ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Implementation of Active Attribute Creation and Efficient Data Retrieval for Dynamic Document Annotation

Deepali Dagale[1], Dr. Poornashankar[2]

P.G. Student, Department of Computer Engineering, Indira College of Engineering and Management, Pune, Maharashtra, India[1]

Professor, Department of Computer Engineering, Indira College of Engineering and Management, Pune, Maharashtra, India[2]

ABSTRACT:Textual information is produced and shared by large number of organisation every now and then. Such textual information contains structured information hidden inside the unstructured text.Manually annotating each document is very tedious task. The main contribution of this paper is Attribute Suggestion for the text document at the time of uploading it. Content value as well as Query workload is used together to suggest the attributes. Suggested attributes along with the values in the adaptive Collaborative Adaptive Data Sharing (CADS) form is considered as the metadata for the document for annotation purpose. The metadata created could be useful for searching the documents in less amount of time. Annotation is done at the time of uploading the document, since user can add metadata at the time of document creation. Proposed algorithm identifies the attributes present in the document. Experimental evaluation shows that proposed algorithm generates good results as compared to existing approach that rely only on content of the document or query workload to suggest the attributes.

KEYWORDS: Annotation, adaptive, CADS, queryworkload

I. INTRODUCTION

In computer system, data mining concept is used to find out the meaningful data from large amount of data. In today's world, large number of organization produce and share large amount of data in different formats. Therefore, Data mining concept is very useful for processing this data.

Metadata is data about the data. Annotating the document means creating metadata of the text document. Annotation is nothing but comments, notes, explanation which can be added to the text documents. Annotation helps to improve future querying the database and improves the searching of the documents. Annotation also helps to rank the documents in database. Various static strategies are available for annotating the documents. But it has many drawbacks which are overcome by dynamic annotation methods.

Collaborative Adaptive Data Sharing (CADS) technique which is dynamic in nature is used to annotate the documents. In CADS adaptive insertion form for each document, it consists of static and dynamic attributes suggested by the proposed algorithm. The static attributes are those which are default attributes. Each and every document which will be uploaded will have these default attributes like creation date, author, and type of document. Dynamic attributes are those which will depend only on current text document.

Bayes theorem is used to find the score of the attributes and rank them. In CADS form, top k-attributes will be visible to the user along with its values. Frequency count (FC) is used to count number of values based on the content of the text. Query Value(QV) is obtained from query workload. Therefore, document is retrieved based on content and query workload and ranking is done using Bayes theorem.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Paper is organized as follows. Section II describes related work regarding dynamic annotation and attribute suggestion. Section III contains the flow diagram which represents the annotation process and algorithm used to suggest the attribute.Section IV presents experimental results showing results. Finally, Section V presents conclusion.

II. RELATED WORK

The review is done to get the insight of the methods, their shortcomings which can be overcome by giving the domain specific information, as input and storing the domain specific keywords in database with meaningful text and giving the related text as output..

Eduardo J. Ruiz, VagelisHristidis, and Panagiotis G. Ipeirotis, proposed a paper, "Facilitating Document Annotation Using Content and Querying Value". In this paper, they proposed CADS (Collaborative Adaptive Data Sharing) platform which is an "annotate-as-you-create" infrastructure that facilitates fielded data annotation. The main contribution of this paper is to use the query workload directly to direct the annotation process, in addition to examining the content of the document. In other words, they are trying to prioritize the annotation of documents towards generating attribute values for attributes that are often used by users to query the database[1].
MohdNaghibzadeh (2010), proposed XCLS (XML Clustering by Level Structure) algorithm to cluster XML documents by considering their structures. This algorithm represents XML structure in a new way called level structure. Level structure only uses the elements or tags of the XML document and ignores their contents and attributes. Using a global criterion for computing similarity is a considerable point in incremental methods[2].
SatishMuppidi, MohdNaghibzadeh (2015), proposed the document clustering with Map/Reduce using Hadoop framework. The algorithm is to sort data set and to convert it to (key, value) pair to fit with Map/Reduce. The operations of distributing, aggregating, and reducing data in Map/Reduce should cause the bottle-necks. MapReduce is a striking framework because it allows users to decompose the data involved in computing documents similarity into multiplication and summation stages separately in a way that it is well matched to effective disk access across several nodes[3].
Habibi, M., Popescu-Belis, A(2015), addresses the problem of keyword extraction from conversations, with the goal of using these keywords to retrieve, for each short conversation fragment, a small number of potentially relevant documents, which can be recommended to participants[4]. A method is used to derive multiple topically separated queries from this keyword set, and clustering technique is used to divide the set of keywords into smaller topically independent subsets constituting implicit queries in order to maximize the chances of making at least one relevant recommendation when using these queries to search over the English Wikipedia.

III. PROPOSED SYSTEM

The proposed document annotation system has two phases:Annotation phase and Search phase. Figure 1 shows the architecture of the proposed document annotation system. The system has three databases: document collection, query collection and annotation database. To create this system, Net Beans IDE using Java for front end designing and MySQL Server as back end is used.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

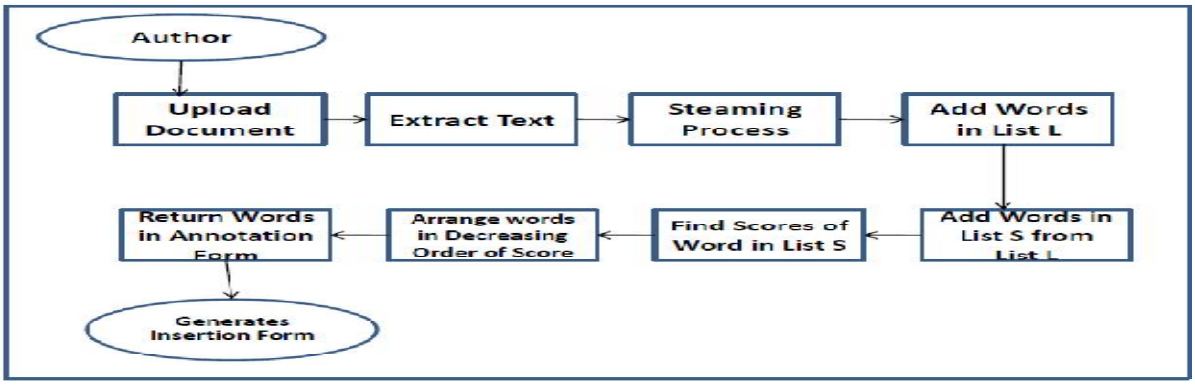

Fig 1: Flow of Annotation Process

In first step, the document is uploaded by user or creator. The document is preprocessed and stop words are removed. Then the proposed algorithm, analyses the text for generating best attributes for document i.e top k-attributes are suggested. Score of an attribute in a document is found by using a probabilistic approach. Total score of an attribute is the product of content score and query score. Attributes with high score is more relevant to that document. Finally attributes will display in an annotation insertion form in the decreasing order of their score.

ALGORITHM

Attribute Suggestion Algorithm(Proposed Algorithm): Used for Suggesting and Ranking the Attributes in Adaptive Insertion Form.

Step 1: Upload Document to be Annotated(eg. Doc5).
Step 2: Find list of keywords from the Document(eg.Doc5) by removing stop words.
Step 3: Get Attribute name for all the distinct keywords obtained in step 2 from Annotation Database.
Step 4: If found in above step, then map it with keywords else ignore that particular keywords.(Attribute Suggestion Process is Completed)
Step 5: For Ranking the Attributes in Adaptive Form, find out the Scores for each Attribute name obtained in step 4 by using Bayes Theorem.
Score = QV .CV

a) Find QV for each Attribute Name.

b) Find CV for the keywords obtained from step 2.

c) Top k-attributes are shown in Adaptive form (where k=2).

Step6: Update the annotation database

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

IV. EXPERIMENTAL RESULTS

The below figures shows the result of my proposed system. Precision and recall is used to evaluate the result.

RUNNING EXAMPLE

Input: Doc5
Content of the Document:need ice at Pembroke near to HS Doral. It is a Wilma or Gustav who is causing issue at west end.
Intermediate Steps:
Keywords = need, ice, peembrok, near, hs, doral, wilma, gustav, causing, issue, west, end

TotalNoOfAttributes_ImplementedAlgorithm = 4
TotalNoOfAttributes_DataFrequencyMethod = 11

Final Output: where k=2
Suggestion for CAD Form:
Attribute = Supplies , Value = Ice
Attribute = City , Value = Peembroke
CADS Form Submitted

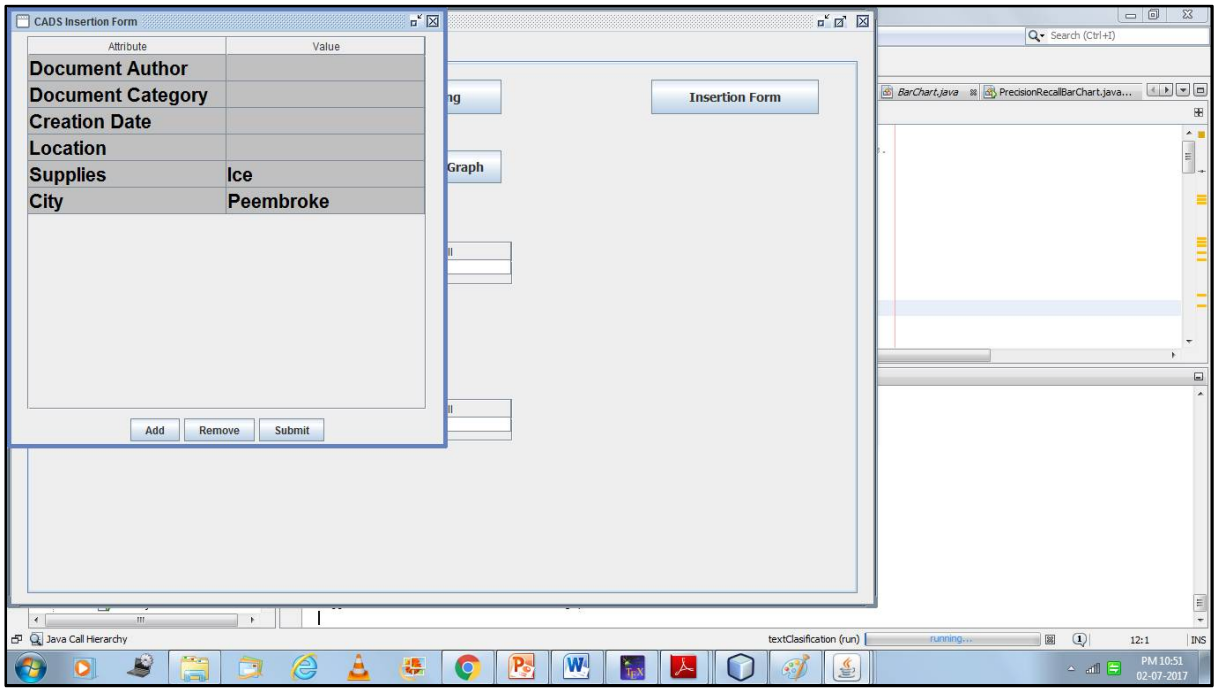

Fig 2: Screen Shot for Annotation Process

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

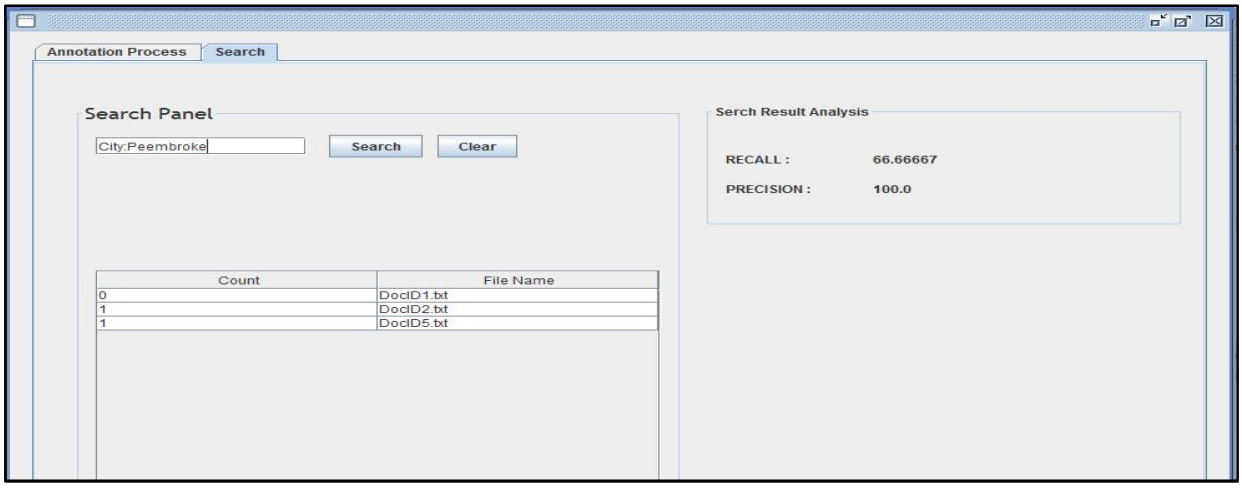

Fig 3: Screen Shot of Search Result

Precision: The precision in document retrieval can be defined as the measurement of the retrieved relevant documents to the query of the total retrieved documents.
Precision measures the accuracy of the retrieval.
Precision=A/B where A is Number of relevant retrieved documents and B is the total retrieved documents.
Recall: The recall in documents retrieval can be defined as the measurement of the retrieved relevant documents to the total database documents.
Recall=A/C where A is Number of relevant retrieved documents and C is Total number of relevant documents in the database.

k- Suggested Attributes Value	Precision(Proposed Algorithm)	Precision(Data Frequency Algorithm)	Recall(Proposed Algorithm)	Recall(Data Frequency Algorithm)
2	100	50	50	9.09
5	40	20	50	9.09
7	28.57	14.28	50	9.09
10	20.0	10.0	50	9.09

Table 1: Evaluation of Existing and Proposed Algorithms

Following graph shows the comparison between proposed and existing system to view the final result.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

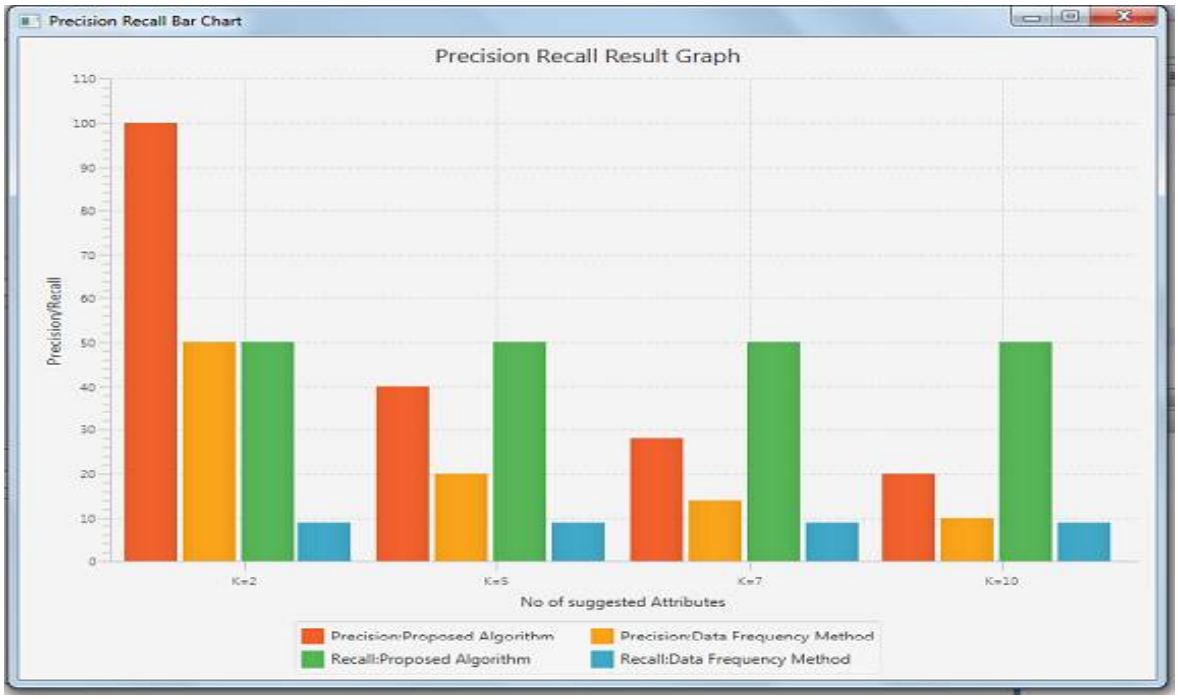

Fig 4: Overall Result of the Proposed System.

V. CONCLUSION

Thus, Bayes Theorem is very effective technique to suggest and rank the attributes in adaptive insertion form. By considering content of the document and query workload together provides good results for suggestion as well as for searching the document after annotating it. Hence, improves the visibility of the document to great extent. In future, the system can process other type of documents such as excel file, image file.

REFERENCES

[1] Eduardo J. Ruiz, Vagelis Hristidis, and Panagiotis G. Ipeirotis, "Facilitating Document Annotation Using Content and Querying Value", IEEE TRANSACTIONS, VOL. 26, NO. 2, 2014.
[2] Mohd. Alishahi, Mohd Naghibzadeh , "Tag Name Structure Based Clustering of XML Documents ", International Journal of Computer and Electrical Engineering, Vol. 2/1, 1793-8163, 2010.
[3] Satish Muppidi, Mohd Naghibzadeh, "Document clustering with Map reduce using Hadoop framework", International Journal on recent and innovation Trends in Computing and communication, Vol. 3/1, 409-413, 2015.
[4] Habibi, M.; Popescu-Belis, A.,"Keyword Extraction and Clustering for Document Recommendation in Conversations" Volume: 23, Pages: 746 – 759, 2015.
[5] Levent Bolelli, Seyda Ertekin , "K-SVMeans: A Hybrid Clustering Algorithm for Multi-Type Interrelated Datasets",7695-3026-5/07, 2007
[6] Y.SureshBabu, K.Venkat Mutyalu, "A Relevant Document Information ClusteringAlgorithm for Web Search Engine",International Journal of Advanced Research in Computer Engineering & Technology,Volume 1, 2012.
[7] V.M.Navaneethakumar, Dr.C.Chandrasekar, " A Consistent Web Documents Based Text Clustering Using Concept Based Mining Model", International Journal of Computer Science Issues, Vol. 9, 2012.
[8] Mr. Sumit D. Mahalle, Dr. Ketan Shah, "Document Clustering by using Semantics", International Journal of Scientific & Engineering Research, Volume 3, 2012.
[9] Chandan Jadon, Ajay Khunteta(2013), "A New Approach of Document Clustering", International Journal of Advanced Research in Computer Science and Software Engineering,Volume 3.
[10] Jiashen Sun; Xiaojie Wang,"Annotation-aware web clustering based on topic model and random walks" International Conference on Cloud Computing and Intelligence Systems (CCIS), Pages: 12 – 16, 2011.
[11] Baghel, R.; Dhir, R., "Text document clustering based on frequent concepts", International Conference on Parallel Distributed and Grid Computing (PDGC),Pages:366-371, 2010.
[12] Durga Toshniwal, Rishiraj Saha Roy(2009), Clustering Unstructured Text Documents Using Naive Bayesian Concept and Shape Pattern Matching, IJACT : International Journal of Advancements in Computing Technology, Vol. 1, No. 1, pp. 52 ~ 63., 2009.[