A Novel Method To Summarize and Retrieve Text Documents Using Text Feature Extraction Based on Ontology

Aradhana R Patil, Amrita A Manjrekar

Abstract— **Data retrieval is a key process of acquiring information as per requirement. Now days, the necessity of proper information has increased. The most basic tools which provide this service are browser. It traverses the data as per user's query and gives the search results of all related information. Hence, it becomes a time consuming process to find required information. In this paper, the focus is done over content based data mining using ontology and text feature extraction. Content based data mining process focuses on domain of the data. *Ontology*, itself is a domain based data set information system that will help to achieve required data retrieval in a more appropriate way. The proposed systemuses*k means clustering algorithm* for creation of flat *clusters.*Flat clusters are the primary classification or clusters of data that are used for various further processing. For more appropriate data retrieval, this system uses *text feature extraction algorithm*. This algorithm will help to reduce the noisy data from data sets. A noise free data will help to perform better data retrieval process.**

Keywords- Domain ontology, Text data, Feature extraction, Flat clusters, information retrieval.

I. INTRODUCTION

Now days, growth of information has become an important process. For usability of large bulk of information, different methods and tools are usedto filter and process the data. Growth ofinformation primarily belongs to internet, because it is a main source of informationretrieval. Obviously, controlling or filtering such a large amount of information is a big issue.Current information retrieval systems are based on similarity of input query, which gives every possible output that matches with that query.

One approach to differentiate this bulk of information, based on their domain is Ontology.'Ontology is a formal explicit description ofconcepts in a domain of discourse (called classes or concepts), properties of each concept describing various features and attributes of the concepts are called slots and a knowledge base consists of the ontology combined with a set of instances.' [4]

It is not always necessary to create new ontology. It is possible to design new system with the help of existing or already designed ontologies.

Since ontology promises its usefulness in various fields, including web documentsextraction, grid computing etc., there is a scope of developing separate ontologies for better and approximate informationretrieval. Clustering of information needs to implement such modules.

Clusteringis similar to classification in that data are grouped. The groups are called clusters [18].

Ms.Aradhana R. Patil, Computer science and technology department, Department of technology, Kolhapur, India (arpcse12@gmail.com)

Ms.Amrita A. Manjrekar, Computer science and technology department, Department of technology, Kolhapur, India (aam_tech@unishivaji.ac.in)

The aim of clustering is to partition objects into clusters. This process of partitioning should have the following properties[16][17]:

Homogeneity within the clusters: data which belong to the same cluster should be as similar as possible.

Heterogeneity between the clusters: data which belong to two or more different clusters should be as dissimilar as possible.

Together these systems viz. Ontology and clustering can be applied for building a module which may give more proper information retrieval. The process of clustering initially creates flat clusters which can be related to the **domain** ontology. Hence it will beuseful for the fast access to the information, according to the domain of input query.

II. RELATED WORK

Yuxiahuang and Ling bian[1] proposed integration technique for heterogeneous tourisminformation for online tour planning using ontology and formal concept analysis (FCA). Mappingbetween two different ontologies i.e. tourist ontology and tourism information provider's ontology isdone.

Feasibility of ontology for solving multi document summarization problem for disastermanagement system is proposed by Lei Li and Tao Li [2]. The concept of domain ontology is used here. Various multi summarization modules are used viz. sentence mapping, sentence representation,generic summarization, and query focused summarization. Author states that domain ontology provesitself as a meaningful framework for semantic representation of textual information.

Qing He [3] proposed brief discussionabout the ontology development.

Kathrin Prantner[4] proposed the ontology designed for e-tourism using semantic webtechnology. Also this paper has stated reviews of the basics of ontology, which seems more particular. The focus of paper is to develop an ontology based over Accommodation,Activities and theInfrastructure for the activities.

BhaskarSinha, Somnath Chandra and MeghaGarg[11] proposed a recent work ofdeveloping ontology for Indian agriculture e- governance data using IndoWordNet which is a semanticweb approach.

An empirical analysis and survey of clustering algorithms for big data is proposed by Fahad, Alshatri, and Tari[13]. Concepts and algorithms related to clustering and a concise survey of existing (clustering) algorithms are studied.

An extremely fast text feature extraction for classification and indexing is proposed by George Forman, Evan Kirshenbaum [15]. In this paper, a fast method for text feature extraction that folds together Unicode conversion, forced lowercasing, word boundary detection, and string hash computation is described.SeemaWazarkar, Amrita Manjrekar [17] proposed Rough k means clustering for the creation of flat clusters. These flat clusters are then used for hierarchical clustering.

978-1-5090-0774-5/16/$31.00 © 2016 IEEE

III. PROPOSED SYSTEM

Internet is always the first preference for retrieval of any kindofinformation.When any user tries to find required information, web browsers searchesresultsaccording to the input query. A user will always get thousands of search results for a single word as an input.

This condition carves for the need of accurate or approximately correct information retrieval from web documents. A system which gives proper information as per user requirement will reduce many efforts and time as well.

In this context, for a text datawhich is primarily used to give a search query, a new approach is proposed here, i.e. collection of data using ontologies. This can be useful for searching more appropriate information according to domain of data. Hence this system is based on accuracy of data retrieval.Objectives of this system are as follows:

- Design a system module which reduce noise from data collection and feature extraction.
- Designing a system that performs creation of flat clusters with the help of text features.
- Data retrieval as per user request.

Figure 1 shows system flow diagram with different stages which are described as follows:

Stage 1: Data set collection

In this stage, the collection of available datasetscan be obtained via internet. Data sets will contain various pdf and word documents.

Standard open data set sources are available on internet, which can be used as reference ontologies. For e.g. various standard data sets such as data.gov.in, Datacatalogs.org, Google public data sets etc. will be used as per requirement.

Stage 2: Preprocessing

This stage performs cleaning of uploaded data. It includes removal of noisy data. Noiseis an error or variance in measured variable such as unnecessary information for e.g. missing values,inconsistent data etc.There are many procedures which cleans the noisy data. Some of them are Binning method,clustering, combined human and computer inspection, regression etc.

For analyzing the collected data, various criteria's can be used for e.g. collected data should be in the standard formats i.e. pdf and word. Also we can limit the data sets up to certain no. of files. This will automatically reduce percentage of noisy data from data sets.

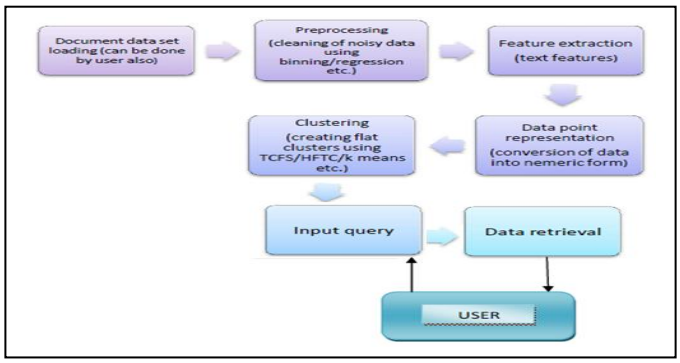

Fig.1:System Flow Diagram

The process of creating features for a given classification instance is called feature extraction. [9]. Textfeatures are to be used in this system for better data representation. Some of the text features are chart/table, appendix, fonts, diagram, graphs, and maps etc.

In this context, approach is followed as: selecting features, loading features and then extracting them. There are varioustext feature extraction systems already available in text mining. Standard text feature extraction algorithm is used for this purpose according to java environment.

Algorithm: Basic text feature extraction
Input: Various data sets consisting noisy data
Output: Feature extracted data
Step 1: Select feature (D, c, and k)
Step 2:V ← ExtractVocabulary(D)
Step 3:L ← []
Step 4:for each t ϵ V
Step 5:do A (t, c) ← ComputeFeatureUtility (D, t, c)
Step 6: append (L, <A (t, c), t>)
Step 7: return FeaturesWithLargestValues (L, k)

For a given class c, we compute a utility measure A (t, c) for each term of the vocabulary and select the k terms that have the highest values of A (t, c). All other terms are discarded and not used in classification. We will introduce three different utility measures in this section: mutual information, A (t, c)= I (Ut, Cc) ; the X^2 Test, A(t, c) = X^2 (t, c) ; and frequency, A(t, c)= N(t, c).

Stage 4: Data point representation

In this stage,extracted features should be converted into input from that will be suitablefor clustering, since clustering algorithm needs information in numeric form. Hence extractedinformation will be converted to numeric form**.**

Stage 5: clustering

Clustering is the most common form of unsupervised learning. This stage will do actual clustering over the extracted information. This will create flat clusters. Flat clustering creates a flat set of clusters without any explicit structure that would relate clusters to each other. Flat clustering is efficient and conceptually simple for primary clustering.

K-means clustering is one of the widely used algorithms for creation of flat clusters. In k-means clustering, a high degree of similarity among the elements in clusters is obtained [18]. Algorithm is stated below

Algorithm: K-means clustering [18].
Input: D={t_1,t_2,….,t_n}//set of elements
k // number of desired clusters.
Output: K // set of flat clusters

Step 1: assign initial values for means m_1, m_2,…..,m_k;
Step 2:Repeat
Assign each item t_i to the cluster which has closest mean; calculate new mean for each cluster;
Step 3:Until convergence criteria is met

As shown in Figure no. 2, a sample flat cluster tree is created for the data set 'document'. It consist two primary clusters as a pdf and word. And they are further classified with the type of documents that belong to themselves.

As stated in k means algorithm, the Cluster mean m_i of $K_i = \{t_{i1}, t_{i2}, \ldots, t_{im}\}$ is calculated as equation (1) below [18] :

$$m_i = 1/m \sum_{j=1}^{m} t_{ij} \quad \ldots\ldots\ldots\ldots (1)$$

Different validation equations are proposed by Fahad, Alshatri, Tari [13]to validate clustersuch as cluster accuracy (CA), normalized mutual information (NMI) etc. Equations are represented in equation (2) and (3) respectively:

$$CA = \sum_{i=1}^{k} \frac{max(C_i|L_i)}{|\Omega|} \ldots\ldots (2)$$

Where $C_i$is the set of instances in the ith cluster; $L_i$is the class labels for all instances in the $i$thCluster, and max $(C_{ij}L_i)$ is the number of instances with the majority label in the i^{th} cluster.

$$NMI = \frac{\sum d_{h,l} \log(|\Omega| d_h, l/d_h c_l)}{\sqrt{(\sum_h d_h \log (d_h/d)) (\sum_l c_l \log(c_l/d))}} \ldots\ldots (3)$$

Where $d_h$is the number of flows in class h, $c_l$is the number of flows in cluster l and d_h, l is thenumber of flows in class h as well as in cluster l.

Stage 6: Input Query

In this system, a query will be provided to obtain required information retrieval. User can upload customized data set or any standard data set as an input.

Stage 7: Data retrieval

This is a final stage where user will get result in form of flat clusters. Data set itself will contain folder of newly created flat clusters.

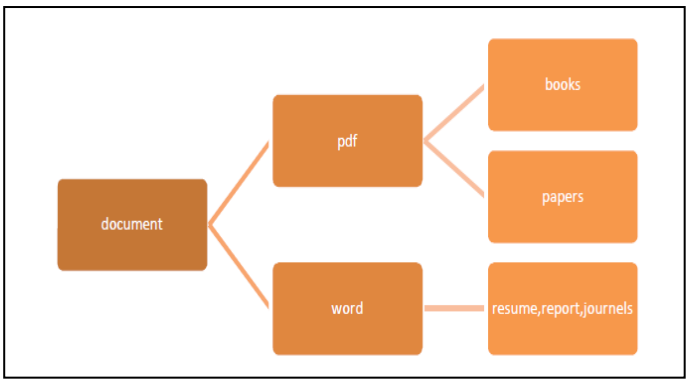

Fig.2: Flat cluster tree

IV. Experimental Results

In this system, user can select any data set directory consisting of the various text files. Those files must be in pdfand word format.

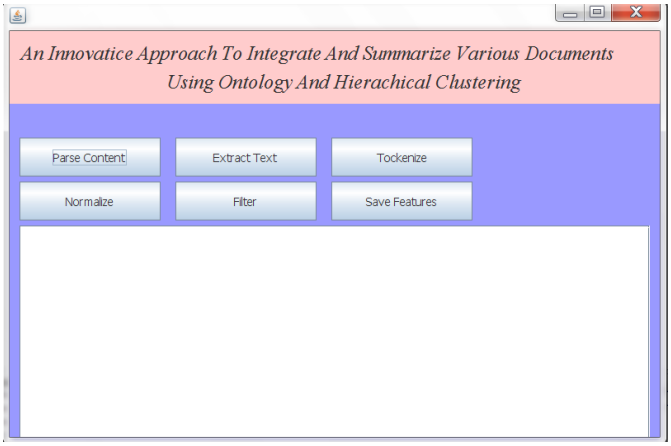

Fig.3: GUI of the proposed system

Preprocessing of given text file is carried out first. This stage performs processing of all the pages and contents of data set. After that, creation of the flat clusters will be done. The folder named 'cluster' is created at the same place where data set directory is present.

Next stage is a text Extraction process in which Parsing and extracting text data carried out. Datapoint representation is the process where various operations are done on given data to achieve some form. Here, it achieves one of the objectives of this system i.e. reducing noise in the text data. This process consists of tokenization, normalization and filtering.

In the tokenization, every single word or part of one sentence is assigned with the token; this is done to achieve primary data classification. The .TKN file is generated in the same folder. Normalization consist locating the noisy data in the given text file. Such as the commas, semicolon, point, repetition of the same word etc. this process generates .NRML file. Filtering includes removing the outliers from the given text file located in previous process. This achieves reduction of noise in data.

When these all processes are completed, the system generates a single file which saves the features of processed data. These features are calculated using text feature extraction algorithm. Figure no.4 shows the file that contains text feature extraction in numeric form. Here, every word is assigned with the unique ID, which is the frequency of that word in given text documents.

```
a:243
novel:3
method:23
to:186
summarize:7
and:288
retrieve:7
text:47
documents:15
using:38
hierarchical:70
clustering:194
based:44
on:70
:671
msaradhana:3
r:3
patil:3
computer:11
science:7
technology:27
department:15
of:400
kolhapur:7
india:7
email:7
arpcsegmailcom:3
msamrita:3
```

Fig.4: File containing extracted Features

Cluster accuracy of implemented system is compared with existing (old) system. The graph on figure no. 5 shows implemented system is more accurate than existing system though we increase no. of files in dataset.

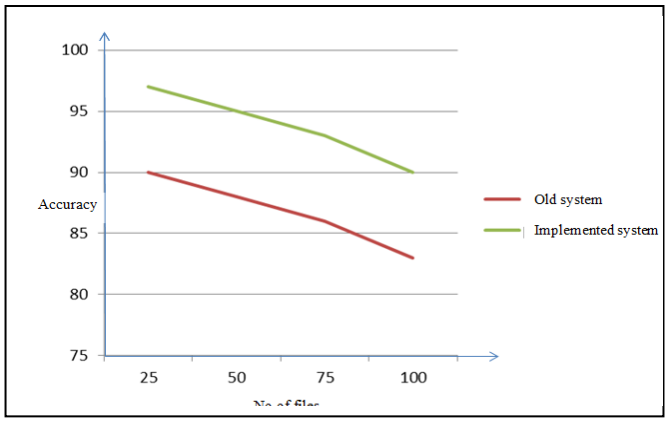

Fig.5: Accuracy Graph

Various data classification techniques are compared with this system. This comparision is made on the basis of data type compatibility. Figure no. 6 shows comparision of NEWSD, uClassify and BDS (buisness document system) with new system. NEWSD is the news classification system. BDS classifies buisness related documents such as pdf and word. uClassify is the system that classifies text data.

type	NEWSD	BDS	Uclassify	Implemented system
Webpages	✓	X	X	✓
Text	X	X	✓	✓
PDF	X	X	X	✓
Word	X	✓	✓	✓

Fig.6: Compatibility table

Time required for retrieval of each type of document differs from each other. Files in documents have different sizes, though the files have same no. of pages. Figure no. 7 shows graph with the variations in time required for data retrieval. For this analysis, 50 files are taken from each document types viz. pdf, word, text and webpages. as a different data sets.

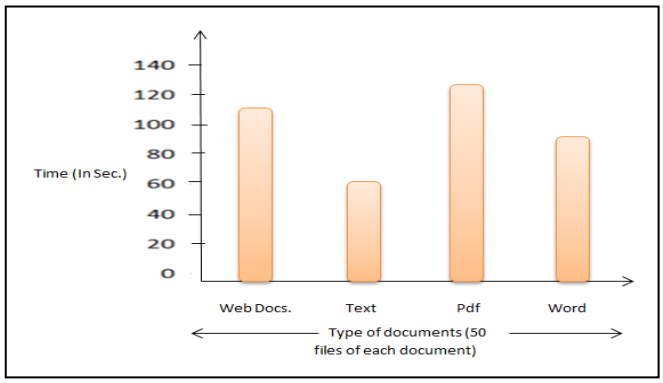

Fig.7: Graph showing time requirement for various document types

V. APPLICATIONS

This system has many useful applications to support existing data retrieval system, few of them are stated below-

- *A Recommendation*- Current browsing system offers the recommendation according to the criteria's which are generally used for classifications. For e.g. a website containing various books information will give recommendations based on the author, subject area and other. This system parses the internal data as the content based data retrieval is a key part of this system. This can be useful for recommendation of books to the user according to the content of the book.

- *Domain based classification*- Use of ontology in this system gives the benefit of domain based content classifications. This feature is also useful for the web data stores to classify their data systems.

- *Digital library management*- Data classification can be more proper and convenient using ontology and hierarchical clustering. Hence this system can be very useful for maintenance of various digital data libraries, which are needed to keep big amount of data. Domain based classification can provide more accuracy in context of classifying data.

VI. CONCLUSION

This system is helpful to summarize and classify the given data set. In context of clustering and retrieval of text data, system provides a fine classification of data set. This system performs creation of flat clusters, preprocessing(text feature extraction) and data point representation (normalization, filtering, tokenization etc.). This system achieves reduction in noisy data that helps for more accurate classification of data sets as flat clusters. In Future, Hierarchical clustering is to be performed over processed data. The advantages of ontology and hierarchical clustering will be helpful for the increase of processing speed of retrieval of information.

REFERENCES

[1]. Yuxiahuang and Ling bian, "Using Ontologies and Formal Concept Analysis to Integrate Heterogeneous Tourism Information" ,IEEE transactions on emerging topics in computing, volume 3,no. 2, June 2015

[2]. Lei Li and Tao Li, "An Empirical Study of Ontology-Based Multi-Document Summarization in Disaster Management", IEEE transactions on systems, man, and cybernetics: systems, vol. 44, no. 2, February 2014.

[3]. Qing He, "Ontology Development", a tutorial report for SENG 609.22., University of Calgary.

[4]. Kathrin Prantner, "On Tour- Ontology", www.deri.org.

[5]. Figueiredo, Julio C.dos Reis, and Rodriguez "Improving Access to Software Architecture Knowledge An Ontology-based Search Approach", International Journal Multimedia and Image Processing (IJMIP), Volume 2, Issues 1/2, March/June 2012.

[6]. Natalya FridmanNoy and Mark A. Musen, "An Algorithm for Merging and Aligning Ontologies: Automation and Tool Support", Stanford Medical Informatics Stanford University Stanford, CA 94305-5479.

[7]. FatenKharbat and Haya El-Ghalayini, "Building Ontology from Knowledge Base Systems", www.intechweb.org.
[8]. Jean Vincent Fonou- Dombeu, and Magda, "Combining Ontology Development Methodologies and Semantic Web Platforms for E-government Domain Ontology Development", International Journal of Web & Semantic Technology (IJWesT) Vol.2, No.2, April 2011.
[9]. Hotho, Stab, Stumme", Ontologies Improve Text Document Clustering", Institute AIFB, University of Karlsruhe, 76128 Karlsruhe, Germany.
[10]. Vincent SchickelZuber and BoiFaltings, "Using Hierarchical Clustering for Learning the Ontologies used in Recommendation Systems", Swiss Federal Institute of Technology EPFL Artificial Intelligence Laboratory Lausanne, Switzerland.
[11]. BhaskarSinha, Somnath Chandra and MeghaGarg, "Development of ontology from Indian agricultural e-governance data using IndoWordNet: a semantic web approach", Journal of Knowledge Management, Vol. 19 Iss 1 pp. 25 - 44(2015).
[12]. Ying zhao and George, "Hierarchical Clustering Algorithms for Document Datasets", Springer paper on Data Mining and Knowledge Discovery, 10, 141–168, 2005
[13]. Fahad, Alshatri, Tari", A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis", IEEE transactions on emerging topics in computing volume 2, no. 3, September 2014
[14]. Gang Liu, Ruili Wang, "A WorldNet-based Semantic Similarity Measure Enhanced by Internet-based Knowledge", 2009.
[15]. George Forman, Evan Kirshenbaum, "Extremely Fast Text Feature Extraction for Classification And Indexing", Conference on Information & Knowledge Management, Napa, CA Oct 27, 2008
[16]. S. V. Wazarkar and A. A. Manjrekar, "HFRECCA for clustering of text data from travel guide articles," *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference"* New Delhi, 2014, pp. 1486-1489.doi: 10.1109/ICACCI.2014.6968349
[17]. SeemaWazarkar, Amrita Manjrekar, "Text Clustering Using HFRECCA and rough k-means, Discovery, Volume 15, Number 40, April 8, 2014.
[18]. Margaret H. Dunham, "DATA MINING (Introductory and Advances)", PEARSON publications 7th edition.
[19]. MasoumehZareapoor,Seeja K. R " Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection", I.J. Information Engineering and Electronic Business, MECS, March 2015.
[20]. Gurpreetkaur, Abhilashsharma. "Feature Extraction techniques for Classification of Emotions in Speech Signals". International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 11, November 2014.