A SURVEY ON FACILITATING DOCUMENT ANNOTATION USING CONTENT AND QUERYING VALUE

Nupoor Gade ¹, Rugraj ²

¹ PG Student, Dept of Computer Engineering, Alard college of engineering and management,
² Professor, Dept of Computer Engineering, Alard college of engineering and management,
Savitribai Phule Pune University, Pune

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, Adaptive technique, Collaborative platform, CADS, Information extraction

[1] 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.
[2] RELATED WORK

Currently available information sharing tools, like content management software annotate document in an ad hoc way. For Google Base, there is predefined template available, which facilitates subsequent information discovery.

Some systems do not have attribute-value annotation would make querying feasible. An annotations strategy that uses attribute-value pairs contains more information than untyped approaches which are more expensive. For such annotations user must be aware of using and applying annotations. In such cases users are not ready to perform the task though system allows user to perform required task.

Such annotations are limited to simple keywords making the analysis and querying data of the data cumbersome. So, there is need of appropriate annotation of the document.

[3] PROPOSED WORK

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 semistructured queries. [Figure-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.

```plaintext
In proposed system, the author generates a new document and uploads it in repository. 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.
```

N. Gade and Rugraj
While extracting attribute names, the adaptive insertion form also extracts the attribute values by employing IE (Information Extraction) Algorithm. In order to extract contains of the text file information extraction (IE) algorithm is used.

**Information Extraction Algorithm:**

Step 1: Select a text file for extraction.
Step 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.
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.

[4] CONCLUSION

This paper surveys work related to document annotation using content and querying value. This paper also surveys Collaborative Adaptive Data Sharing platform (CADS) for fielded data annotation. Proposed system aims to minimize the cost of annotating documents. The advantage of proposed system is query based searching.
REFERENCES


