Survey on Unstructured Document Annotation using Content and Query Value Based

Pranjali Raut*  
M. Tech. (CSE)  
VMIT, Nagpur, India

Gurudev B Sawarkar  
Associate Professor (CSE)  
VMIT, Nagpur, India

Abstract

A large data is generated in different organization which is in textual format. This contain eloquent amount of structured information, which dwell underlying in the unstructured text. Document annotation is the task of adding metadata information in the document which is useful for information extraction. Many information mining algorithms facilitate the extraction of structured information from raw data but which is costly, inefficient and also shows contaminated results. We present, an adaptive technique that Collaborative Adaptive Data Sharing platform (CADS) for document annotation and use of query workload to direct the annotation process. Paper proposes survey on facilitates the generation of structured metadata by identifying documents containing information of interest. Such information is further useful for querying the database. Experimental assessment shows superior results compared to approaches that rely only on the textual content or only on the query workload, to recognize attributes of importance. Paper is a survey on document annotation techniques.

Keywords: Annotation, CADS, METADATA adaptive technique.

*Author for correspondence pranjalir8@gmail.com

1. Introduction

Organizations generate huge amount of unstructured data. Advanced growth in data collection and storage technology made it possible to arrange this data at lower cost. Our goal is exploiting this stored data, in order to extract useful and actionable information. To get summarized search information is our requirement and to get this we arrange data in smart way. Annotation is one of the best techniques to arrange significant information and get effective search result. Annotations of documents are comments, notes, explanations, or other types of external remarks that can be attached to a Web document or to a selected part of a document. As they are external, it is possible to annotate any Web document independently, without needing to edit the document itself. From a technical point of view, annotations are usually seen as metadata, as they give additional information about an existing piece of data. Annotations of documents can be stored locally or in one or more database servers. When a document is searched, content of queries value each of these database servers, requesting the annotations related to that document in Web server database. There are many annotation techniques are present that are based on attribute...
value pair. The strategies based on attribute value pair are effective method of document annotation. But there is restriction that document should be in structured format when using these systems. Also user has internal knowledge of attributes of document, as there are number of attributes because of them it will be difficult and infeasible to identify such attributes and its difficult approach to facilitate document annotation. Hence algorithm should focus on those documents that contain words that are used during query. If we ignore contents of document then it will be unable to find out required information that’s why document feature extraction is done on documents. We proposed Collaborative Adaptive Data Sharing Platform (CADS) which is an “annotation as user created”. It assists fielded data annotations. In this the direct use of key contribution for the query workload by using direct annotation process and also 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 where in users will have to deal only with relevant results.

2. Literature Review

Jeffery et al. proposed pay-as-you-go user feedback for data space systems 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 they assumed that data sources already contain structured information and the problem is to match the query attributes with the source attribute. Saleem et al. 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. They proposed a disaster management model which works well at some extent but it is not considering the effective retrieval. Clemen et al. worked on probabilities of particular uncertain event which helps us to find out annotation and attributes. Franklin et al. proposed a solution to Laplace smoothing to avoid zero probabilities for the attributes that do not appear in the workload. It helps us to converge towards accuracy. Tsoumakas & Vlahavas proposed an ensemble method for multi-label 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.

Heymann et al. proposed a solution for prediction of tags for particular object. We can adopt this for out suggesting annotation concept. Song et al. worked on real-time automatic tag recommendation which exactly works with the same way we want for out document annotations. Ponte & Croft proposed 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. Eck et al. in their paper “Automatic Generation of Social Tags for Music Recommendation” promoted same kind of auto suggestions of tags. But this was
dedicated to the musical data. Sigurbjornsson & Zwol in their paper “Flickr Tag Recommendation Based on Collective Knowledge” suggested tags for images /snapshots on Flickr and guides us for web based system structure tag recommendations. Jain & Ipeirotis presented 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.

Wu et al. used technique of closed sequential patterns in text mining. It improves the performance of text mining. Pattern taxonomy model is developed to improve the effectiveness. It uses closed patterns in text mining effectively. Term-based methods and pattern based methods is used to improve the performance of information filtering. Yin et al. suggested social tagging by incremental process. It proposes probabilistic models. Probabilistic tag recommendation systems were introduced. It uses Bayesian approach. It is focusing only on content and not the query workload that reflects the user interest. Russell et al. proposed 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. Chen et al. focused on system for form design, data entry and data quality assurance. Using existing data set of form, USHER derives a probabilistic model using the questions of the form. It is closely related to CAD form in our system. Using Usher we can identify dependencies across attributes. Jayapandian & Jagadish in their papers “Automated Creation of a Forms-Based Database Query Interface” and “Expressive Query Specification through Form Customization” found that CADS, a technique to extract query forms form existing queries in a dataset that are fires on database using 'querability' of column. Miah et al. extracted algorithm based on Integer Programming formulation of the problem. It takes significant amount of time for processing for small workload but provide optimal and nearest solution.

3. Scope of Study
The project is designed with the fundamental knowledge in data mining, basics of unstructured data & natural language processing concept. The programme is developed using probabilistic methods & algorithms to integrate information from query workload. The main use of our system is mainly that when users of author perform query based search, they could get minimum and distinct accurate results where it could be easy for retrieval data from the database.

4. Objectives of Study
The primary aim of this project is to develop 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. We create principled probabilistic methods and algorithms to seamlessly integrate information from the query workload into the data annotation process, in order to generate metadata that are not just relevant to the annotated document, but also useful to the users querying the database. We present extensive experiments with real data and real users, showing that our system generates accurate suggestions that are significantly better than the suggestions from alternative approaches.
5. Research Methodology

This system suggests CADS, an annotation-as-you-create infrastructure that facilitates fielded data annotations. The goal of CADS is to lower the cost creating annotated documents that can be useful for common semi-structured queries. Figure 1 represents the workflow 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.

![Fig. 1: CADS Workflow](image)

In proposed system, the author generates a new document and uploads it in storage. After uploading the document, CADS analyses the text and creates adaptive insertion form as shown in Figure 2. The form contains the best attribute names which are present in the document and information needed for query workload and most probable values of the attributes given in the document. The author has ability to check the form, modify the metadata if it is necessary and finally submit the document for storage.

![Fig. 2: Adaptive insertion form](image)

While retrieving attribute names, the adaptive insertion form also retrieves the attribute values by using IE (Information Extraction) Algorithm. In order to retrieve contains of the text file information extraction (IE) algorithm is used.

A. 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 exclusive 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 $d$.

**B. QV, CV Computation and Combining Algorithm**

Step 1: Enter the queries for retrieving the document.

Example: location='Mumbai' and year=2012.

Step 2: Disjoin the queries and pass it to database for retrieving.

Step 3: Check all annotated 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.

**C. MODULES**

- Registration
- Login
- Document Upload
- Search Techniques
- Download Document

![Fig. 3: IE Algorithm](image)

**6. Result and Discussion**

We 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. We create principled probabilistic methods and algorithms to seamlessly integrate information from the query workload into the data annotation process, in order to generate metadata that are not just relevant to the annotated document, but also useful to the users querying the database. We present extensive experiments with real data and real users, showing that our system generates accurate results that are significantly better than the results from alternative approaches & with respect to the query workload by up to 70%. It also improves the efficiency of searching.

**7. Conclusion**

This paper surveys work related to document annotation using content and querying value. This project is proposing adaptive methods to suggest relevant, recommended attributes to annotate a document while also trying to satisfy the user querying, searching needs. Our solution is based on a probabilistic framework that views the confirmation in the document content and the query workload. The main advantage of our application is mainly that when users perform query based search, they could get minimum and distinct outcomes/results where it could be easy for
retrieval. Experiments shows using these techniques, workload of application can reduce by large amount. It improves the annotation process and visibility of documents.

References