Improving Annotations in Digital Documents using Document Features and Fuzzy Logic

Priyanka C. Ghegade[1], Prof. Vinod S. Wadne[2]

[1] PG Student, Department of Computer Engineering
, JSPM's ICOER, Maharashtra, India
[2] Asst. Professor, Department of Computer Engineering
, JSPM's ICOER, Maharashtra, India

---***---

Abstract - ***A web is biggest repository of unstructured information, it is becoming a challenging job to align the document in a finite format. In order to fulfill need of users , libraries must provide relevant and useful data. To do this any one has to study the document and then have to format it in desired manner. But this process of formatting document content can take much time to finish the job with desired accuracy. So many systems are proposed to do this, but most them are losing their semantics. The paper contains framework that presents an efficient and effective approach for automatically annotating document content within information rich texts using document features and fuzzy logic. So proposed system put forwards an innovative idea of annotation of documents using the extracted features present in the documents like title feature, sentence length, term weight, sentence position, sentence to sentence similarity, proper noun, numerical data, thematic word, etc. Then the features are extracted using proposed algorithm and then these document features are provided as input to the fuzzy logic to achieve the best possible summary and to maintain accuracy in annotations, Based on this by using weighted feature value technique document annotation applied to the hugh amount of the digital documents. Our approach clearly maintaining the accuracy as well as semantic of the annotation in document in all possible conditions and scenarios.***

Key Words: *Feature extraction, NLP, Fuzzy logic, Annotation.*

1. INTRODUCTION

Due to increase in information in computer industry, there is need of arranging digital documents in proper way. So that digital libraries able to fulfill the informative need of current as well as future users. There are number of application domains are there that generates and share the generated information for e.g. newspapers, social networking groups, media channels etc. There are number of annotating tools and annotation techniques that annotate documents and make it available to provide relevant data. Microsoft sharing tool is a one of the efficient sharing tool which enable the user to share the information and tag or annotate it. The process of annotating document content is done using ad hoc method. Annotation assigned by annotation techniques to document is used to gives tag to the objects and thus it allows the users to organize the documents. There is another sharing tool like Google base. Google base is database provided by the Google in which user can add any types of data i.e. textual data, pictures, videos etc. It accepts the data in any format. It allows the users to define the attributes, also it enable the users to select attribute values from the readymade templates. But these annotation or tag provided to document requires knowledge discovery in various documents i.e. on large amount to discover the best annotation.

There are some annotations strategies are presents that making use of attribute-value pairs. It has been found that this is one of the effective as well as useful methods of annotation. While using this system a care is taken that the document should be in structured format. If user using annotation using attribute and value pair then user should be more upstanding in there annotation work. Users should have knowledge of insight attributes of documents also when to use these attributes. But if the numbers of attributes are in range of tons then it will be infeasible and complicated to use such approach. Also it will create more loads on relevant system. Even if the option is given in the system that allows the users to tag(annotate) the document using attribute and value pair method then users will have less interest in doing the given task. Such difficulties will results in very poor annotation also it will be restricted to the simple keywords only. Such poor annotation will cumbersome the data and the system too.

Limitation of current system:

- Tagging is noisy, user is missing tags or annotate with incorrect annotation.
- Attribute on this domain are poorly correlated so searching for correlations can be misguiding.
- The existing system had number of issues and disadvantages unaddressed and also prediction rate was not very appreciable.

Nowadays most of the people using web applications, so that results in demand for relevant data and increases need of annotation. Proliferation in the web applications i.e. flicker application increases the need of annotation. Also the approach getting much attention in the research, industry and academics. Some systems are there that can give the automatic annotation suggestion based on the context. Automatic annotation identification is selected a one of the important research topic. When the online recommendation of tagging is done at that time from the given vectors of documents the posterior probabilities of classes are calculated at first stage. Then by studying the combined possibilities of the tags and documents annotations are created for the respective documents based on the ranking of the cluster. A study has been shows that the average amount of time required to generate annotation is only 1.1 sec.

Proposed system is providing annotation for digital documents but provided annotations should be semantic, relevant and effective. One can get relevant document while searching. Hence the algorithm should focus on only those documents that contain the query words. If the contents of the documents are ignored while doing the task then it will be the waste of time as it will be unable to find the exact documents. Hence feature extraction can be done on the documents before applying the annotation algorithms. It is useful to generate semantic annotations. In feature extraction method main features can be selected to conclude the task: Proper nouns, title feature, sentence length, sentence position, sentence to sentence similarity, numerical data, term weight, and thematic word. Before extracting pre-processing is done to remove things which unwanted from the documents as it increases the operation time.

In pre-processing three processes such as stop word removal, special symbol removal and stemming is done to bring the documents in more structured form. In pre-processing most of the systems are focuses on issue at a time i.e. either accuracy of the annotation done or its efficiency.

However for efficient extraction of information ontologies can be used. Normal systems are unable to reference the geographical features. These features can be easily referenced by the ontologies.

The rest of the paper content is organized as. Section 2 discusses related work i.e. literature survey and section 3 presents the design of our approach. The details of the results and discussions about proposed framework are presented in section 4. Sections 5 provide hints of some extension of our approach as conclusion.

2. LITERATURE SURVEY

A lot of research work has been done on annotating digital documents and improve the accuracy of providing relevant document rate. Some of the work of Research of recent year is described below.

Nowadays numbers of technologies are there and today's digital era of these technologies is on glance. On every day number of techniques are explored as well as implemented by people. With this new technologies and enhancement number of organizations as well as channels also generate and presented to share the information like video, image and audio etc. With this new information still text is prioritizing and choosing as a great way to share the information by the people. Also the large amount of the data of the digital library also exists in the form of text only. Hence nowadays the number of documents to be processed is increasing tremendously. Therefore having no control over meaningful data is like having no information. The same problem is known and called as an information overloading unless suitable technique or methods were implemented to share, organize, indexing and retrieval. Hence natural language processing (NLP) is emerged as an area to deal with textual data. The input data given to NLP should be with neat and proper format.

To bring the output pre-processing is required on the input data. Normalization is one of such method used for pre-processing. It is the process of bringing the word to its original form i.e. root form. This can be done by the stemming process. Normally normalization is done to find and remove the suffixes attached to the words in order to find the occurrences of the same word repeatedly in specific context (e.g. going and go gives the same meaning, computing and computed also have same meaning). Again the suffixes to be searched should be known in advance. Sometimes it may possible that normalization carried by the stemming process will change the meaning of the word (computing and computed will give compute), a solution on this is to use lemmatization. But a condition that doesn't have the linguistic knowledge in prior will support the stemming as a best method.

In the paper[1] " Facilitating Document Annotation Using Content and Querying Value " Author Eduardo J. Ruiz, Vagelis Hristidis, and Panagiotis G. Ipeirotis(2014) describes Innovative and effective document annotation technique which include Collaborative Adaptive Data Sharing platform is used as an annotate as you create document infrastructure that facilitates fielded data annotation. The goal of Collaborative Adaptive Data Sharing platform is to encourage as well as decrease the cost of creating annotated documents that can be further useful for commonly issued semi structured queries. Proposed system's goal is to encourage the annotation of the documents at time of creation of that document, while the creator of the document is still in the phase of document generation, the same techniques can also be used for post generation of document annotation. System works in this way the author generates a new document and uploads it to the repository. Once document is uploading, CADS analyzes the document text as well as creates an adaptive insertion form. The form generated by CADS contains the best attribute names suggested for given the document text and the information need that is query workload, and the most attribute values suggested for given the document text. The author can check the form, modify the generated metadata as necessary according to document and submit the annotated document for storage in database.

In the paper[2] "A Probabilistic Model for Personalized Tag Prediction" Author D. Yin, Z. Xue, L. Hong, and B.D. Davison(2010) describes Probabilistic tag recommendation systems have a similar goal and objective like facilitating document annotation using content and querying value. However the main difference is that proposed system i.e. a probabilistic model for personalized tag prediction uses the query workload in model, reflecting the user interest. A Bayesian approach is used as well as integrates three factors an ego-centric effect, content of web page and environmental effects. There are two methods intuitive calculation and learning Optimization provided for parameter estimation.

In the paper[3]"Towards Ontology-based Information Extraction and Annotation of Paper Documents for Personalized Knowledge Acquisition" Author Benjamin Adrian, Heiko Maus , Malte Kiesel , and Andreas Dengel presents an annotation approach that makes use of ontologies for the feature extraction purpose. Annotation plays an important role when there is low or none textual information is available. Thus annotation of the multimedia objects such as images, videos is an interesting area of study. In "Flickr Tag Recommendation based on Collective Knowledge" [4] annotation recommendation algorithm used by flicker is presented. Flicker is online photo share community that allows the users to share the photos. The algorithm has the facility to give annotation to the photos. These tags are based on the content of the images.

As discussed in previous section, annotation deals with the hugh unstructured data. This hugh data deals with the many problems like inconsistency, missing data, noisy data. So to deal with all those problems preprocessing is used. Preprocessing plays an important role in the area of document annotation because it greatly reduces the amount of generated data. The reduced data size strongly improves the performance of the system as the time gets saved for the processing of such extra information. In the paper[5] " A Comprehensive Approach Towards Data Preprocessing Techniques & Association Rules" Jasdeep Singh Malik, PrachiGoyal, Mr.Akhilesh K Sharma Gives detail information of all the phases involved in the process of preprocessing. Preprocessing mainly consist of four phases as Data cleaning, Data integration, Data transformation, Data reduction.

Data cleaning is a technique of observing and thus replacing the data which is being missed. Data integration deals with collecting the data from the multiple sources and putting it to one place, so that it will be cost effective to retrieve the data. Data transformation comprised of converting or transferring the data to most suitable form for the better information retrieval. Data reduction is used for removing or cutting the data size by replacing the less important data.

Given below is the figure showing all the mentioned methods of preprocessing.

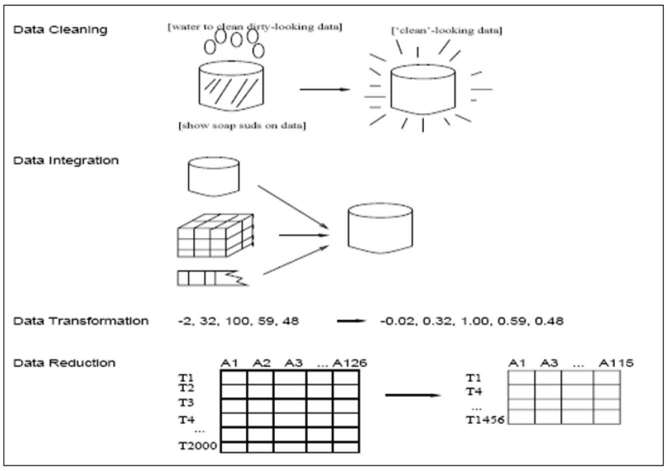

Fig- 1: preprocessing methodology

Feature extraction is the normal process observed in the field of data mining. It helps the system to select the important data by discarding the data which will not going to contribute in the process of answer extraction. Below are the two basic techniques used for the feature extraction Principal Component Analysis and Linear Discriminant Analysis.

Figure shows the generalized steps observed in the feature extraction process.

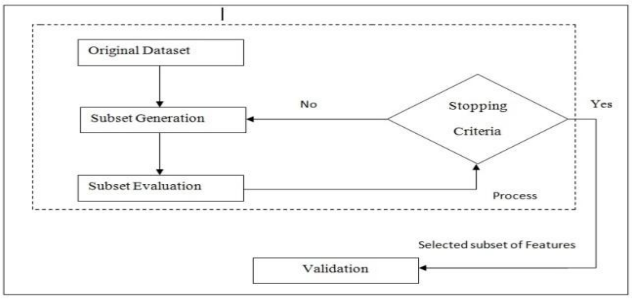

Fig -2: Feature extraction process

Paper [6] "Feature extraction for classification in the data mining process." Author Pechenizkiy, Mykola, SeppoPuuronen and Alexey Tsymbal illustrates the feature extraction techniques observed in the classification process. The main reason behind the paper is to show how to overcome the problem of the curse of dimensionality. Here different feature extraction approaches based on Eigen vector are discussed. Here author proposed a decision support system which combines the feature extraction and classification approach. The feature extraction used here is based on the principle component algorithm.

For selection of best summarization method, Paper [7] "A survey of text summarization extractive techniques." Author Gupta, Vishal, and Gurpreet Singh DLehal. proposed a deep survey on 12 best text summarization methods. Here the problem statement of each method is well narrated. Also the techniques used for the implementation is being good represented. Apart from this mathematical equations if needed are elaborated. Advantage of each system along with the disadvantage is deeply presented. So by reviewing the paper one can simply come to the conclusion for the selection of best method to be used. In this paper total 14 features that can be extracted from documents are presented by the author. So it will get simple for the intended person to select the best feature as per their applications in which summary is needed to generate.

[8] Narrates a framework for document annotation known as GoNTogle. GoNTogle is completely based on the web semantics and the information retrieval techniques. Here the annotation is done by taking the help of ontologies. It supports both i.e. manual and automatic annotation scheme. Apart from the document annotation a searching option is also provided by the author. For the flexible searching mechanism a keyword based and semantic based searches are combined to form brand new search technique.

3. PROPOSED SYSTEM

Here we describe system for annotation in digital document using document feature with fuzzy logic. There are some mentioned steps as shown in figure 3

Step 1:Here user submits document set which are read into string and send to the preprocessing step.

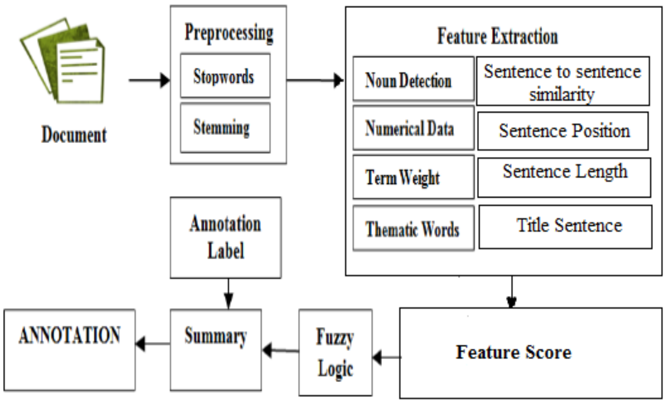

Fig- 3: Overview of the proposed system

Step 2: Preprocessing This is the step where preprocessing is conducted, where string is processed to its basic meaning
words by the following four main activities:

- *Sentence segmentation* is technique where boundary detection is done and in that separate inputed text into sentence.
- *Tokenization* is a process of separating inputed query into individual or separate words.
- *Stop word removal* :In any text document narration the conjunction words i.e. is ,an ,the does not play that much role in the document meaning , so by discarding these words (like: is, the, for, an) from the documents which greatly reduces the time as well as processing overhead.
- *Stemming:* Documents contains many of the elongated words of English language not only fail to provide proper meaning in the given condition but also they increases the time required for computation. So it is necessary to bring the words to their base form by replacing its extended.

Step 3:-Applying Feature extraction: *As* mentioned in earlier segments the features of a text play a very great role in semantic categorization techniques. So our proposed system makes use of different features of document, which are extracted from query entered by the user as mentioned below.

3.1) Noun detection

Noun extraction from the entered query plays a vital role in identification of the perception of the query. This is done by comparing each word of the query with the dictionary collected for almost 1, 00,000 words of English language. The details of this process can be shown in the below algorithm.

3.2) *Term weight.*

The most repetitive words in text are obviously the important words. So system identifies the list of most repeated words and considers some top n elements (where n is user defined) as the important word for text to store in vector.

3.3) Thematic Word

The number of thematic word in sentence, this feature is essential because terms that arise frequently in a document are perhaps related to matter. The count of thematic words means the words with maximum number probable and also relative to sentence. Here in proposed system we are considering top 5 most repeated word. The feature score for thematic word feature is calculated as the proportion of the number of thematic words that present in the sentence over the maximum value of summary of thematic words feature in the sentence.

$$F(s) = \frac{NoofThematicWords \in s}{Max(NoofThematicWord)}$$

3.4) Numerical Data

The number of numerical data in sentence, sentence that holds numerical data is important and it is most probably included in the document summary. The feature score for numerical data feature is intended as the proportion of the number of numerical data present in sentence over the length of particular sentence.

$$F(s) = \frac{NoofNumericalData \in s}{SentenceLength\,(l)}$$

3.5) Title sentence

The word in sentence that also occurs in title gives high score. Title sentence is identified by using calculation of the number of matching between words present in a sentence and the words in the title of document. We are calculating the feature score for title by taking ratio of the number of words present in the sentence.

3.6) Sentence Length

This sentence length feature is beneficial for filtering short sentences such as journalist names, datelines, venues, time commonly found in news articles. Such type of short sentences are not predictable to belong to the summary. Here We using sentence length, which is calculated as the proportion of the number of words present in the sentence over the number of words present in the sentence of the document whose length is long.

3.7) Sentence to sentence similarity

This feature is a similarity among sentences. For each sentence S, the similarity between S and each other sentence is computed by the cosine similarity measure with a resulting value between 0 and 1 Now, the score of this feature for a sentence S is acquired by computing the proportion of the summary of sentence similarity of sentence S with each other sentence over the maximum of summary.

3.8) Sentence position

location in text gives the rank of the sentences. This sentence position feature can contain several type of stuffs such as the location or position of a sentence in the document section as well as paragraph, etc., suggested the very first sentence is having highest ranking. The score for sentence position feature: here we are considering the first 5 sentences in the each paragraph. This feature score is intended as the succeeding equation.

Step 4: Applying Fuzzy Logic - The aim of text summarization is based on extraction method of sentence selection. One of the methods to get the appropriate sentences is to consign some numerical measure of a sentence for the Summary known as sentence weighting and then select the best ones. Therefore the features score of each and every sentence that we termed in the prior section are used to acquire the significant sentences. In this section, we use method to extract the essential sentences: Fuzzy Logic method based Text Summarization. Fuzzy is a system whose base is rule. Main function of fuzzy interference system is rule and result quality in fuzzy based system depends on fuzzy rule. Improve better result with the help of fuzzy logic The system involves of the following core Steps:

Step A: In the fuzzifier, Inputs are consider as crisp inputs which are taken from result of the feature extraction.

Step B: After fuzzification step, the inference engine referring to the rules contains fuzzy IF-THEN rules.

Step C: This is final step, in this we get the final sentence score. In inference engine step, important part is to define fuzzy IF-THEN rules. The essential sentences which are important are extracted by using these rules according to our features criteria. Sample of IF-THEN rules are described below.
IF (0.81< NoWordInTitle) and (0.81< SentenceLength) and (0.81< TermFreq) and (0.81< SentencePosition) and (0.81< SentenceSimilarity) and (0.81< NoProperNoun) and (0.81< NoThematicWord) and (0.81< NumbericalData) THEN (Sentence is important).
After this process all the text sentences are ranked in a descending order according to their scores. A set of uppermost score sentences are extracted as a text summary.

Step 5:In this step annotation labels are extracted from text summary by string comparing and displayed to the user.
The whole proposed system is expressed mathematically in the below model.

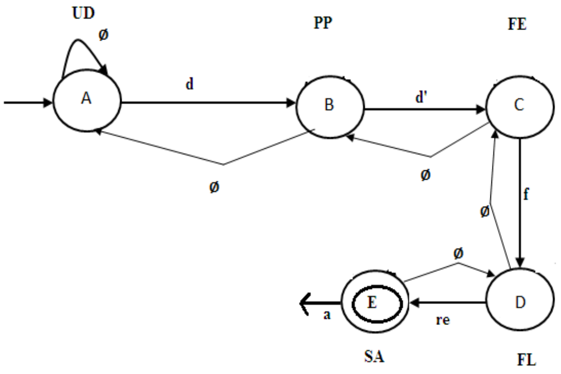

Fig-4: Deterministic Finite Automata

A deterministic finite automata M is a 5-tuple, (Q,∑,δ,q0,F) consisting of

- a finite set of states (Q)={A,B,C,D,E}
- a finite set of input symbols called the alphabet(_)={d,d1,p}
- a transition function (δ : Q * ∑ → Q)={ UD,PP,FE,FL,SA}
- a start state (q_0 ϵ Q)={q0}
- a set of accept states (F ⊆Q)={q4}

WHERE,
d=document
d'=document after preprocessing
f = extracted features
re=results of fuzzy logic module
a =annotation
UD=Uploading Document
PP=Preprocessing Loaded document
FE=Feature Extraction
FL=Fuzzy Logic
SA=Summary analysis

Derivation (δ) is defined as following transition table.

States	D	d'	f	re	a	Ø
A	B	Ø	Ø	Ø	Ø	A
B	Ø	C	Ø	Ø	Ø	A
C	Ø	Ø	D	Ø	Ø	B
D	Ø	Ø	Ø	E	Ø	C
E	Ø	Ø	Ø	Ø	Ø	D

Table-1:DFA

4. RESULTS AND DISCUSSIONS

For analyzing effectiveness and accuracy of proposed framework some experiments are conducted on windows java based machine. To measure the performance as well as accuracy of the system we set the bench mark by selecting real world documents as the input to the system.

To determine the performance of the system, we examined how many relevant documents are annotated on the basis of our technique.

To measure the effectiveness of proposed system we are considering precision and recall as the best measuring techniques. So precision can be defined as the ratio of the relevant Attribute suggested for the annotated documents to the total number of irrelevant and relevant Attribute suggested for the annotated documents. Measures result usually expressed as a percentage. This techniques of measurement gives the information about the relative effectiveness of the system.

Whereas Recall is the ratio of the number of relevant Attribute suggested for the annotated documents to the total number of irrelevant Attribute suggested for the annotated documents This measures gives the information about absolute accuracy of the system.

The advantage of having the two for measures like precision and recall is that one is more important than the other in many circumstances.

For more clarity let we assign

- A = The number of relevant Attribute suggested for the annotated documents,
- B = The number of irrelevant Attribute suggested for the annotated documents, and
- C = The number of irrelevant Attribute suggested for the annotated documents.

So,

Precision = (A/ (A+ C))*100
Recall = (A/ (A+ B))*100

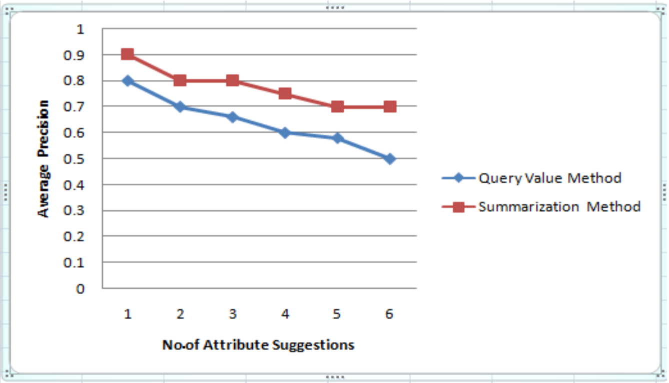

Fig-5: Average precision of the suggested attribute Comparisons

In Fig. 5, we observe that average precision for the suggested attribute of the proposed method of summarization yields more precision compare to the method suggested by [1]

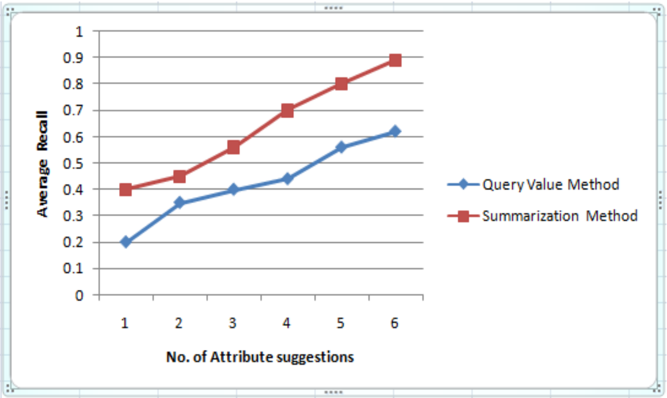

Fig-6: Average Recall of the suggested attribute Comparisons

In Fig.6 , we observe that average Recall for the suggested attribute of the proposed method of summarization yields more precision compare to the method suggested by [1]

5. CONCLUSIONS

The proposed system of document annotation successfully reads all types of txt documents. Then our system properly preprocesses the inputed text data to reduce the overheads of the unwanted text from the documents. As the main step of our system, our method extracts the main document features from the finely pre-processed data like sentence length, noun, sentence to sentence similarity, thematical word, sentence position, numerical data and top repeated words etc. As these features are indicating the best possible meta data of the document. Then fuzzy logic classifies the features based on the if- then rules to get the perfect summary of the document.

This summary contributes further for extraction of the important annotation from the document which is obviously one of the meaningful annotations for the given set of documents.

The proposed system can be enhancing to work for huge amount of documents in distributed network where thousands of documents can be set for annotation in a single go.

ACKNOWLEDGEMENT

I would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance. I'm also thankful to reviewer for their valuable suggestions and also thank the college authorities for providing the required infrastructure and support.

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, FEBRUARY 2014.

[2]D. Yin, Z. Xue, L. Hong, and B.D. Davison, "A Probabilistic Model for Personalized Tag Prediction," *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery Data Mining*, 2010.

[3]Benjamin Adrian1 , Heiko Maus1 , Malte Kiesel1 , and Andreas Dengel1,2 "Towards Ontology-based Information Extraction and Annotation of Paper Documents for Personalized Knowledge Acquisition"

[4] Börkur Sigurbjörnsson, "Flickr Tag Recommendation based on Collective Knowledge".

[5] Jasdeep Singh Malik, PrachiGoyal, 3Mr.Akhilesh K Sharma 3 "A Comprehensive Approach Towards Data Preprocessing Techniques & Association Rules" Assistant Professor, IES-IPS Academy, Rajendra Nagar Indore – 452012 , India

[6]Pechenizkiy, Mykola, SeppoPuuronen and Alexey Tsymbal. "Feature extraction for classification in the data mining process." (2003).

[7]Gupta, Vishal, and GurpreetSingh DLehal. "A survey of text summarization extractive techniques." Journal of Emerging Technologies in Web Intelligence 2.3 (2010): 258-268.

[8]Giannopoulos, Giorgos, et al. "GoNTogle: a tool for semantic annotation and search." The Semantic Web: Research and Applications. Springer Berlin Heidelberg, 2010.376-380.

BIOGRAPHIES

	Mr. Vinod S. Wadne received the B.E and M.E. degrees in computer Engineering And working As Assistant professor In Department of Computer Engineering at Imperial College Of Engineering And Research, Wagholi, Pune. He has 12 year teaching Experience. His areas of interest in research is Data Mining.
	Priyanka C. Ghegade received the B.E. degree in computer Engineering from Pune university and pursuing master of Engineering in computer Engineering from Imperial College Of Engineering And Research, Wagholi, Pune and Working as Assistant professor In Department of Computer Engineering at HSBPVT's COE college. She has 1 Year of experience. Her areas of interest in research is Data Mining.