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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

Aspect Based Document Annotation and Retrieval

Anumol Babu , Rose V Pattani

Post Graduate Student, Dept. of CSE, Mangalam College of Engineering, Ettumanoor, Kottayam, Kerala, India

Assistant Professor, Dept. of CSE, Mangalam College of Engineering, Ettumanoor, Kottayam, Kerala, India

ABSTRACT: Document annotation is the process of adding metadata information in the document which is used for information extraction. Collection of textual data contains certain amount of structured information which is hidden in unstructured text. So that it is always difficult to find relevant information. Annotations are presently extracted structured attributes within the document using the entity extraction technique .Annotations extracted using this technique loses the quality of topics and the attributes cannot be grouped in a meaningful sense. In order to solve this aspect mining methods are proposed for annotation. Aspects mean a condition. Aspects are obtained from a Latent Topic Modeling algorithm. The topics generated will give meaningful concept from the document. The experimental evaluation shows that this approach suggests more number of attributes for annotation and also retrieves more number of documents compared to the existing approaches.

KEYWORDS: Document annotation, Entity extraction, Attributes, Topic Modeling, Aspects

I. INTRODUCTION

Today organizations generate textual descriptions of their products, services etc. Such data contains some amount of important information which is hard to find out. Information extraction algorithms are normally used for extracting the relevant information. But it is expensive and inaccurate. So annotations are used .Annotations are comments or explanations that can be attached to a document. In many application purpose domains user can create and share the necessary information including social networking site, Disaster Management System etc. Information sharing tool like Microsoft share point permit user to generate and share the document and also to tag them. In the same way user can define attributes for their objects in Google base [6]. Hence annotation process can useful for information discovery.

Existing annotation systems results in very basic annotations. Such simple annotations make the analysis and retrieval of the data difficult to manage. Users are often limited to plain keyword searches, or make use of very basic annotation fields such as document type, creation date. Annotation systems that apply attribute-value pairs are normally more expressive, as they can contain more information. But the users should know the underlying schema and field types to use. The tasks become complicated and difficult to handle.

Annotations are presently extracted relevant information within the document using the entity extraction technique. Still it loses the quality of topics. So we propose aspect mining [2] methods for annotation. An aspect means a condition. It extracts aspects/meaningful concept from the document which is used for annotating the document.

II. RELATED WORK

Annotations using CADS (Collaborative Adaptive Data Sharing platform),[1] is an “annotate as you create” infrastructure which promote fielded data explanation to direct the annotation system. CADS system uses proposed adaptive technique to suggest attribute to annotate document .Query Workload, along with examining the content of the document is used. The attribute can improve visibility of the document while searching. For information sharing and user interaction many user-centered platforms are now available, Extracting information from these large bodies of texts is useful and challenging. A generative probabilistic aspect mining model (PAMM)[2] is used for identifying the

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

aspects/topics relating to class labels or categorical meta-information of a corpus. Some system which uses collaborative annotation of objects being derived from the user created tags [3] to annotate the new objects. The user can produce a label for entities. Previous research on label prediction system concentrates on getting better its accuracy or on managing the process, while neglecting the efficiency issues. Data quality is main problem in huge collection of databases. USHER [4] improves data quality dynamically. USHER is used for form designing and data entering and assuring data quality. Using existing data set of form, USHER derives a probabilistic model using questions form On every steps of the data entry process it helps to generate predictions and find error probabilities of the form. It will helps to reduce questions ask by user, and better performance of query search. A PMM (Poisson Mixture Model) two way model is anticipated to structural design of the document distribution into the act of mixing components in the middle of all clusters and combined words into the word clusters randomly. PMM is used for efficient document classification. A new document is categorized to help of the mixture model which based on its probabilities so the tags (labels) are suggested according to ranks [5].

III. PROPOSED SYSTEM

The proposed architecture of the system mainly focuses on attribute suggestion. Attributes are suggested either by QV-CV computation and document scoring or topic modelling.

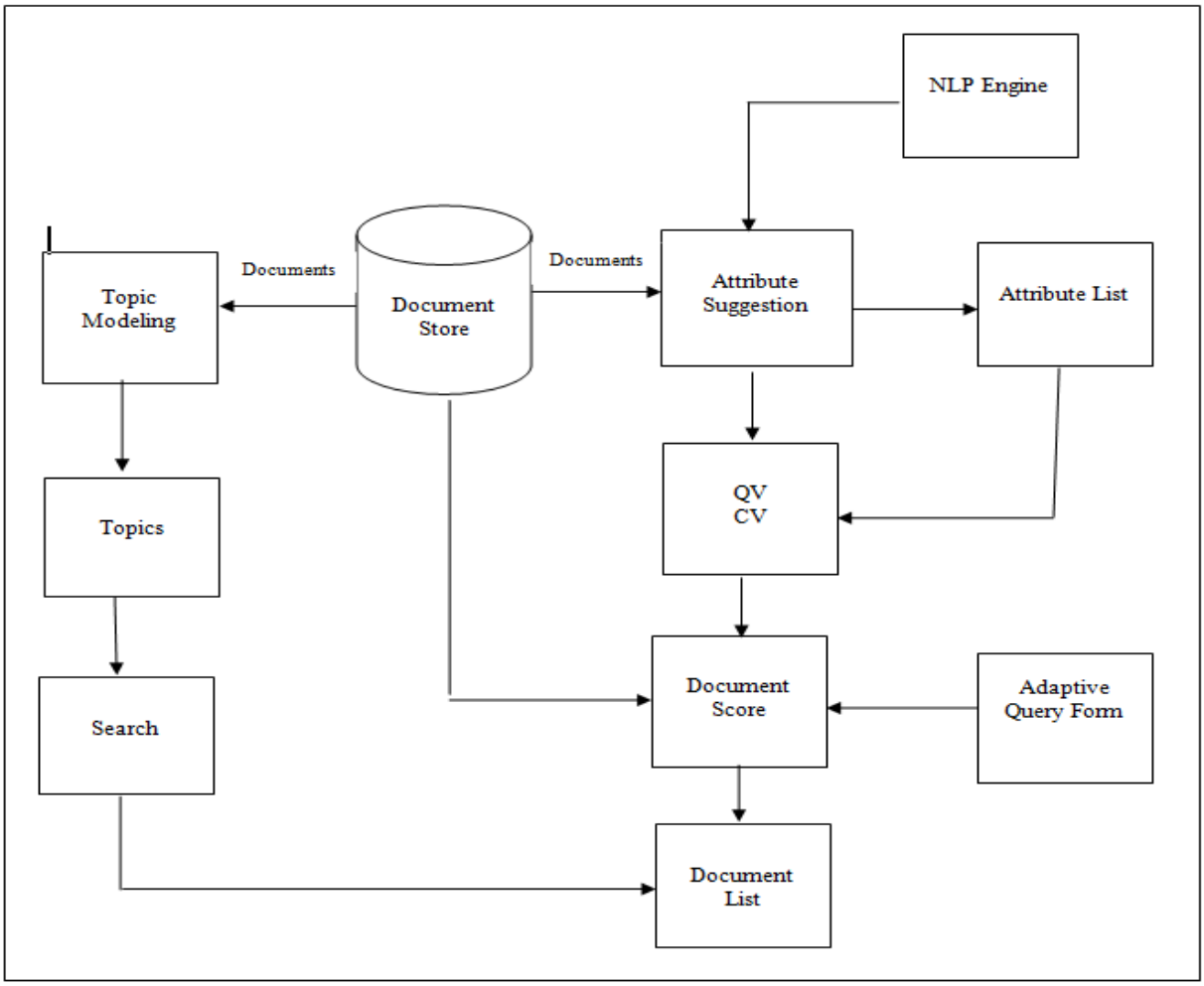

Fig 1. System Architecture

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

The Fig.1 represents the architecture of the proposed system. In this system receives documents from the document store as input. Then it performs Attribute suggestion based on NLP to find attribute list for annotation. System uses Content Value (CV) and Query Value (QV) computation of the attributes and assigns a document score based on the combined CV and QV. Finally the system returns documents having annotations satisfy query conditions ordered by decreasing score values using adaptive query form. The system also uses meaningful aspects from the documents for annotation by topic modelling. The search becomes efficient using the topics.

The proposed system mainly consist of the following modules

1. Document Management
   Here documents are uploaded to the data store with basic annotation such as document type, date and the location. Word documents and Pdf documents can also be processed using Word converters and Pdf converters

2. Corpus Preprocessing
   Here documents are pre processed before extracting the attributes by doing the process of stopword removal and stemming. A stop word is a commonly used word in a language which are filtered before processing of textual data.Stemming is the process of reducing word to their stems by removing prefixes and thus eliminating tag-of-speech and other verbal or plural inflections

3. Attribute Identification
   Here structured attributes within the documents are identified by doing the processes of information extraction, Querying Value computation and Content Value computation. Information extraction (IE) is the process of automatic extraction of relevant information from semi-structured or unstructured documents. It identifies the key phrases and relationships within the textual data. Information extraction commonly done by means of natural language processing (NLP). From the extracted attribute list Content Value (CV) and Querying Value (QV) of each of the attributes are calculated

4. Document Scoring
   Here document score is calculated by joint utilization of the QV and CV. Retrieve each attribute from the document and Score of each attribute is computed from the conditional independence of Bayes Theorem

5. Aspect Based Annotation
   Here aspects are obtained using a Latent Topic Modelling algorithm. LDA [Latent Dirichlet Allocation] implementation is used in the algorithm..Topics related with the document is extracted and this extracted aspects/topics from the document is used in subsequent information discovery. Aspects having highest topic distribution are used for annotation purpose

IV. PSEUDO CODE

CV Computation

Step 1: Parsing document text.
Step 2: For each term, compute its contribution in the document.
Step 2: Find its probability over the entire document

QV Computation

Step 1: Retrieve each attribute from list of query values
Step 2: Find its probability over the entire workload
Document Score Calculation

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

Step 1: Document score is calculated by multiplying the QV Value with CV Value

Combining QV and CV

Step 1: Retrieve each attribute from list of query values
Step 2: Get the Content Value for attribute.
Step 3: Calculate the threshold value as a function of CV and QV.
Step 4: If the attribute has Score value greater than the threshold value, then the attributes are suggested for annotation

Latent topic modeling algorithm

Step 1: Randomly select a distribution over topics
Step 2: For each word in the document
a. Randomly select a topic from the distribution over topics
b. Randomly select a word from the corresponding topic
Step 3: End

V. SIMULATION RESULTS

In the experimental evaluation different annotation methods such as ordinary annotation and aspect based annotation are compared. Ordinary annotation uses NLP for annotation of documents. Aspect based annotation uses topic modeling for annotation of documents.

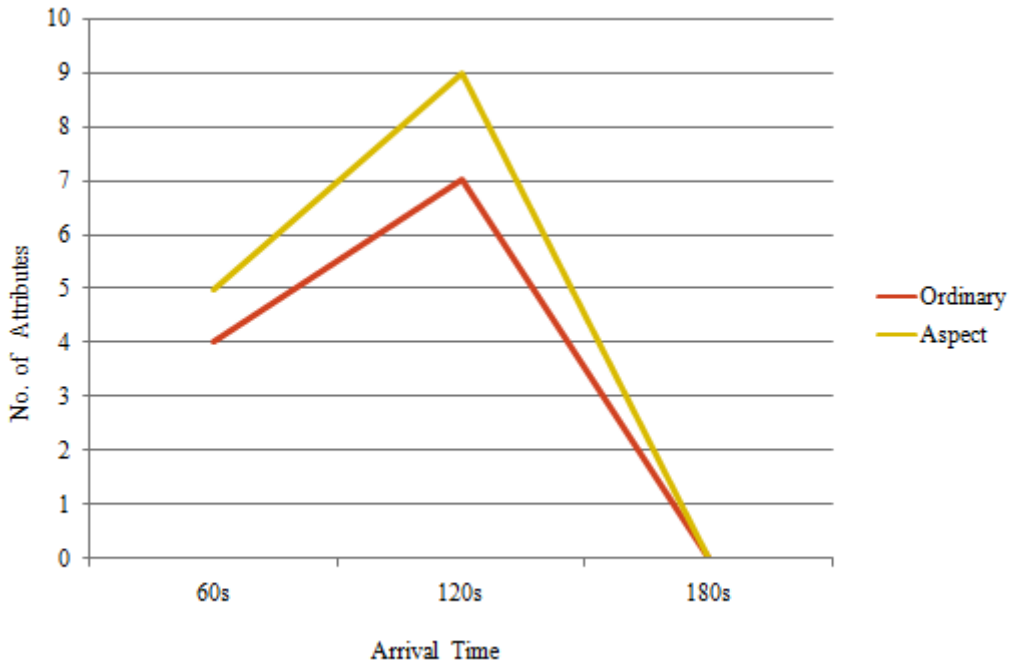

Fig 2. Arrival Time vs No. of Attributes

Fig. 2 represents the analysis of no. of attribute suggestion using the two annotation methods. Different experiments are done using different documents. The graph shows that aspect based annotation suggests more number

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

of attributes compared to the ordinary annotation. The number of attribute suggestion increased the visibility of the document

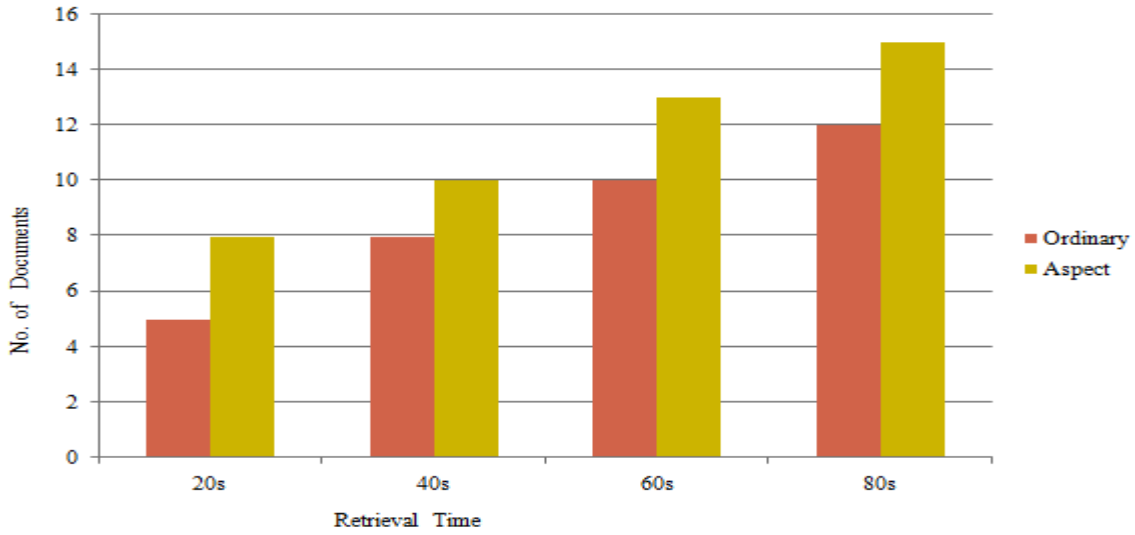

Fig 3. Retrieval Time vs No. of Documents

Fig. 3 represents the analysis of no. of documents retrieved using the two annotation methods. Different experiments are done. The graph shows that aspect based annotation retrieves more number of documents compared to the ordinary annotation. The number of attribute suggestion is high using aspect based annotation so that it improves the searching process. The efficiency is high using aspects in the document

VI. CONCLUSION AND FUTURE WORK

Aspect mining methods are used for the annotation of documents. Rather than the ordinary entities extracted, a lot of other topics related with the document is extracted. The topics generated will give aspects/meaningful concept from the document. These made the annotation process faster, effective and accurate than other annotation process. It improves the searching process. The data also helps to summarise the document.

. In future the system can process other type of documents such as excel documents, image files etc.

REFERENCES

1. Eduardo J. Ruiz, Vagelis Hristidis, and Panagiotis G. Ipeirotis, "Facilitating Document Annotation Using Content and Querying Value" , IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 2, pp. 336-349, February 2014
2. Victor C. Cheng, C.H.C. Leung, Jiming Liu, and Alfredo Milani, "Probabilistic Aspect Mining Model for Drug Reviews", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 8, pp.2002- 2013, August 2014
3. L. Hong, D. Yin, Z. Xue and B.D. Davison, "A Probabilistic Model for Personalized Tag Prediction", Proc. ACM SIGKDD Int'l Conference Knowledge Discovery Data Mining, pp. 959-968, July 2010
4. K. Chen, H. Chen, N. Conway, J.M. Hellerstein, and T.S. Parikh, "Usher: Improving Data Quality with Dynamic Forms", IEEE Transactions On Knowledge And Data Engineering, Vol. 23, No. 8,pp.1138-1153, August 2011
5. Y. Song, J. Li, W. -C. Lee, C.L. Giles Z. Zhuang, H. Li and Q. Zhao, "Real-Time Automatic Tag Recommendation" Proc. 31st Ann. Int'l ACM SIGIR Conference Research and Development in Information Retrieval (SIGIR '08), pp. 515- 522, 2008
6. "Google" , Google Base, http://www.google.com/base, 2011.
7. D. Blei , "Probabilistic topic models". Communications of the ACM, Vol. 55,No. 4 , pp.77-84, April 2012

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

BIOGRAPHY

Anumol Babu is a Post Graduate student in Department of Computer Science & Engineering, Mangalam College of Engineering, Ettumanoor, Kottayam. She received Bachelor of Technology (B.Tech) degree from University College of Engineering, Thodupuzha, Kerala, India.

Rose V Pattani is Assistant Professor in Department of Computer Science & Engineering, Mangalam College of Engineering, Ettumanoor, Kottayam. She received Bachelor of Technology (B.Tech) degree from **SCMS School of Engineering & Technology**, Angamaly, Kerala, India and **Master of Engineering** (M.E) degree from K.C.G College of Engineering,Chennai.