ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 3, March 2014

A Survey of Text mining Using User Required Information

Vikash Kumar Singh[1], Prof.(Mrs.) L.H.Patil[2]

Student, M.Tech(CSE) PIET, Nagpur.India[1]

Assistant Professor, CSE Dpt. PIET, Nagpur.India[2]

Abstract: Now days, many companies for the sake of promoting their product and services sham the respective services content of their products amongst the organizations. This textual content contains significant amount of structured of useful information which is buried by the unstructured text or content. And therefore the text mining from that unstructured data is not effective and gives the ineffective results on to the clients and hence affects the business. In a typical organization, only 20% of the information that exists is well formed – living in relational databases or legacy mainframe transactional systems. This information which we refer to as structured information. In the same typical organization, 80% of the information that exists is often unstructured information. Our approach relies on the idea that many structured and useful information arrives at the time of implementation of subject and if this information is mentioned in our sharing content then this help in effective text mining and obtain the data of interest and desired results.

General Terms: Unstructured data, structured data, structured information retrieval, extracted data search model.

I. INTRODUCTION

A large number of organizations recently generate and share textual descriptions of their products, and services, such collections of textual data contain significant amount of structured, information, which remains buried in the unstructured document. Large companies may have presences in various places, each of which generate a large volume of unstructured data. For example, insurance agencies may have data from thousands of local branches. Further, large organizations have complex data structure with or without schemas.

Unstructured data can take many forms like word documents, spread sheets, email messages, blogs, pictures, movies. Unstructured data by nature is raw data, data mining or "analysis" of the UD to arrive at the results or statistics that will be placed in the structured world equivalent to business rules. In my opinion, they should unstructured data mining should contain the document information, and possibly a few other key notions. The mining engine should be capable of "clustering" terms together to form an idea, a context. Data mining is the process of semi automatically and analyzing large databases to find useful patterns. Data mining process attempts to discover rules and patterns from the data. The Unstructured data analysis and mining is much more than this. Unstructured Data can be scattered, complex and different structures, different schemas. The tools available for data mining techniques may or may not be very useful to extract and represent the structured information out of unstructured data.

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**.** 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 allows users to define attributes for their objects. This annotation process can provides subsequent information discovery. Many annotation systems allow only "untyped" keyword annotation for instance, a user may annotate a Lather report using a tag such as "Storm Category. The information is because of the sheer volume of structured and unstructured information that is available. There is too much information, even within a department or team, for any one person to keep on top of it all. To deal with the information glut, users need to be able to filter and personalize the information that is relevant to them. For example, a salesperson might care only about information related to his or her target prospects and customers. This might include internal information, proposals, financial information, sales history, credit information, product information, buying patterns, as well as external information about competitors in his accounts, etc.

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 umber some. 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, the users are often unwilling to perform this task: The task not only requires considerable effort but it also has unclear usefulness for subsequent searches in the future: who is going to use an arbitrary, undefined in a common schema, attribute type for future searches? But even when using a predetermined schema, when there are tens of potential fields that can be used, which of these fields are going to be useful for searching the database in the future?

Such difficulties results in very basic annotations, if any at all, those are often limited to simple keywords. Such simple annotations make the analysis and querying of the data. Users are often searches, or have access to very basic annotation fields. There are clear benefits to bridging the gap between structured and unstructured data within the enterprise and presenting it to end users in the appropriate context. However, it is often easier said than done. There are a number of technical reasons why it is very difficult to achieve this integration. There are also a number of less technical and more cultural reasons, both historical and priority-related, that have stymied this integration. In many cases, there are separate, on-going efforts to integrate the structured data with other structured data, and to integrate the unstructured content with other unstructured content. Going forward, we must take a different approach to managing enterprise information.

II. RELATED WORK

2.1 Extracting semantic annotations and their correlation with document components:

Digital document can preserve of information in the form of digital content. Searching this digital content requires time and computing resources. These Techniques are required to efficient process these digital documents. This Metadata and semantic annotations can augment the overall search process and provide a foundation to build intelligent applications by using the documents in the repository. In this paper, I am proposing an approach for generation of context aware metadata to enhance search for the scientific publications and also prove the impact of compound words on semantic metadata. Our main contribution of our work is to correlate these structured extracted semantic annotations information with the document components. This process allows for accessing the document. for example, searching a document centered around a scientific claim by differentiating be taken author's claims and statements about related systems mentioned in different document components. The approach utilizes the syntactic and semantic measures to increase the quality of the extracted semantic annotations and to bring improvements in precision of search results.

2.1.A: Ranking and scoring semantic document annotation:

Semantic makes computer understands meaning of queries. This state of technology will assist human in querying rich documents based on their intention. I define rich document as semantic document in terms of its description, which contains exact statements and related statements. Sometimes, some of the search engines are lack of ranking and scoring features.

In this paper, we modify the algorithm to ranking and scoring the semantic document annotation based on document richness. We apply the modification algorithm into a research prototype retrieval engine, Pico Doc, to experiment its ability in ranking and scoring documents Documents annotation. The result shows a modified with related spreading concept yields promising results in retrieving related annotated document.

2.2 Semantic Multimedia Document Adaptation with Functional Annotations:

The diversity of presentation contexts for multimedia documents requires the adaptation of document specifications. In this work, we have proposed a semantic adaptation framework for multimedia documents. This framework covers the semantics document of the document composition and transforms the relations be taken multimedia objects according to adaptation constraints. In this paper, I show that relying on document composition alone for adaptation restricts the set of relevant candidate solutions and may even divert the adaptation from the author’s intent. Hence, I propose to introduce functional annotations to guide the adaptation process. Theses annotations allow refining the role of multimedia objects in the document. I show that SMIL documents could embed functional annotations. These multimedia documents are then adapted thanks to an interactive adaptation tool.

2.3 Advances in collaborative annotation in semantic management environment:

Providing solutions to problems associated with mythological creation, management and information extraction search in an annotation archive is the core of this study. Information extraction from unstructured archives grows at a relatively slow space but annotations associated with archives grow geometrically because of the diversity of reflections on documents emanating from different authors and with time. Information annotation by creator of document is generally connected to a definite document, specific individuals or a single time. Annotation can be seen as an informal way for individuals who do not freely have initial rights for a document to "publish" their thoughts on a subject of interest. Publishing one's thoughts using annotations does not involve publication protocols such as copyright issues. Where there is freedom of expression through annotation, the flexibility and frequencies of "publishing" one's views on a subject are bound to increase. This flexibility and simplicity in expression entails a systematic management of an annotation archive.

The creation of an annotation database is often seen as the human activity that can embed the function of its creator (who is also a document user), the original document and time. It means that a database of annotations based on three parameters (creator, document and time) may include divergent annotations as a result of multiple documents and human factors. With participation of diverse users, there can be divergent interpretations of subjects of interest based on varying thoughts of users. With change of time, a user's opinion on a subject can change. The question that quickly comes to mind is how can a database growing geometrically, with divergent reflections (annotations), by divergent users with considerable length of time be created and searched effectively in a collaborative environment?
We consider creation and the exploration of an annotation database by combining the concept of semantic technology with the topic maps data model. Each word - - used by users in annotation creation benefits from the potential of semantic technology based on topic maps to resolve the difficulty in management. More precisely, our attention in this study is the creation and exploitation of annotation databases to improve information research. Our TMSUMS platform benefits of combining the SUMS-based semantic logical model with the topic maps-based semantic physical data model. As one of the key issues, we brought to light, the problem related to annotation creation in a collaborative environment. Thereafter, we introduce scenarios of information search in an annotation database constructed on specific parameters. we demonstrate how difficult it to search for meaningful information from such an annotation archive/database in a normal situation. Our proposal is a search through such a database with the concepts of semantic technology and topic maps data model to demonstrate how such a search can be improved. Our conception concludes how several elements of such annotation system illustrate how to build semantic management based on topic maps the data model. We figure out how annotation management can be improved following this approach.

III. PROPOSED METHODOLOGY

In this paper, we propose CADS Collaborative Adaptive Data Sharing platform),which is an "annotate-as-you-create" infrastructure that facilitates fielded data annotation .A key 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 other words We are trying to prioritize the annotation of documents towards generating attribute values for attributes that are often used by querying users.

Cads Objective:

CADS stand for Collaborative Adaptive Data Sharing platform

1. Facilitates effective and effortless data annotation at insertion-time Leverages these annotations at query-time
2. Learns with time the information demand which is then used to create adaptive insertion and query forms.

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: a model that considers both components conditionally independent and a linear weighted model. In this we first convert the untrusted document into structured format by using following algorithms.

- Snowball techniques
- Proteus techniques
- Known tall techniques.

For searching these documents from database we used following searching techniques.

- Top-k ranked document search algorithm.
- Shark search algorithm

IV. CONCLUSION

Text data contains some valuable information which are remains buried in the unstructured doc. Unstructured data not be fitted into relational table. We explore a technique for extracting such table data that is easy to accessible to the users. On the basis of our experiment we found that the execution of this system is more efficient than previous one.

REFERENCES

[1] Google,"Googlebase, http://www.google.com/base," 2011.
[2] S. R. Jeffery, M. J. Franklin, and A. Y. Halevy, "Pay-as-you-go user\feedback for data space systems," in ACM SIGMOD, 2008.
[3] 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 '08, 2008.
[4] A. Jain and P. G. Ipeirotis, "A quality-aware optimizer for information extraction," ACM Transactions on Database Systems, 2009. 21st annual international ACM SIGIR conference on Research and development \in information retrieval ser. SIGIR '98. New York, NY, USA: ACM, 1998, pp. 275–281.
[5] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in Proceedings of the
[6] M. Shepherd. Extracting Meaningful Metadata. DRTC Workshop on Semantics, 2003.
[7] K. Cardinaels, M.E. Duval. Automatic Metadata Generation: The Simple Indexing Interface, IW3C, 2005.
[8] X. Hu, T. Young Lin, Il-Y. Song, X. Lin, I. Yoo, M. Lechner, M. Song. Ontology-Based Scalable and Portable Information Extraction System to Extract Biological Knowledge from Huge Collection of Biomedical Ib Documents. Proceedings of the IEEE/WIC/ACM International Conference on Ib Intelligence., 2004.
[9] B. L. Grand and M. Soto. Visualisation of the semantic Ib: Topic maps visualisation. volume 00, page 344, Los Alamitos, CA, USA, 2002. IEEE Computer Society. R. H. H. The topic maps handbook. In Empolis Arvato Knowledge Management. Getersloh, Germany, 2003.
[10] V. D. Krötzsch Markus and V . Max. Wikipedia and the semantic Ib: The missing links. In Wikimania, volume 43, pages 907.928, Frankfurt, Germany, 2005.
[11] Iyengar S. S. and Brooks R. R "Distributed Sensor Networks" CRC press 'Inc., pp 1188 Dec 28, 2004.