Document Annotation by Content and Querrying Value

Arati S. Kailuke
Bachelor Student, Computer Science & Engineering, DES's COET Dhamangaon Rly, India.

Dipika D. Kalsait
Bachelor Student, Computer Science & Engineering, DES's COET Dhamangaon Rly, India.

Tubasamer M. Pathan
Bachelor Student, Computer Science & Engineering, DES's COET Dhamangaon Rly, India.

Manisha D. Sakharkar
Bachelor Student, Computer Science & Engineering, DES's COET Dhamangaon Rly, India.

Abstract – The paper is mainly about the generation of the structured metadata by identifying documents which contain targeted information and this information is going to be subsequently useful for querying the database. We proposed adaptive techniques to suggest relevant attributes to annotate a document, 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.

Index Terms – Probabilistic framework, Collaborative Adaptive Data Sharing platform(CADS), datasets, annotations.

1. INTRODUCTION

There are many application domains where users create and share information; for instance, news blogs, scientific networks, social networking groups, or disaster management networks. Current information sharing tools, like content management software (e.g., Microsoft SharePoint), allow users to share documents and annotate (tag) them in an ad-hoc way. Similarly, Google Base [1] allows users to define attributes for their objects or choose from predefined templates. This annotation process can facilitate subsequent information discovery. In Dataspaces, users provide data integration hints at *query* time. The assumptionin such systems is that the data sources *already* contain structured information and the problem is to match the query attributes with the source attributes. Many systems, though, do not even have the basic "attribute-value" annotation that would make a "pay-as-you-go" querying feasible. Annotations that use "attribute-value" pairs require users to be more principled in their annotation efforts. Users should know the underlying schema and field types to use; they should also know when to use each of these fields. With schemas that often have tens or even hundreds of available fields to fill, this task become complicated and bulky. This results in data entry users ignoring such annotation capabilities. Even if the system allows users to arbitrarily annotate the data with such attribute-value pairs, this task is difficult to perform for users.

A 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 this, we are trying to prioritize the annotation of documents towards generating attribute values for attributes that are often used by querying users.

2. RELATED WORK

2.1 Collaborative Annotation:

There are several system that favor the collaborative annotation of objects and use previous annotations or tags for annotating new objects. There have been a significant amount of work in predicting the tags for documents or other resources (web pages, images, videos) [1], [2], [3], [4], [5]. Depending on the object and the user involvement, this approaches have different assumptions on what is expected as an input, Though the goals are similar as the expect to find missing tags that are related with the object. We argue that our approach is different as we use the workload to increase the document visibility after the tagging process. Compared with the other approaches precision is a secondary goal as we expect that the annotator can improve the annotations on the process. In another way, the discovered tags assist on the tasks of retrieval instead of simply bookmarking.

2.2 CAD:

Collaborative Adaptive Data Sharing platform (CADS), which annotate data as we create infrastructure that facilitates fielded data annotation. A 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. This project is to prioritize the annotation of documents toward generating attribute values for attributes that are used by querying users. The goal of CADS is to encourage and lower the cost of creating nicely annotated documents that can be useful for commonly issued semi structured queries. Our goal is to encourage the annotation of the documents at creation time, while the creator is still in the "document generation" phase, even though the techniques can also be used for post generation document annotation. In this, the author generates a new document and uploads it to the database. After uploading the document, 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.

3. PORPOSED MODELLING

In this paper we are performing document annotation by using CADS technique .The workflow of the CAD is as follows:

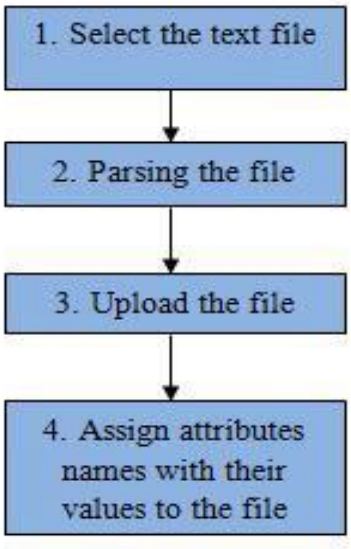

Fig 3.1: CADS PROCESS

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. In the rest of the paper the term data usually refers to a document; other types of data are also possible, but we focus on documents for simplicity. Fig. 3.2 presents a typical CADS workflow.

In the insertion phase the submission of a new document to be included in the database. After uploading the document, CADS analyzes the text and creates an adaptive insertion form with the set of the most probable ⟨attribute name, attribute value⟩ pairs to annotate the new document. The user fills this form with the required information and submits it. The final stage consists of the storage of the associated document and metadata in the CADS database.

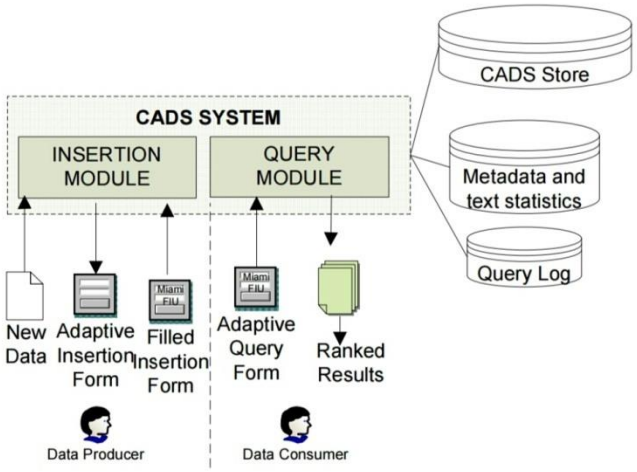

Fig: 3.2:CADS WORKFLOW

In the query phase, the adaptive query form is given to user which supports ⟨attribute name, attribute value⟩ conditions. Initially, before CADS has began learning the information demand through processing the query workload, the query form only specifies the default attributes.

4. RESULTS AND DISCUSSIONS

In this to address the problem of the probabilistic nature of the annotations, previous work on ranking under uncertainty must be adapted for the hybrid filter/ranking model of CADS querying. This incurs efficiency and scalability issues, which require execution algorithms to achieve real-time responses. We proposed CADS, a Collaborative Adaptive Data Sharing platform, which is a next-generation data sharing platform where the annotation and combination occur at both the data insertion (production) and querying (consumption) actions. A goal of CADS is to leverage the information demand to create adaptive insertion and query forms. We believe that CADS has a great potential to improve many collaboration environments, and hence it is beneficial to follow research directions that will allow the realization of CADS.

5. CONCLUSION

We proposed adaptive techniques to suggest related attributes to annotate a document, while trying to satisfy the user querying needs. The system is based on a probabilistic framework that considers the evidence in the document content and 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. This system suggest attributes that improve the visibility of the documents with respect to the query workload. That is, we show that using the query workload can greatly improve the annotation process and increase the utility of shared data.

REFERENCES

[1] R. Jeffery, M. J. Franklin, and A. Y. Halevy, "Pay-as-you-go user feedback for data space systems," in ACM SIGMOD, 2008.
[2] K. Saleem, S. Luis, Y. Deng, S.-C. Chen, V. Hristidis, and T. Li, "Towards a business continuity information network for rapid disaster recovery," in International Conference on Digital Government Research, ser. dg.o 2008.
[3] A. Jain and P. G. Ipeirotis, "A quality-aware optimizer for information extraction," ACM Transactions on Database Systems, 2009.
[4] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ser. SIGIR '98.
[5] R. T. Clemen and R. L. Winkler, "Unanimity and compromise among probability forecasters," Manage. Sci., vol. 36, pp. 767–779, July 1999.
[6] C. D. Manning, P. Raghavan, and H. Sch¨utze, Introduction to Information Retrieval, 1st ed. Cambridge University Press, July 2008.
[7] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on amazon mechanical turk," in Proceedings of the ACM SIGKDD Workshop on Human Computation, ser. HCOMP '10. New York, NY, USA: ACM, 2010.
[8] R. Fagin, A. Lotem, and M. Naor, "Optimal aggregation algorithms for middleware," J. Comput. Syst. Sci., vol. 66,pp. 614–656, June 2003.
[9] K. C.-C. Chang and S.-w. Hwang, "Minimal probing: supporting expensive predicates for top-k queries," in ACM SIGMOD, 2002.
[10] G. Tsoumakas and I. Vlahavas, "Random k-labelsets: An ensemble method for multilabel classification," in Proceedings of the 18th European conference on Machine Learning, ser. ECML '07. Berlin, Heidelberg: Springe Verlag, 2007.

Authors

Arati S. Kailuke

Final year Computer Science & Engineering student from DES's COET Dhamangaon Rly, India.

Dipika D. Kalsait

Final year Computer Science & Engineering student from DES's COET Dhamangaon Rly, India.

Tubasamer M. Pathan

Final year Computer Science & Engineering student from DES's COET Dhamangaon Rly, India.

Manisha D. Sakharkar

Final year Computer Science & Engineering student from DES's COET Dhamangaon Rly, India.