Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 85 (2016) 45 – 53

Procedia
Computer Science

International Conference on Computational Modeling and Security (CMS 2016)

DSPAA: A Data Sharing Platform with Automated Annotation

Keerthana.I.P[a,*], Aby Abahai.T[b]

[a]PG Scholar, Dept. of CSE, Mar Athanasius College of Engineering, Kothamangalam-686666, Kerala, India
[b]Assistant Professor, Dept. of CSE, Mar Athanasius College of Engineering, Kothamangalam-686666, Kerala, India

Abstract

Document annotation and search are two important factors that need to be considered in data sharing platforms. A large unstructured text document contains substantial amount of structured attribute information. Important information is very difficult to find in these documents. Current ad-hoc or predefined annotation of the shared data causes inadequate search, retrieval and analysis capabilities. In this paper we propose a new approach that supports the generation of the structured annotation in the form of attribute name and attribute value pairs from unstructured document. A new data sharing platform DSPAA (Data Sharing Platform with Automated Annotation) is proposed, where the document annotation occurs when the author uploads a document and it is based on a probabilistic framework that considers the attributes in the document content and the query collection. The new system also performs semantic annotation of document using WordNet database. When a user submits a search query, then the system will search for documents in the annotation database and rank the selected documents by using Vector Space model. From experiment results it is clear that the system generates superior results at a rate faster than traditional document retrieval strategies.

Keywords: Content Score; Document Annotation; Document Ranking; DSPAA; Query Score; Semantic Annotation; Vector Space Model; WordNet

1. Introduction

The metadata or annotation is such data, which describes other data or gives information about the data. For example, letters or characters in a text are data, but the number of letters in a text is the metadata. If a narrower sense is implied and the term is used in connection with file types, the metadata means, for example, such information as

* Corresponding author. Tel.: +91-8593916878.
*E-mail address:*keerthanaprakash3@gmail.com

doi:10.1016/j.procs.2016.05.175

the name and the title of a file, its author, keywords to the contents of a file or the date of the saving. An obvious positive effect of metadata in files is: they allow cataloguing and browsing of data according to certain general criteria in a simpler and more precise way.

Data management tools like Microsoft's SharePoint permit users to share documents and tag them for some specific case. SAP NetWeaver permit users to annotate documents, share and do simple keyword based queries. Similarly in Google Base, users can specify their own <attribute name, attribute value> pairs in addition to the ones proposed by system. But, suggested attributes in Google Base are fixed for each category. This fixed or predefined annotation of the shared data causes problems like schema explosion or inefficient data annotation, which in turn guide to unsatisfactory analysis and search performances. A scenario is complicated where the author has to fill a number of fields at time of uploading a particular document. Hence users often avoid such annotations. Such problems results in very simple annotations that is often limited to small keywords. Users are often limited to plain keyword searches, or have access to very basic annotation fields. Such annotations cause the analysis and querying of the data complicated. Annotations which use <attribute name, attribute value> pair requires users to be more principled in their annotation work. Here users must have clear idea in applying and using the attributes.

Data Sharing Platform with Automated Annotation (DSPAA) is a new document sharing platform which performs annotation of document during document upload phase and supports fielded data annotation. This system uses the query collection to direct the annotation process, in addition to inspecting the content of the document. The goal of DSPAA is to create efficiently annotated documents that can be useful for usually issued semi-structured queries of user. This paper is structured as follows: in section 2, the works related to document annotation is discussed. Section 3 presents the implementation details of the proposed system. In section 4 performance analysis of the proposed system is discussed. Finally the section 5 gives the conclusion for the work.

2. Related Works

The information management challenges today stem from organizations based on a large number of heterogeneous, related data sources, but having no way to manage their dataspaces [1] in a convenient, integrated, or principled fashion. They propose dataspaces and their support systems are used for data management. A common problem of database systems is that they are hard to query for users discomfort with a formal query language. To handle this problem in [2],[3] form-based interfaces and keyword search have been proposed, combining the two for creating an approach that provides best result. At query time, a user issues standard keyword search queries, instead of returning tuples, the system returns forms relevant to the search query. The user then creates a structured query by using one of these forms and submits it to the system. With large number of data sources available over the web integration of them is an important problem. Integration of the hidden sources is integration of their query interfaces. An interactive, clustering-based approach to matching query in interfaces is discussed in [4].

There are huge amounts of text on the internet that are neither grammatical nor formally structured. These sources of data, called posts are full of useful information for agents searching the Semantic Web, but they miss the semantic annotation to make them searchable. By leveraging their common attributes, called reference sets, it can annotate these posts despite their lack of grammar and structure [5]. In [6], a method for semantic annotation of web pages is introduced and performed semantic annotation by using web patterns. This method is based on extraction of patterns, which are characteristic for a particular domain. They have annotated pages in a database with regard to patterns so there is information about which patterns are contained on each page.

In [7], highlights the challenges in two scenarios – the Deep Web and Google Base. Traditional data integration techniques are not valid in the case of such heterogeneity and scale. They propose new data integration architecture called PAYGO, which is based on the concept of dataspaces and emphasizes pay-as-you-go data management for achieving web-scale data integration. In [8] describe three recent extraction systems that can be operated on the entire Web. In [9], a tool called KMAD which tells the quality of document or its usefulness based on annotations is presented. Collective sentiments of annotators are classified as negative, positive and objectivity. In [10], introduces the use of Wikipedia for automatic keyword extraction and word sense disambiguation. Given an input document, the system identifies the important concepts in the text and automatically links these concepts to the corresponding Wikipedia pages. In [11],[12] CADS, a Collaborative Adaptive Data Sharing platform, where the information demand of the community is exploited to annotate the data at insertion-time. In [13], describes GoNTogle, a framework for document annotation and retrieval, built on top of Semantic Web and IR technologies. GoNTogle supports ontology-based annotation [14] for documents of several formats, in a fully collaborative environment.