IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727,
PP 41-45
www.iosrjournals.org

Facilitating document annotation using content value and query value

Vaishnavi Mhaske[1], Gauri Kale[2], Rashmi Dashpute[3], Ashwini Yeole[4]
[1,2,3,4]*(Information Technology, MCOERC/ Savitribai Phule Pune University, India)*

Abstract : *Document annotation is the task of adding metadata information in the document which is useful for information extraction. In many applications domain textual data contains significant amount of structured information which is in unstructured text. So that it is always difficult to find relevant information. This paper proposes, an adaptive technique that facilitates the generation of structured metadata by identifying documents containing information of interest. Such information is further useful for querying the database. This paper proposes survey on Collaborative Adaptive Data sharing platform (CADS) for document annotation and use of query workload to direct the annotation process. A key novelty of CADS is that it learns with time the most important data attributes of the application, and uses this knowledge to guide the data insertion and querying.*
Keywords – *Annotation, CADS, Key-pair value, Metadata, Unstructured data.*

I. INTRODUCTION

Annotations are comments, notes, explanations, or external remarks. Annotations are metadata, as they give additional information about data. If the documents are properly annotated it is possible to improve quality of searching. Lack of appropriate annotations makes it hard to retrieve it and rank it properly. Existing annotations makes the analysis and querying of data cumbersome. Therefore this paper surveys, Collaborative Adaptive Data Sharing platform i.e. annotate-as-you-create infrastructure. This facilitates fielded data annotation. The key goal of proposed system is to lower the cost of document annotation and provide query workload to direct the process of annotation. Summarized output on searching particular document is prime requirement nowadays. To get such summarized search output, we have to maintain documents / data in smart way. Annotation technique is one of the best featured techniques to manage such documents and get effective search result. Attribute – value pairs are generally more meaningful and significant as they can contain more information than un-typed approaches. Efforts to keep such decent maintenance of such annotate documents user has to take extra efforts. A scenario is cumbersome, complicated and tedious where there are number of fields to be filled at time of uploading a particular document. Hence end user frequently ignores such annotation capabilities. User is still unresponsive and ignoring task though system offers the facility to randomly annotate the data with attribute-value pairs. Along with this there it also has unclear usefulness for subsequent searches in the future. Such difficulties finally tend to very basic annotations, if any at all, that are often limited to simple keywords. Such simple annotations make the analysis and querying of the data cumbersome. It's the fact that this effective but ignored attribute – value paired annotation scheme can bring smooth searching and maintenance and this motivated us to work on Collaborative Adaptive Data Sharing platform (CADS), which is an "annotate-as-you create" infrastructure that facilitates fielded data annotation. The contribution of our system is the direct use of the query workload to direct the annotation process, in addition to checking the content of the document. Along with this contribution we are also working on phrase extraction process to build knowledge out of text. CAD provides cost effective and good solution to help efficient search result. The goal of CADS is to support a process that creates nicely annotated documents that can be immediately useful for commonly issued semi-structured queries of end user.

II. EXISTING TECHNOLOGY

K. Saleem, S. Luis, Y. Deng, S.-C. Chen, V. Hristidis, and T. LiIn have considered the Crisis Management and Disaster Recovery have gained immense importance in the wake of recent man and nature inflicted calamities. They proposed a solution or model for pre-disaster preparation and post disaster business continuity/rapid recovery. In case of disaster need of rapid information retrieval and sharing increases. This paper proposed a disaster management model which works good at some extent but it is not considering the effective retrieval. S.R. Jeffery, M.J. Franklin, and A.Y. Halevy proposed a paper Pay-as-You-Go User Feedback for Dataspace Systems, this system propose a system which is a line of work towards using more expressive queries that leverage annotations is the "pay-as – you – go " querying strategy in data spaces. In data spaces users provide data integration hints at querying time. But in this paper it is assumed that data sources already contain structured information and the problem is to match the query attributes with the source attribute.

G. Tsoumakas and I. Vlahavas have proposed a paper Random K-Labelsets: An Ensemble Method for Multilabel Classification which focuses on ensemble method for multilabel classification. The RAndom k-labELsets (RAKEL) algorithm constructs each member of the ensemble by considering a small random subset of labels and learning a single-label classifier for the prediction of each element in the power set of this subset. In this way, the proposed algorithm aims to take into account label correlations using single-label classifiers that are applied on subtasks with manageable number of labels and adequate number of examples per label. Using this we can take into account the correlation between tags for annotations. But in this collaborative annotation is missing. P. Heymann, D. Ramage, and H. Garcia-Molina have proposed a paper "Social Tag Prediction" giving solution for prediction of tags for particular object. We can adopt this for out suggesting annotation concept. J.M. Ponte and W.B. Croft have proposed a paper "A Language Modeling Approach to Information Retrieval", they have considered this information retrieval scenario and proposed a solution to analyze the content. They proposed a approach to retrieval based on probabilistic language modeling. Their approach to modeling was non-parametric and integrates document indexing and document retrieval into a single model. But in these making prior assumptions about the similarity of document is not warranted. D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green: proposed a paper "Automatic Generation of Social Tags for Music Recommendation. This paper promotes same kind of auto suggestions of tags. But this is dedicated to the musical data. We are using text based documents. A. Jain and P.G. Ipeirotis, propose a paper "A Quality-Aware Optimizer for Information Extraction," This paper presents Receiver Operating Characteristic (ROC) curves to calculate the extraction quality and selection of extraction parameter. Automated information extraction (IE) algorithms used to extract targeted relations or characteristic of the document. In this case we should process only documents that actually contain such information. When we process documents that do not matched with the predefined targeted information and we use automated information extraction algorithms to extract such annotation. We often face a significant number of wrong positives results, which may lead to significant quality problem in the data annotation.

III. PROPOSED SYSTEM

This paper proposes, Collaborative Adaptive Data Sharing platform (CADS). CADS is nothing but annotate-as-you-create infrastructure that facilitates fielded data annotations. The aim of CADS is to minimize the cost creating annotated documents that can be useful for commonly issued semistrucured queries. Fig.1 represents work flow of CADS. The CADS system has two types of actors: producers and consumers. Producers upload data in the CADS system using interactive insertion forms and consumers search for relevant information using adaptive query forms. CAD's basic objective is to create very structured annotated document to trigger efficient search in minimal execution cost. Also for semi-structured queries of user CAD generate most useful output. Also CAD adopt the strategy in which document is annotate at time of creation while crater is still in "document generation" phase, even though the techniques can also be used for post-generation document annotation.

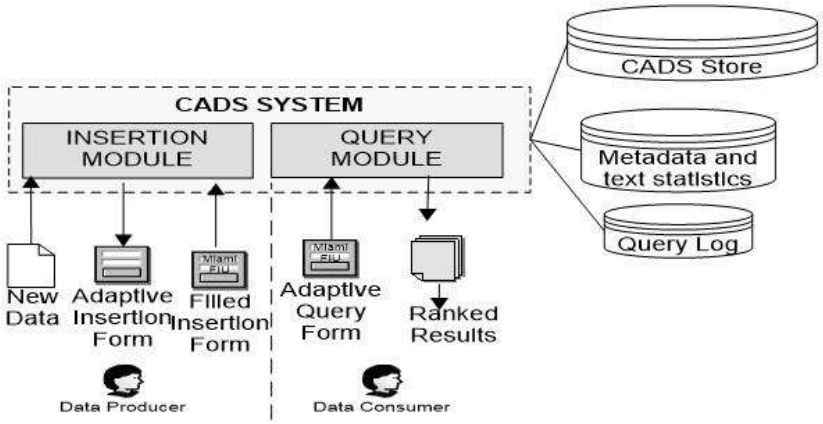

Fig. 1: CAD System

In our scenario, the author generates a new document and uploads it to the repository. After the upload, CADS analyzes the text and creates an adaptive insertion form. 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.

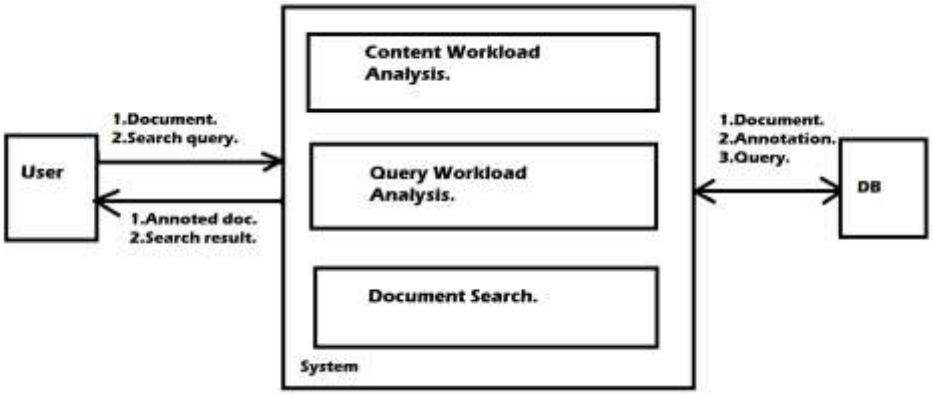

Fig.2: Proposed System Architecture

Flow of the proposed system

1. User first selects the document to upload it on the server. Before uploading the actual document our system analyze the document and get informative data from it.
2. To get data in annotation form in key and value pair.
3. To analyze the data we first use STOP word method.
4. After STOP word we use STEMMER method to filter data
5. After this we calculate the frequency count.
6. Then we apply Bayes algorithm to suggest annotations from filtered data.
7. After this we generate a CAD form (Collaborative Adaptive Data) which is having annotations suggested by the system. Along with the system suggestions user can add his own annotations for particular document before uploading. These annotations help us to find same document when we search it. While searching, users fire some queries; these search queries are registered by our system and feed to Bernaulli Algorithm to querying value analysis. Later result of Bernaulli's algorithm is also used to suggest annotations.

IV. ALGORITHM

Information Extraction Algorithm:

Step 1 : Select a text file for extraction.
Step 2 : Parse the text file. Ignore stop words 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.
Step 3 : Upload the file on server.
Step 4 : Then fill all the annotations which are relevant to the document which can be useful for query based searching.

The key contribution of this work is the "attribute suggestion" problem, which accounts for the query workload, and identifies the attributes that are present in the document, but not their values. There are two conflicting properties for indentifying and suggesting attributes for a document d. The attribute must have high querying value (QV) with respect to the query workload W. The attribute must have high content value (CV) With respect to d.

QV, CV Computation and Combining Algorithm:

Step 1 : Enter the queries for retrieving the document
Example: location='Pune' and year=2010
Step 2 : Split the queries and pass it to database for retrieving
Step 3 : Check all related results and show the related results to user.
Step 4 : For much efficient and accurate results, users should try to enter maximum queries they can.

V. EXPERIMENTAL WORK & MATHEMATICAL MODELING

In our project first we create GUI in net beans. We have create login ,registration, upload data, CAD form, Search and Result form with proper connectivity. We use JSON Parser for parse the data. We have use java platform for front end. And for database we use windows 7 OS. We have use Amazon review dataset.

Mathematical/logical model of proposed system using

β = { UI , O , Pr }

Where ,

β = Final System Model for User
UI = User Input
O = Output
Pr = Process

Where,

UI = **{** F , De , η , CAD , Q**}**

Where,

F = File / Documents
De = File Description
η = Annotations
CAD = Collaborative Adaptive Data Form
Q = User Queries
Pr = **{ STE , STO , Ba , Be , F , Ob}**

Where,

STE = Stemmer Algorithm Process
STO = Stop words algorithm
Ba = Bayes Method
Be = Bernaullie's Method
F = Frequency Count
Ob = Processed Objects
O = **{** SR , PD **}**

Where,

SR : Searched result
PD : Processed Document

VI. CONCLUSION

Our system provides solution to annotate the document at time of uploading and also works on user's querying needs. Our proposed architecture works on the content of document and also analyze the user queries. User queries and document content are the two basic source to generate the annotation. Along with annotation document pattern mining is the technique that helps the user to map document with frequent pattern and use pattern at the time of searching. The annotation and pattern matching technique provides flexible and complete solution for document tagging and searching

Acknowledgements

We would like to thanks all anonymous researchers, scientist, inventors and engineers.

REFERENCES

Journal Papers:

[1] Eduardo J. Ruiz, Vagelis Hristidis, and Panagiotis G. Ipeirotis,"Facilitating Document Annotation Using Content and Querying Value", *IEEE TRANSACTIONS, VOL. 26, NO. 2,* FEBRUARY 2014.

[2] R. Fagin, A. Lotem, and M. Naor, "Optimal aggregation algorithms for middleware," *J. Comput. Syst. Sci., vol. 66, pp. 614–656,* June 2003.
[3] S.R. Jeffery, M.J. Franklin, and A.Y. Halevy, "Pay-as-You-Go User Feedback for Dataspace Systems," *Proc. ACM SIGMOD Int'l Conf. Management Data,* 2008.
[4] J.M. Ponte and W.B. Croft, "A Language Modeling Approach to Information Retrieval," *Proc. 21st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '98), pp. 275-281,*
[5] R.T. Clemen and R.L. Winkler, "Unanimity and Compromise among Probability Forecasters," *Management Science, vol. 36, pp. 767-779,*
[6] C.D. Manning, P. Raghavan, and H. Schu¨ tze, Introduction to Information Retrieval, first ed. Cambridge Univ.
[7] M. Franklin, A. Halevy, and D. Maier, "From Databases to Dataspaces: A New Abstraction for Information Management," *SIGMOD Record, vol. 34, pp. 27-33,*
[8] J. Madhavan et al., "Web-Scale Data Integration: You Can Only Afford to Pay as You Go," *Proc. Third Biennial Conf. Innovative Data Systems Research (CIDR),* 2007.
[9] M.J. Cafarella, J. Madhavan, and A. Halevy, "Web-Scale Extraction of Structured Data*," SIGMOD Record, vol. 37, pp. 55-61,*
[10] M. Jayapandian and H.V. Jagadish, "Automated Creation of a Forms-Based Database Query Interface," *Proc. VLDB Endowment, vol. 1, pp. 695-709,*
[11] Vagelis Hristidis, Eduardo Ruiz," CADS: A Collaborative Adaptive Data Sharing Platform", *School of Computing and Information Sciences, Florida International University.*
[12] R. Fagin, A. Lotem, and M. Naor, "Optimal aggregation algorithms for middleware," *J. Comput. Syst. Sci., vol. 66, pp. 614–656,* June 2003.