Also, given the fact the efficiency of searching will be faster because of using the query-based searching technique.

Query-based searching will be the future in information retrieval as this searching techniques may be applied on other file formats like .docx,.pdf,.xml etc which can give users better,faster and accurate results and will also increase the performance. This application can surely give a huge boost to mainly in text mining which can be thought of as an changing trend or technology.

VI. Acknowledgements

We take this opportunity to thank Prof Jaya.R.Suryawanshi,Nitin.S.Ujgare & Vaishali.S.Pawar for their valuable guidance and for providing all the necessary support to accomplish this research. We would like to extend our gratitude towards our beloved Principal J.T.Pattiwar for his great support.

VII. REFERENCES

[1]. Eduardo J. Ruiz, Vagelis Hristidis, Panagiotis G. Ipeirotis ,"**Facilitating Document Annotation using Content and Querying Value**", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING VOL.PP NO.99 2013.

[2] Vagelis Hristidis, Eduardo Ruiz," **CADS: A Collaborative Adaptive Data Sharing Platform",** School of Computing and Information Sciences, Florida International University.

[3] M. J. Cafarella, J. Madhavan, and A. Halevy, "**Web-scale extraction of structured data,**" SIGMOD *Rec.*, vol. 37, pp. 55–61, March 2009.

[4] A. Halevy, Z. Ives, D. Suciu, and I. Tatarinov, "**Schema mediation in peer data management systems,**" in Data Engineering, 2003. Proceedings. 19th International Conference on, March 2003, pp. 505 – 516

[5] R. Fagin, A. Lotem, and M. Naor, "**Optimal aggregation algorithms for middleware,**" J. Comput. Syst. Sci., vol. 66, pp. 614–656, June 2003.

`