International Journal of Advanced Technology in Engineering and Science www.ijates.com
Volume No 03, Special Issue No. 01, March 2015 ISSN (online): 2348 – 7550

FACILITATING DOCUMENT ANNOTATION USING METADATA

[1] R.Nalini, [2] M.Parimala

[1] PG Scholar, [2] Assistant Professor, Department of CS & IT,

Dhanalakshmi Srinivasan College of Arts & Science for Women, Perambalur, Tamilnadu (India)

ABSTRCT

A large number of organizations today generate and share textual descriptions of their products, services, and actions .collections of textual data contain significant amount of structured information, which remains buried in the unstructured text. While information extraction algorithms facilitate the extraction of structured relations, they are often expensive and inaccurate, especially when operating on top of text that does not contain any instances of the targeted structured information..I present a novel 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 subsequently useful for querying the database. Identify the metadata when such information actually exists in the document, instead of naively prompting users to. As a major contribution of I present a algorithm that identify structured attributes that are likely to appear within the document, by jointly utilizing the content of the text and the query workload. That our approach generates superior results compared to approaches that rely only on the textual content or only on the query workload, to identify attributes of interest.

Keyword: CAD Platform, Coordinate Matching, Information Extraction, Keyword Search, Metadata.

I. INTRODUCTION

Clustering algorithms are typically used for exploratory data analysis, where there is little or no prior knowledge about the data. This is precisely the case in several applications of *Computer Forensics*, including the one addressed in our work. From a more technical viewpoint, our datasets consist of unlabeled objects the classes or categories of documents that can be found are *a priori* unknown. Moreover, even assuming that labeled datasets could be available from previous analyses, there is almost no hope that the same classes would be still valid for the upcoming data, obtained from other computers and associated to different investigation processes. More precisely, it is likely that the new data sample would come from a different population. In this context, the use of clustering algorithms, which are capable of finding latent patterns from text documents found in seized computers, can enhance the analysis performed by the expert examiner.

The rationale behind clustering algorithms is that objects within a valid cluster are more similar to each other than they are to objects belonging to a different cluster. Thus, once a data partition has been induced from data, the expert examiner might initially focus on reviewing Representative documents from the obtained set of clusters. Then, after this preliminary analysis, it may eventually decide to scrutinize other documents from each cluster. By doing so, one can avoid the hard task of examining all the documents but, even if so desired, it still could be done.

In a more practical and realistic scenario, domain experts are scarce and have limited time available for performing examinations. Thus, it is reasonable to assume that, after finding a relevant document, the examiner could prioritize the analysis of other documents belonging to the cluster of interest, because it is likely that these are also relevant to the investigation.

Clustering algorithms have been studied for decades, and the literature on the subject is huge. Therefore decided to choose a set of representative algorithms in order to show the potential of the proposed approach, namely: the partition K-means and K-melodies. Thus, as a contribution of our work, we compare their relative performances on the studied application domain using five real-world investigation cases conducted by the Brazilian Federal Police Department. In order to make the comparative analysis of the algorithms more realistic, two relative validity indexes have been used to estimate the number of clusters automatically from data.

II. LITERATURE SURVEY

2.1. Facilitating Document Annotation Using And Querying Value:

- ❖ In this paper 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.
- ❖ In other words are trying to prioritize the annotation of documents towards generating attribute values for attributes that are often used by querying users.

2.1.1 Disadvantage

- ❖ Unavailability of proper information to different levels of query.” Coordinate matching” by inner product similarity.
- ❖ It does not provide more accurate data only get the files from only exact name required.
- ❖ Unavailability of proper information rarely differentiates the search results.

2.2. Towards a Business Continuity Information Network for Rapid Disaster Recovery

Most of the recent work has been conducted for crisis management under terrorist attacks and emergency management services under natural disasters with private business continuity and disaster recovery a secondary concern. In this paper, we propose a model for pre-disaster preparation and post-disaster business continuity/rapid recovery. The model is utilized to design and develop a web based prototype of our Business Continuity Information Network (BCIN) system facilitating collaboration among local, state, federal agencies and the business Community for rapid disaster recovery. We present our model and prototype with Hurricane Wilma as the case study.

2.2.1 Disadvantages

We then utilize our model for the implementation of a web based Business Continuity Information Network (BCIN) that creates a disaster management dataspacebased on the communication among the stakeholders and enables businesses, emergency management community and NGOs to effectively communicate, identify and assist in the execution of preparation and recovery plans; identify user relevant and location specific disaster preparation and recovery resources along with the business/employee loan and assistance programs facilitating intelligent decision support; and dynamically disseminate location and user specific information regarding key inhibitors to preparation and recovery process such as open/closure status of schools, businesses, transportation, roadways and emergency services etc.

III. SYSTEM ARCHITECHU

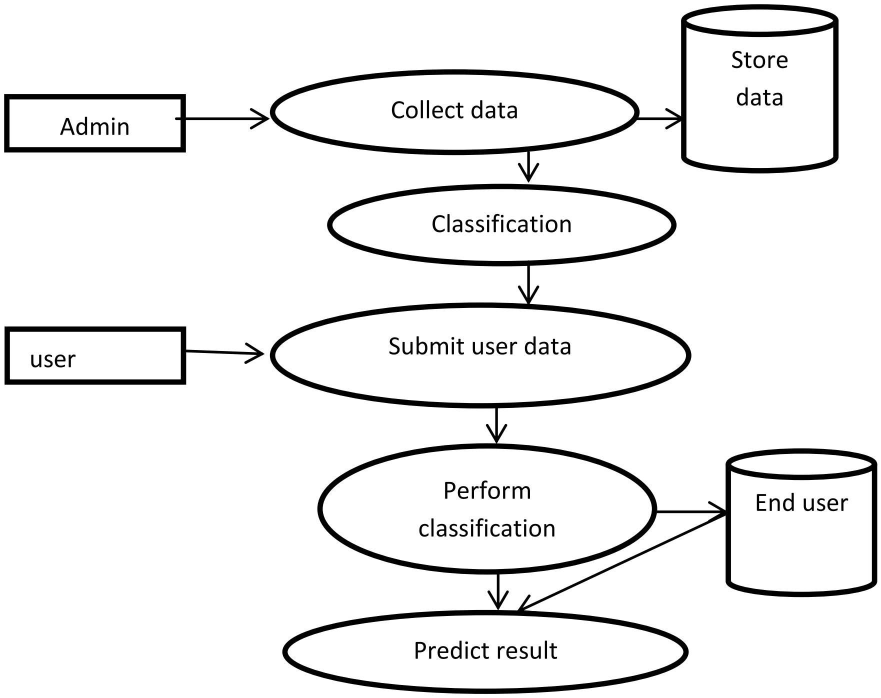

Figure: 3.1 System Architecture Diagram

IV.METHODOLOGY

4.1. Registration

In the registration phase the new user can register the details and get the service, if there is any new user they can create the new login id, in registration the new use must give full details about the name and other details. Finally they will get the user name and password.

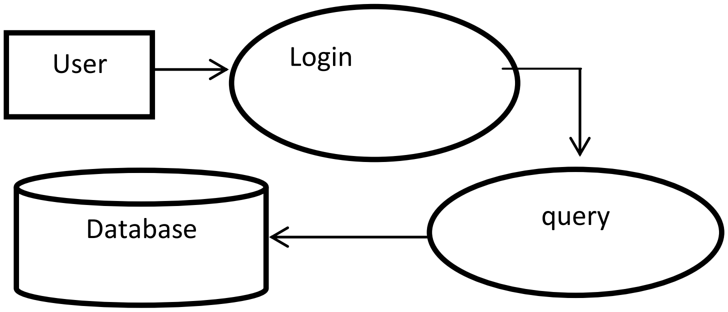

Figure: 4.1.1Registeration Diagram

4.2. Login

In this module, any of the above mentioned person have to login, they should login by giving their email id and password. This Module is a portal module that allows users to enter a User Name and Password to log. This Module displays a *username* and password Login form to perform authentication with user ID and password. If the user enters a valid username/password combination they will be granted access to additional resources on your website.

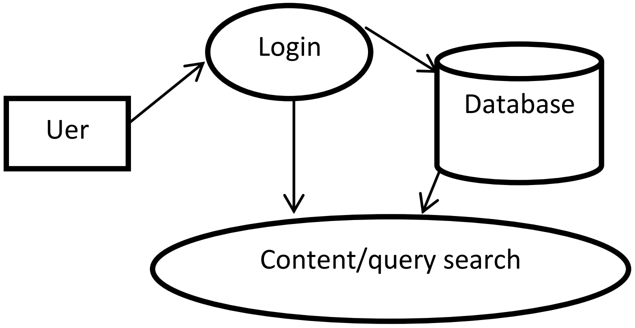

Figure: 4.2.1 Login Diagram

4.3. Document Upload

In this module Owner uploads an unstructured document as file (along with Meta data) into database, with the help of this metadata and its contents, the end user has to download the file. It has to enter content/query for download the file.

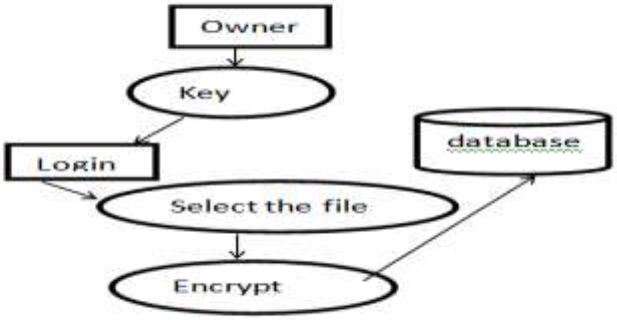

Figure: 4.3.1.Document upload

4.4. Search Techniques

Here we are using two techniques for searching the document

1) Content Search,

2) Query Search.

4.4.1 Content Search

It means that the document will be downloaded by giving the content which is present in the corresponding document. If its present the corresponding document will be downloaded, otherwise it won't.

4.4.2. Query Search

It means that the document will be downloaded by using query which has given in the project. If its input matches the document will get download otherwise it won't.

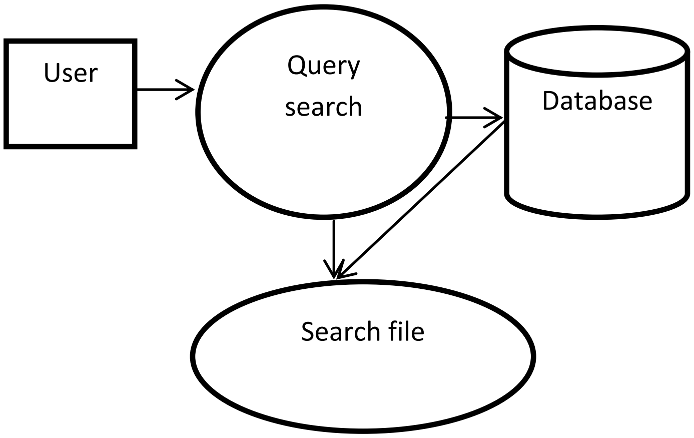

Figure: 4.4.1.Search Technique

4.5. Download Document

User has to download the document using query/content values which have given in the base paper. It enters the correct data in the text boxes, if it's correct it will download the file. Otherwise it won't.

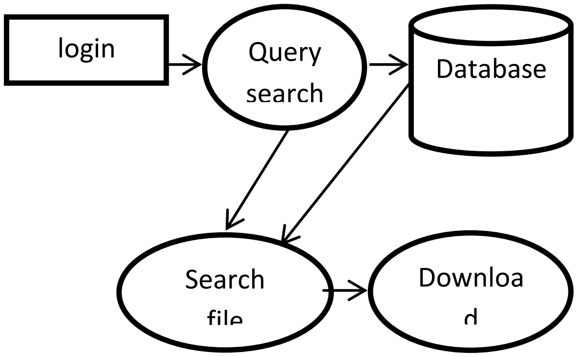

Figure: 4.5.1 Download Document

IV. CONCLUSION AND FUTURE WORK

The adaptive techniques to suggest relevant at-tributes to annotate a document, while trying to satisfy the user querying needs. Our solution is based on a probabilistic framework that considers the evidence in the document content and the query workload. We present two ways to combine these two pieces of evidence, content value and querying value: a model that considers both components conditionally independent and a linear weighted model. Experiments shows that using our techniques, we can suggest attributes that improve the visibility of the documents with respect to the query workload by up to 50%. That is, we show that using the query workload can greatly improve the annotation process and increase the utility of shared data. In Future, The form contains the best attribute names given the document text and the information need (query workload), and the most probable attribute values given the document text. The author (creator) can inspect the form, modify the generated metadata as necessary, and submit the annotated document for storage. Our efforts focus not only on identifying the potential annotations fields that exist in complete and optimal annotations for document , but also to rank them and display on top the most important ones. Since the goal of annotations is to facilitate future querying, we want the annotation effort to focus on generating annotations useful for the queries in the query workload.

V. ACKNOWLEDGEMENT

The author deeply indebted to honorable Shri A.SRINIVASAN (Founder Chairman), Shri P.NEELRAJ(Secretary) Dhanalakshmi Srinivasan Group of Institution , Perambalur for giving me opportunity to work and avail the facilities of the college campus . The author heartfelt and sincere thanks to Principal Dr.ARUNA DFINAKARAN, Vice Principal Prof.S.H.AFROZE HoD Mrs.V.VANEESWARI, (Dept of CS & IT) Project Guide Mrs.A.PARIMALA, (Dept of CS & IT) of Dhanalakshmi Srinivasan College of Arts & Science for Women, Perambalur. The author also thanks to parents, Family Members, Friends, Relatives for their support, freedom and motivation.

REFERENCES

1. S.R. Jeffery, M.J. Franklin and A.Y. Halevy, &ldquo, Pay-as-You-Go User Feedback for Data space Systems,&rdquo, *Proc. ACM SIGMOD Int',l Conf. Management Data,* 2008.
2. K. Saleem, S. Luis, Y. Deng, S.-C. Chen, V. Hristidis and T. Li, &ldquo, Towards a Business Continuity Information Network for Rapid Disaster Recovery, & rdquo, *Proc. Int', l Conf. Digital Govt. Research* 2008.
3. A. Jain and P.G. Ipeirotis, & ldquo,A Quality-Aware Optimizer for Information Extraction,&rdquo, *ACM Trans. Database Systems,* vol. 34, article 5, 2009.
4. J.M. Ponte and W.B. Croft, &ldquo,A Language Modeling Approach to Information Retrieval,&rdquo, *Proc. 21st Ann. Int',l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ',98),* pp. ,275-281, http://doi.acm.org/10.1145/290941.291008.

5. R.T. Clemen and R.L. Winkler, &ldquo,Unanimity and Compromise among Probability Forecasters,&rdquo, *Management Science*, vol. 36, pp. ,767-779, http://portal.acm.org/citation.cfm?id=81610.81609, July 1990.
6. C.D. Manning, P. Raghavan and H. Schü,tze, *Introduction to Information Retrieval*, first ed. Cambridge Univ. Press, http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&,path=ASIN/0521865719, July 2008.
7. P.G. Ipeirotis, F. Provost and J. Wang, &ldquo, Quality Management on Amazon Mechanical Turk, &rdquo, *Proc. ACM SIGKDD Workshop Human Computation* http://doi.acm.org/10.1145/1837885.1837906, 2010.
8. R. Fagin, A. Lotem and M. Naor, &ldquo,Optimal Aggregation Algorithms for Middleware,&rdquo, *J. Computer Systems Sciences*, vol. 66, pp. 614-656, http://portal.acm.org/citation. Cfm? Id= 861182.861185, June 2003.
9. K.C.-C. Chang and S.-w. Hwang, &ldquo, Minimal Probing: Supporting Expensive Predicates for Top-K Queries, &rdquo, *Proc. ACM SIGMOD Int',l Conf. Management Data*, 2002.

BIOGRPHY

Nalini.R, Dhanalakshmi Srinivasan college of Arts and Science for women, Perambalur Tamilnadu. Received her affiliated by BDU, Trichy.B.sc cs in Department of cs Puthanampatti, Trichy, Tamilnadu.Area of Interest network, Java.

Parimala .M Received M.C.A, M.E Degree in Computer Science and Engineering. Currently working as Assistant Professor in Department of Computer Science in Dhanalakshmi Srinivasan College of Arts and Science for women Perambalur, Tamilnadu. Published a Book Named “A Small Pickup From Computer Concepts”. Research areas are Networking, Web Technology, and Mobile Computing.