ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

NLP Based Document Annotation Using Content and Query Value

Jayalakshmi, Liji Samuel

Post Graduate Student, Dept. of CSE, Sree Buddha College of Engineering for Women, Elavumthitta, Pathanamthitta, Kerala, India
Assistant Professor, Dept. of CSE, Sree Buddha College of Engineering for Women, Elavumthitta, Pathanamthitta, Kerala, India

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. Here, 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 work proposes Collaborative Adaptive Data sharing platform (CADS) for document annotation and use of query workload to direct the annotation process in a disaster management system. Here a document can be annotated using the content value (attribute value) and query value (attribute name) by the user and the user can view or download the annotated document. These are done using the process such as stopword removal, stemming, tokenization and workload extractor. Here, NLP is used to make the annotation process easier. Another contribution is that the query-based searching can be applied on other file formats like .docx, .Pdf which can greatly improve the annotation process and increase the utility of shared data. This can surely give a huge boost to mainly in text mining which can be thought of as a changing trend or technology.

KEYWORDS: document annotation; information retrieval; structured data; NLP; content value; query value

I. Introduction

Organizations today encounter textual data while running their day to day business. Electronic text, call centre logs, social media, corporate documents, research papers, application forms, service notes, emails, etc. are the sources of data. This data may be accessible but remains untapped due to the lack of awareness of the information wealth an organization possesses or the lack of methodology or technology to analyse this data and get the useful insight. Any form of information that an organization possesses or can posses is an asset and can get insight about its business by exploiting this information for decision making. It holds information about their customer, partners and competitors. The data about customers could give them insight about how to provide better services to its customers and increase their customer base. At the same time its partners can provide insights about how to maintain better relationships with its partners and forge new and valuable relationships. The data about its competitors can help them stay ahead of its competitors. However, not all the data that an organization possesses is tapped to get these insights. The reason being that major portion of data is in the unstructured form and its processing will be very difficult; the way structured data is processed to get the useful and desired insight. Therefore several techniques can exploit this potential by uncovering hidden value from this data. This is where text mining techniques find its value and can be helpful in discovering useful and interesting knowledge from this data. Businesses use such techniques to analyse customer and competitor data to improve competitiveness.

Due to problem of discovering important information from the data deluge that the organizations are facing today, text mining is essential. The origin of Text mining is often considered to have from data mining; however a few of the techniques have come from various other disciplines like information science and information visualization. Text mining strives to solve the information overload problem by using techniques from data mining, machine learning, NLP, Information Retrieval, and Knowledge Management.

Annotating a document is the comments, text and explanations which are added to the part of the document or the whole document [1]. User can create and share the necessaryinformation in many application purpose domains includingsocial networking site, Disaster Management System. The Microsoft share point, which is an Informationsharing tool allow users to create and share the document and also to tag them. In the same way user can

DOI: 10.15680/IJIRCCE.2015. 0311033

define attributes for their objects in Google base [2]. Hence annotation process can helpful for information discovery. Many annotation systems already contain "untyped" keyword annotation: for example, a user may tag the weather report using the annotation "Storm Name Kathreena". Attribute-value pairs for annotation are expensive because they contain more information compared to untyped approach. In this scenario, the above information can be rearranged as (Storm Name, Kathreena).

Many systems, though, do not even have the basic "attribute-value" annotation that would make a "pay-as-you-go" [3] querying feasible. Annotations that use "attribute-value"pairs require users to be more principled in theirannotation efforts. The schema and field types to use should be known by the users; they should also know whento use each of these fields.This results in dataentry users ignoring such annotation capabilities. Even ifthe system allows users to arbitrarily annotate the data withsuch attribute-value pairs, the users are often unwilling toperform this task. The task not only requires considerableeffort but it also has unclear usefulness for subsequent searches in the future.

To solve the problem, "annotate-as-you-create" infrastructure is used that facilitates the fielded data annotation. This proposed system is CADS (Collaborative Adaptive Data Sharing) platform. A key goal of system is the direct use of the query workload to direct the annotation process, in addition to examining the content of the document. Here a document can be annotated using the content value (attribute value) and query value (attribute name) by the user and the user can view or download the annotated document. These are done using the process such as stopword removal, stemming, tokenization and workload extractor. Here, NLP is used to make the annotation process easier. Another contribution is that the query-based searching can be applied on other file formats like .docx, .Pdf which can greatly improve the annotation process and increase the utility of shared data. This can surely give a huge boost to mainly in text mining which can be thought of as a changing trend or technology.

II. Related work

Social tags have recently emerged as a popular way to allow users to contribute metadata to large and dynamic corpora. Social Tag Prediction [4] system consists of users u $\in$U, tags t $\in$T, and objects o $\in$O. Then call an annotation of a set of tags to an object by a user a post. A post consists of one or more t_i, u_j, o_k triples. Here imagine that every object o has a vast set of tags that do not describe it, a smaller set of tags which do describe it, and an even smaller set of tags which users have actually chosen to input into the system as applicable to the object. The first set of tags negatively describes the object, the second set of tags positively describes the object, and the last set of tags currently annotates the object. Social tags are user generated keywords associated with some resource on the Web. Automatic Generation of Social Tags for Music Recommendation [5] was proposed by Paul Lamere, Stephen Green. In the case of music, social tags have become an important component of "Web2.0" recommender systems.It allow users to generate playlists based on user-dependent terms such as chill or jogging that have been applied to particular songs. A set of boosted classifiers are used to map the audio features onto social tags collected from the Web. The resulting autotags furnish information about music might be untagged or poorly tagged and it allows for insertion of previously unheard music into a social recommender. This avoids the "cold-start problem" common in such systems. Autotags can also be used to smooth the tag space from which similarities and recommendations are made by providing a set of comparable baseline tags for all tracks in a recommender system. Online photo services such as Flickr and Zooomr allow users to share their photos with family, friends, and the online community at large. Flickr Tag Recommendation based on Collective Knowledge [6] provides the services is that users manually annotate their photos using so called tags, which describe the contents of the photo or provide additional contextual and meaningful information. Based on the analysis, it evaluates tag recommendation strategies to support the user in the photo annotation task by recommending a set of tags that can be added to the photo. The result of the empirical evaluation shows that it can effectively recommend relevant tags for a variety of photos with different levels of exhaustiveness of original tagging. A DBMS is a generic repository for the storage and querying of structured data. It offers a suite of interrelated services and guarantees that enables developers to focus on the specific challenges of their applications, rather than on the recurring challenges involved in managing and accessing large amounts of data consistently and efficiently. From Databases to Dataspaces: A New Abstraction for Information Management [7] was proposed byMichael Franklin, Alon Halevy, and David Maier. Dataspaces are not a data integration approach; rather they are moreof a data co-existence approach. The main aim of dataspace is to provide base functionality over all data sources, inspite of its integration. Similar to existing desktop search systems, it can provide keyword search over all of its data sources. When more sophisticated operations are required, Additional effort must be applied for sophisticated operations such as relational style queries, data mining, or monitoring over certain sources, to integrate the sources in

DOI: 10.15680/IJIRCCE.2015. 0311033

an incremental, "pay-as-you-go" fashion. A dataspace must deal with data and applications in a wide variety of formats accessible through many systems with different interfaces. It is required to support all the data [8] in the dataspace rather than leaving some out, as with DBMSs. A dataspace must offer the tools to create tighter integration of data in the space as necessary.

III. PROPOSED ALGORITHM

The proposed system mainly consists of two modules:Admin and User. A document can be annotated using the content value(attribute value) and query value(attribute name) by the user and the user can view or download the document. These are done using the process such as stopword removal, stemming, tokenization, and workload extractor. Using the QV-CV computation, score of each attribute is found by the conditional independence of Bayes theorem. It is mainly used for disaster management.

PHASE 1: ADMIN MODULE

Phase 1 consist of six stages. They are:

- User Registration Approval

  New user registration request was approved by the admin.
- Document Uploader

  Based on the type of the document such as flood/earth quake/storm/hurricane, date it is occurred and the location it is happened; admin can upload the document by browsing the file location.
- Stopword Management

  Stopwords are words which are filtered before or after processing of textual data. There is not one definite list of stop words which all tools use, if even used. Some tools specifically avoid removingthem to support phrase search. The most common stop words found in the text are "the", "is", "at", "which” and "on" etc. These kinds of stop words can sometimes cause problems when looking for the phrases that include them. Some search engines remove some of the most common words from the query on orderto improve performance. Here new stopwords can be inserted and also existing one can be deleted.
- Information Extraction

  Information extraction identifies the key phrases and relationships within the textual data. Information extraction infers the relationships between all the identified people, places and time from the text to extract the meaningful information. For handling huge volumes of textual data Information extraction can be very useful. The meaningful information is collected and stores in the data repositoriesfor Knowledge discovery, mining and analysis.

Information Extraction involves the following steps:

Step 1: Load the document

Here each line of a document is compared with the stopword list and removes the stopword fromthe document.

Step 2: Attribute Extraction

Here each word is separated from a line. Then calculate the total number of words. After thatthe relevance is measured by finding the count of words from the total words. With a threshold value, it can display the word and its relevance.

Step 3: Stemming

Stemming is the process for reducing inflected (or sometimes derived) words to their stem, base orroot form, generally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root. Algorithms for stemming have been studied in computer science since 1968. Many search engines treat words with the same stem as synonyms as a kind of query broadening, a process called conflation. Stemming programs are commonly referred to as stemming algorithms or stemmers.

Step 4: Tokenization

NLP is a large area, which includes topics like text understanding and machine learning. Tokenization is the process of breaking piece of text into smaller pieces like words, phrases,symbols and other elements which are called tokens. Even a whole sentence can be considered asa token. During the tokenization process some characters like punctuation marks can be removed. The tokens then become an input for other processes in text mining like parsing.

POS (Part-of-speech)tagging also known as grammatical tagging or word category disambiguation is the process of assigning a word in the text corresponding to a particular part of speech like noun, verb, pronoun, preposition, adverb, adjective or other lexicalclass marker to each word in a sentence. The input to a tagging algorithm is a string of words of a natural language sentence and a specified tagset. The output is a single best POS tag for each word. For this OpenNLP library is used. Table1 shows various POS tags.

Notation	Description
CC	Coordinating conjunction
CD	Cardinal number
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PRP	Personal pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
VB	Verb, base form
VBD	Verb, past tense
VBN	Verb, past participle

Table1: Summarization of POS tag notation

Step 5: Workload Extractor

NLP tool make use of NameFinder object. It contains <location>, <date>, <time>, <percentage>,<organization>, <money> and <person>.

PHASE 2: USER MODULE

User module consists of user querying. Here document type, date and location are taken on page load. User querying is based on the format q_i is a triplet ($A_{j,}$ p, V). Attribute can be of noun, verb, date, location, time, percentage, organization etc. Attribute values can be taken from the attribute suggestion list. After giving these values, the user can view or download the document.

NLP based synonyms can be used for retrieving documents with similar types. This will increase the searching process and the number of documents annotated can also be increased. Using the word and Pdf converters, Pdf and word documents can also be used for annotation process**.**

IV. PSEUDO CODE

QV-CV computation is done as follows:

Step 1: Retrieve each attribute from the document

Score of each attribute is computed from the conditional independence of Bayes Theorem with an equation

$$\text{Score }(A_j) = \frac{p(A_j\,|\,W)}{1-p\left(A_j|W\right)} \cdot \frac{p(d_t|A_j)}{p(d_t|\overline{A_j})}$$

Step 2: QV Computation
Step 2.1: Count the attributes
Step 2.2: Store each attribute into another variable
Step 2.3: Check the count of each attribute
Step 2.4: Compute the probability of each attribute over the workload
Step 2.5: Divide this probability with its negation
This is the first term of score i.e, $\frac{p(A_j\,|\,W)}{1-p\left(A_j|W\right)}$
Step 3: CV Computation
Step 3.1: Count the value of each attribute
Step 3.2: Store each attribute into another variable
Step 3.3: Check the count of each attribute's value with its attribute and value
Step 3.4: Compute the probability of each value over the total count
Step 3.5: Divide this probability with its negation
This is the second term of score i.e, $\frac{p(d_t|A_j)}{p(d_t|\overline{A_j})}$
Step 4: Calculate the threshold value $\tau = F(\overline{CV}, QV(A_j))$ [1] where $\overline{CV}$ is the maximum possible CV for the unseen attributes and $QV(A_j)$ is the QV of A_j.
Step 5: If the A_k has Score $(A_k) > \tau$, it retrieves the inferred attributes
Step 6: End.

V. Simulation Results

Annotating a document makes the search faster. A document can be annotated using the content and query value. If it is done with NLP, the searching becomes more efficient. NLP based synonyms also made the search more efficient.In this part graphs are shown to prove its efficiency.

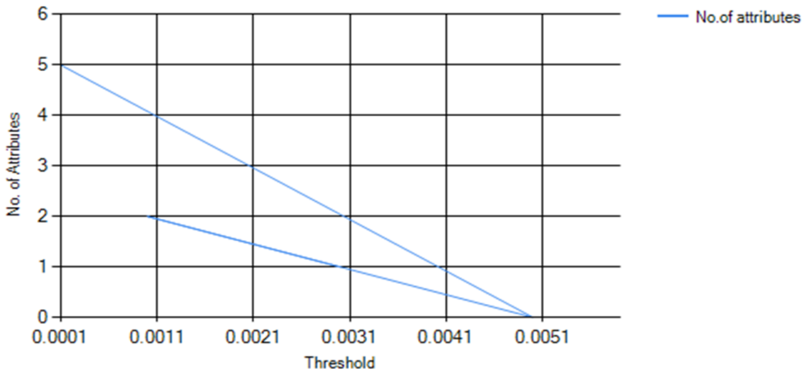

Fig.1: Threshold vs. No. of Attributes

Fig.1 shows an analysis of Threshold vs Number of Attributes. Several experiments have been conducted on different threshold value for a document. The value of threshold is in between 0 and 1.For threshold value 0.001, number of attributes is 2; for threshold value 0.003, number of attributes is 1; for threshold value 0.005, number of attributes is 0; for threshold value 0.0001, number of attributes is 5. Thus by reducing the threshold value, number of attributes can be increased. Thus the threshold act as a factor for filtering attributes and it affects the output of document annotation.

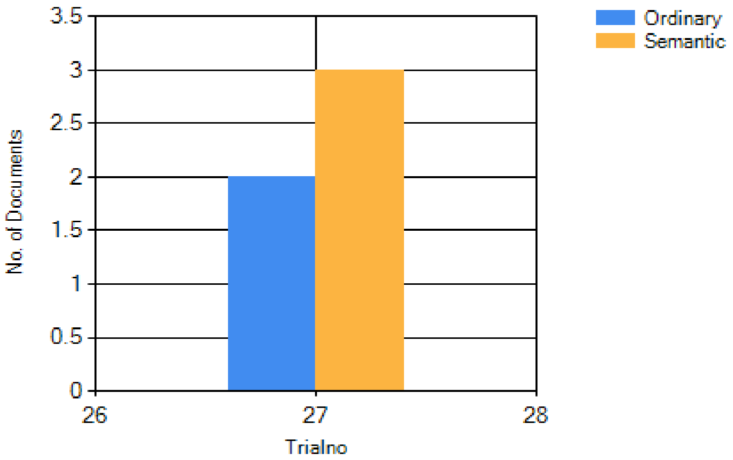

Fig.2: Trial no. vs. Number of documents

Fig.2 shows an analysis of Ordinary Annotation vs. Semantic Annotation based on NLP. Ordinary annotation means annotation of text documents, Pdf and word documents using NLP. Semantic annotation is based on the synonyms of attribute value on NLP. In semantic annotation, the synonyms can improve the annotation process. Graph shows that the semantic annotation can increase the number of documents to be annotated on a single search.

VI. Conclusion and Future Work

Annotating a text document using the content value and query value is done here. Here both values are combined for the annotation process. NLP based techniques are used for this processing such as stopword removal, stemming, and tokenization based on POS tagging. All these made the annotation process more effective, faster, and accurate than other annotation process and improve the searching process. NLP based synonyms made the annotation process more effective so that many similar documents can be annotated. In this work, word documents, Pdf documents can also beprocessed using Word converters and Pdf converters which increase the efficiency of this process. By annotating using the images in a document can be done as a future work.

References

1. Eduardo J. Ruiz, Vagelis Hristidis, and Panagiotis G. Ipeirotis, "Facilitating Document Annotation Using Content And Query Value", IEEE Transactions On Knowledge And Data Engineering, vol. 26, no. 2, February 2014.
2. "Google", Google Base, http://www.google.com/base, 2011.
3. S.R. Jeffery, M.J. Franklin, and A.Y. Halevy, "Pay-as-You-Go User Feedback for Dataspace Systems", Proc. ACM SIGMOD Int'l Conf. Management Data, 2008.
4. P. Heymann, D. Ramage, and H. Garcia-Molina, "Social Tag Prediction", Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '08), pp. 531-538, http://doi.acm.org/10.1145/1390334.1390425, 2008.
5. D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green, "Automatic Generation of Social Tags for Music Recommendation", Proc. Advances in Neural Information Processing Systems 20, 2008.
6. B. Sigurbjorrnsson and R. van Zwol, "Flickr Tag Recommendation Based on Collective Knowledge", Proc. 17th Int'l Conf. World Wide Web (WWW '08), pp. 327-336, http://doi.acm.org/10.1145/ 1367497.1367542, 2008.
7. M. Franklin, A. Halevy, and D. Maier, "From Databases to Dataspaces: A New Abstraction for Information Management", SIGMOD Record, vol. 34, pp. 27-33, http://doi.acm.org/10.1145/ 1107499.1107502, Dec. 2005.
8. M.J. Cafarella, J. Madhavan, and A. Halevy, "Web-Scale Extraction of Structured Data", SIGMOD Record, vol. 37, pp. 55-61, http://doi.acm.org/10.1145/1519103.1519112, Mar. 2009.

BIOGRAPHY

Jayalakshmi is a Post Graduate student in Department of Computer Science & Engineering, Sree Buddha College of Engineering for Women, Mahatma Gandhi University. She received Bachelor of Technology (B.Tech) degree in 2007 from Mahatma Gandhi University, Kottayam, Kerala, India. Her research interests are Data Mining, Networking etc.