INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume2, Issue6)
Available online at: www.ijariit.com

An Improved Hierarchical Clustering for Information Retrieval System

Ila Shrivastava*
Computer Science and Engineering,
Acropolis Institute of Technology and Research,
Indore (Madhya Pradesh)
ilashrivastava.hind@gmail.com

Rahul Moriwal
Computer Science and Engineering,
Acropolis Institute of Technology and Research
Indore (Madhya Pradesh)
rahulmoriwal@acropolis.in

***Abstract*— Now in these days the information need is increasing rapidly in our day to day life therefore a large number of users are accessing data from search engine. The search engines are composed with three major components user query interface, search algorithm and the ranking process. During search process the system evaluate the user input query and the database documents according to best fit documents are retrieved. The retrieved document is then ranked according to the user query relevance thus most near document of the user query is listed first. The available technique are provides the ranked listing of documents.**

In this presented work first the recently developed text document retrieval models are evaluated and then after a traditional model of document retrieval is enhanced with help of supervised classification technique. The proposed data model of the document search first finds the document's word probability using the Bayesian classification approach then after the data is normalized to find the similar length of text document features. These document features are used to make training of neural network .The neural network processes the input training features and makes training for the documents pattern. This data model is used to predict the user input data patterns from the existing set of data.

The implementation of the proposed technique is performed using the JAVA development technology after implementation of the desired document retrieval technique the performance of the system is estimated in terms of accuracy, error rate, memory consumption and the time consumption. According to the evaluated results the performance of the algorithm is found more optimum. Thus the given model is more adoptive as compared to the traditional approaches available.

***Keywords*— Information, Text retrieval, Neural network, Data mining, Classification Algorithm.**

I. INTRODUCTION

Information retrieval is considered as a research field of computer science that deals with organising and searching "information". Information retrieval is the form of presenting the most applicable information to the searcher. Information retrieval System provides users with access to immense collection of stored information. System composed of Structure, analysis, storage and probing of information which involves collection like many data types such as text, images, audio, and video and so on.... Searches can be depend upon metadata or on full-text indexing. An automated information retrieval system reduces what has been called "information overload". Web search engines are the most perceptible Information Retrieval applications. To find the useful information from the large data source is a complicated and much frustrating task. Therefore efficient and effective content predicted systems are required in order to enhance the traditional way of information retrieval.

In this era of computational intelligence most of the data resources are available in text format. The amount of data is too big to analyse and finding the important knowledge is too complex. Therefore, that is an interesting domain of research and development. The text categorization includes a wide range of applications that is component of knowledge process and natural language processing in artificial intelligence. The content mining, artificial intelligence and content mining techniques that include sundry applications such as semantics analysis, compiler design, document and sizably voluminous data analysis are the sub domain of text and semantic analysis.

The proposed work is to improve the technique of automatic text categorization and extraction process. Fundamentally, the text mining is a domain of unsupervised learning techniques as the data is always found in unstructured way. Additionally, text categorization includes the sundry semantically and statistical issues. Therefore proposed work presents a hybrid approach to optimize the process of text categorization and retrieval using the unsupervised learning techniques. This technique is employed for finding the essential features form the large text documents for preserving the computational resources during text analysis

II. PROPOSED WORK

Text mining is an interesting domain for knowledge processing and business intelligence. In this domain the unstructured data is mined for finding expensive and fruitful data discovery. Therefore, in order to explore the domain of text mining and to find the appropriate the solution for text mining and information discovery required to involve the following work to include in the proposed study.

- ***Study of various text mining techniques*:** in this phase of the study work various text mining techniques are observed by which the guidelines can be obtain to find the problem domain and solution domain of the work.
- ***Design and implementation of the new text mining technique*:** in this phase a new text processing algorithm is implemented for finding the relevant text from rich text data sources which will be more efficient and accurate.
- ***Performance study:*** the performance of the proposed system is evaluated in this phase in terms of accuracy, memory consumption, categorization time, and error rate

A. PROBLEM DOMAIN

In information retrieval system there are two different approaches are used for extracting the information from the data base. In first technique the model is responsible for data retrieval and in second method the query semantics played essential role. Thus in order to find more significant documents according to the user input queries the different retrieval systems were developed but previous technologies were not able to provide the specific search result in terms of relevance feedback, Symantec gap, ranking & re-ranking. Therefore we proposed an efficient and intelligent technique of search and query processing that compare the input user query to the available documents text and more nearer documents are retrieved. Finally the ranking process is used to enhance the listing of search results by the search algorithms .we focused over the issues to handle the text based data retrieval system

B. SOLUTION DOMAIN

The proposed solution is aimed to resolves the issues listed in above section therefore the following solution is suggested to implement accurate and efficient classification scheme.

Firstly to implement a dynamic pre-processing system which include following steps as

- Upload the file using upload function.
- Then the file is pre-processed. In this step the HTML tags, scripts, styles are removed from the file and it is saved as a text document.
- Then the text is tokenized.
- After tokenizing the text, the stop words and special characters are removed.

After these the proposed data model of the document search first finds the document's word probability using the Bayesian classification approach then after the data is normalized to find the similar length of text document features. These document features are used to make training of neural network. The neural network processes the input training features and makes training for the documents pattern. This data model is used to predict the user input data patterns from the existing set of data.

In addition of that the given retrieval model only processes the web documents and the text documents. Thus that is also required to enhance the traditionally available technique for different other formats such as PDF documents, PPT slides, HTML documents and rich text documents. In order to implement such kind of method a new technique

using hybrid techniques of back propagation neural network, Bayesian classifier and KNN is proposed. The detailed implementation technique of the proposed methodology is given in next section.

C. SIMULATION ARCHITECTURE

In order to understand the proposed technique of document retrieval model the given figure 3.1 helps. In this given diagram the document processing of the system is demonstrated. First the document is provided to the bay's model where the input documents are evaluated and their features are computed.

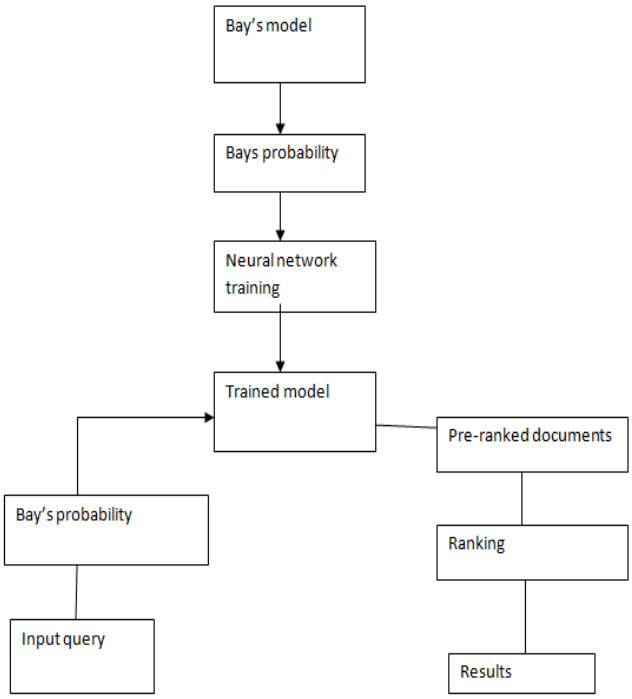

Fig 1 Simulation Architecture

The given figure 1 shows the simulation architecture of proposed system in this context the entire system is divided into a set of modules. The entire modules are described as:

1) Bayesian classifier: The Naive Bayes classification algorithmic rule is a probabilistic classifier. It is predicated on probability models that incorporate robust independence postulations. The independence postulations conventionally don't have an effect on reality. So they're thought of as naive. You can derive probability models by utilizing Bayes' theorem (proposed by Thomas Bayes). Predicted on the nature of the probability model, the training of the Naive Bayes algorithm program is done in a much supervised learning way. In straightforward terms, a naive Bayes classifier assumes that the value of a concrete feature is unrelated to the presence or absence of the other feature, given the class variable.

2) BPN Algorithm:

The implementation of neural network is defined in two phases' first training and second prediction: training method utilizes data and designs the data model. By this data model next phase prediction of values is performed.

Training:

1. Prepare two arrays, one is input and hidden unit and the second is output unit.
2. Here first is a two dimensional array W_{ij} is used and output is a one dimensional array Y_i.
3. Original weights are random values put inside the arrays after that the output is given as**.**

$$x_j = \sum_{i=0} y_i W_{ij}$$

Where, y_i is the activity level of the j^{th} unit in the previous layer and W_{ij} is the weight of the connection between the i^{th} and the j^{th} unit.

4. Next, action level of y_i is estimated by sigmoidal function of the total weighted input.

$$y_i = \left[\frac{e^x - e^{-x}}{e^x + e^{-x}}\right]$$

When event of the all output units have been determined, the network calculates the error (E) given in equation.

$$E = \frac{1}{2}\sum_i (y_i - d_i)^2$$

Where, y_i is the event level of the j^{th} unit in the top layer and d_i is the preferred output of the j_i unit.

The detailed processes of neural network implementation is discussed in this section, the estimated words and their probability to be find in a document is produced in neural network where the neural network perform training on the input data of fixed features. After the neural network training the model enabled to classify the data therefore the trained data model is used for extracting the text documents.

Now when the user provides the query for retrieving the documents then the user query tokens and their probability for each document is estimated first and then after the query tokens and their probability is provided as input to the neural network. Neural network processes the input probability of user input query. The neural network produces the initial documents as retrieved documents. Now the KNN algorithm is taken place to rank the retrieved documents. Thus the KNN accepts the two inputs first the user query tokens as the query scenario and the available documents as the database scenario. The algorithm finds the most nearer searched documents as top of search results. The KNN algorithm for document retrieval is given as:

3) K**-nearest-neighbour (KNN) Algorithm**

The K-nearest-neighbour algorithm measures the distance between a query scenario and a set of scenarios in the data base. The distance between these two scenarios is estimated using a distance function d(x,y), where x, y are scenarios developed through features, like [20]

$$X = \{x_1, x_2, x_3, \ldots\}$$
$$Y = \{y_1, y_2, y_3, \ldots\}$$

The frequently used distance functions are absolute distance measuring using:

$$\mathbf{d_A(x,y)} = \sum_{\mathbf{i=1}}^{\mathbf{N}} |\mathbf{x_i - y_i}|$$

And second is Euclidean distance measuring with:

$$\mathbf{d_A(x,y)} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \sqrt{\mathbf{x_i^2 - y_i^2}}$$

To find closest instance

The overall KNN algorithm is running in the following steps:

1. Store the output values of the M nearest neighbours to query scenario Q in vector r = $\{r_{1,\ldots\ldots},r_m\}$ by repeating the following loop M times:
    a. Go to the next scenario S_i in the data set, where I is the current iteration within the domain {1……P}
    b. If Q is not set or q < d (q, S_i): q ←d (q, S_i), t ←O_i
    c. Loop until we reach the end of the data set.
    d. Store q into vector c and t into vector r.
2. Calculate the arithmetic mean output across r as follows:

$$\bar{r} = \frac{1}{M}\sum_{i=i}^{M} r_i$$

3. Return r as the output value for the query scenario q

III.PROPOSED ALGORITHM

The proposed system architecture is divided into two major parts first training and second testing of application. in the first module train the classifier using domain keyword and second module is the test module.

The proposed methodology of text document information retrieval technique can be summarized through the algorithm steps given below.

Training of data model

1. for each document in the database
2. $word\ frequency[] = \frac{number\ of\ times\ a\ word\ appeared}{total\ number\ of\ words\ avialbe}$

3. End for
4. Sort the word frequency data
5. Save to database
6. Initialize the neural network
7. train network using computed work probability

Query processing

1. Input user query
2. find probability of each token in user query
3. input to the trained neural network
4. classify the input query pattern
5. apply KNN for rank the neural network results
6. return ranked results

IV RESULTS ANALYSIS

After implementing the desired algorithm for document retrieval system the performance of the system is estimated in order to find efficacy in terms of resources consumption (i.e. memory consumption and time consumption). Additionally, the precision and error rate are provided to improve the efficiency of the system.

***1. ACCURACY*:** The accuracy of the system is the measurement that demonstrates how efficiently a document recognizes their domain in order to categorize under a specific domain. The accuracy of the system is measured during different experiments and different documents, and best obtained results are listed in this section..The following formula is utilized for calculating the accuracy of the desired proposed system

$$accuracy = \frac{total\ correctly\ classified\ data}{total\ samples\ avilable\ to\ classify} X100$$

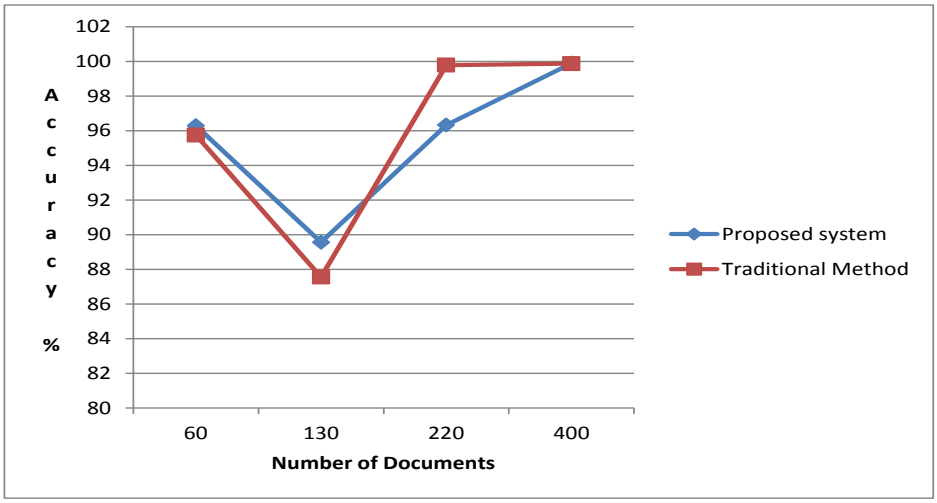

Figure 4.1 Accuracy of system

The comparative accuracy of the proposed system is given in Figure 4.1 that shows the precision of the proposed model of document retrieval system. In the above diagram X axis involves different number of documents on which experiments performed with the system and the Y axis involves the obtained performance of algorithm in terms of precision. According to the evaluated results the performance of the proposed technique enhances their precision as the amount of data for training is increases in the data base. Thus the proposed data model is relevant for sizably voluminous data.

2. ERROR RATE: Error rate of the system reflects the outcome is how far from the existing solution, therefore the error rate of the system can be given using the below formula

$$error\ rate = 100 - accuracy$$

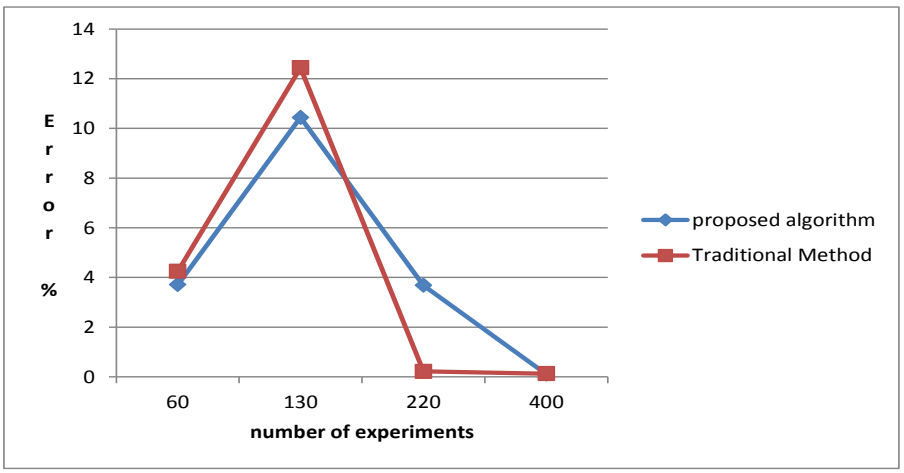

Figure 4.2 Error rate

The figure 4.2 shows the comparative performance of the proposed system and traditional IR model in terms of error rate. The figure contains the error rate in Y axis and the X axis contains the different number of documents by which experiments performed with the algorithm with increasing amount of data. According to the evaluated performance the error rate of the algorithm is improved as the size of data is increases for evaluation thus the proposed model is adoptive model and reflect the efficient performance even when the data is increases in data base.

***3. MEMORY USAGE*:** The amount of memory consumed during the process execution is known as the memory consumed. The figure 5.3 shows the memory usage by the system that is calculated on the basis of peak working set.

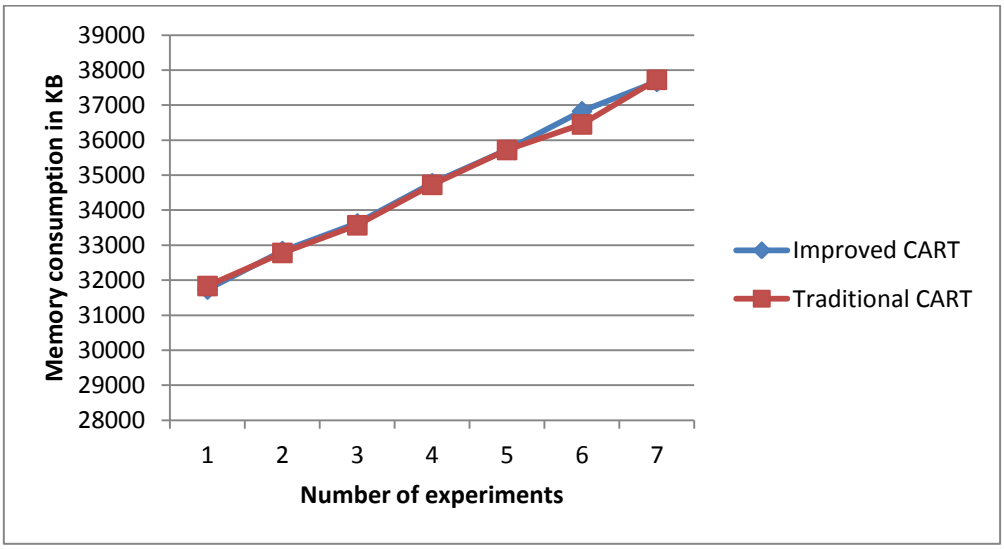

Figure 4.3 Memory usage

The amount of main memory consumed during the processing of data using the proposed algorithm is given using figure 4.3. In this diagram the memory consumption of the algorithm is given using Y axis which contains the memory consumption of the algorithm in terms of KB (kilobytes) and X axis contains number of experiments performed with the system. According to the obtained results the memory consumption of the algorithm is not much effected with the amount of data increases in database but that are slightly increases as with the amount of processing data

***4. RETRIEVAL TIME*:** The total amount of time is consumed during the retrieval of documents are known as clustering time or time complexity of the selected algorithm. The figure 5.4 shows the time consumption for finding and learning of the data to retrieve the accurate text from database. The time is given using figure 4.4

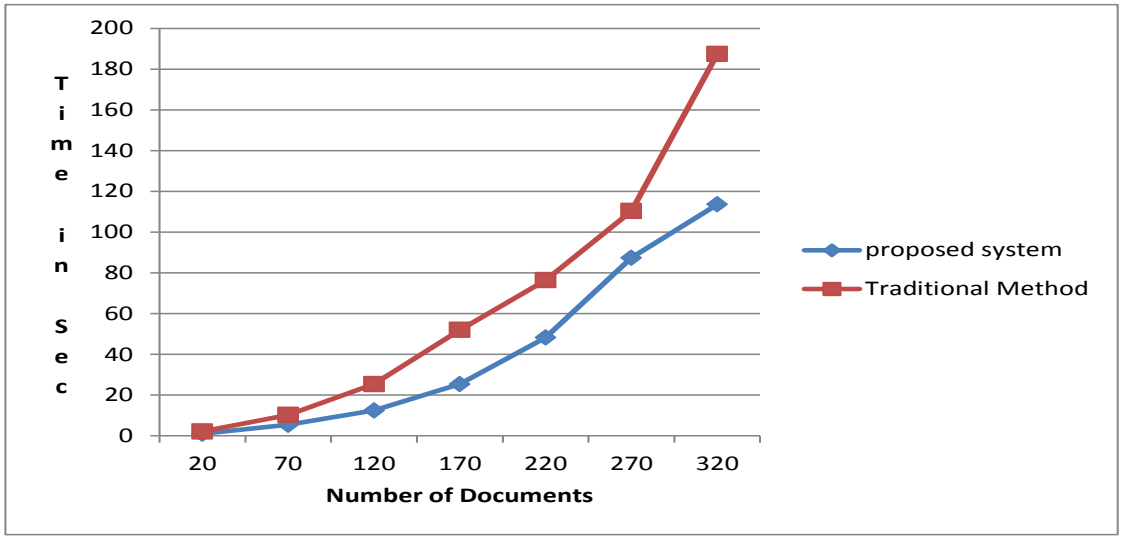

Figure 4.4 Time consumption

The given figure 4.4 shows the comparative performance of the proposed SOM based document retrieval data model and traditional technique of information retrieval in terms of the time complexity. In order to reveal the performance of the proposed algorithm the X axis contains the different amounts of documents that are used for experimentation and the Y axis contains the amount of time consumed for data analysis. According to the obtained results the amount of time for retrieving data is higher if the amount of data in training data is large. Thus that can be achieved time complexity of the proposed system is optimal and consumes an adoptable duration.

CONCLUSION

The cyber world predicated information retrieval system supports the information extraction from the raw set of data existing in the web directory. Therefore a machine learning technique is utilized for extracting information efficiently and accurately. In order to obtain more such model various recent contributions on the information extraction is evaluated and a promising model is observed as given in [3]. In the model described here, the author recommended to implement a neural network based technique that helps to learn the documents patterns and according to the document patterns the information from the web database is extracted.

In the available model the three issues are focused for rectification. First the supporting of different file formats of data such as docx, PDFs, pptx, html and text etc... while second issue is the unsupervised algorithm that provides less precise results as compared to the supervised learning concept. Thus the proposed system incorporates the back propagation neural network as the supervised learner of the search algorithm with the help of Bayesian classifier. Additionally the use of third party API the text extraction techniques are implemented to support more than one format of data.

The proposed technique is implemented with the help of Java technology and their performance is evaluated. Finally the performance estimation of the proposed clustering technique is quantified in terms of of memory consumption, search time, accuracy and their error rate..The system is adaptable due to its high accurate predictive results and the less resource consumption in terms of time and space involution the performance of the implemented technique is summarized using the table 5.1.

S. No.	Parameters	Remark
1	Accuracy	Accuracy is increase as the amount of data is training database is increases

2	Error rate	The error rate of the algorithm is decreases as the amount of data in training set is increases
3	Memory consumption	The linearly increasing memory consumption with the amount of data
4	Time consumption	The time consumption of training is increases as the amount of data is increases

Table 5.1 performance summary

The implemented algorithm is evaluated in different performance parameters and their performance found optimum during different data formats search. Thus the proposed technique is much adoptable as compared to the traditional data model.

FUTURE WORK

The key objective of the proposed work is to enhance the traditional information retrieval model for improving relevancy and the resource consumption. The basic objective of the proposed work is accomplished by implementing the hybrid approach of neural network and Bayesian classifier. But the accuracy of the given model can be improvable thus in near future the given model is refined more for enhancing the recognition accuracy.

REFERENCES

[1]B.V.Rama Krishna, B. Sushma, "Novel Approach to Museums Development &Emergence of Text Mining", ISSN 2249-6343, International Journal of Computer Technology and Electronics Engineering (IJCTEE), Volume 2, Issue 2,2008

[2] H. P. Luhn, "A Business Intelligence System", Volume 2, Number 4, Page 314 (1958), Nontopical Issue,IBM Research Journals

[3]Larbi GUEZOULI, Amine KADACHE, "Information retrieval model based on neural networksusing neighborhood", 2012 International Conference on Information Technology and e-Services

[4]Andreas Hotho, Andreas Nurnberger, Gerhard Paaß, FraunhoferAiS, "A Brief Survey of Text Mining", Knowledge Discovery GroupSankt Augustin, May 13, 2005

[5]Umajancy.S, Dr. Antony SelvadossThanamani, "AN ANALYSIS ON TEXT MINING –TEXTRETRIEVAL AND TEXT EXTRACTION", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 8, August 2013

[6]MilošRadovanović,MirjanaIvanović, "TEXT MINING: APPROACHES AND APPLICATIONS", Abstract Methods and Applications in ComputerScience (no. 144017A),Novi Sad, Serbia,Vol. 38, No. 3, 2008, 227-234

[7]Vishal Gupta,Gurpreet S. Lehal, "A Survey of Text Mining Techniques andApplications",JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, NO. 1, AUGUST 2009

[8]P.Bhargavi, B.Jyothi, S.Jyothi, K.Sekar, "Knowledge Extraction Using Rule Based Decision TreeApproach", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.7, July 2008

[9]G. KoteswaraRao and ShubhamoyDey, "DECISION SUPPORT FOR E-GOVERNANCE: A TEXT MINING APPROACH", International Journal of Managing Information Technology (IJMIT) Vol.3, No.3, August 2011

[10]Daniel Ramage, Christopher D. Manning,Susan Dumais, "Partially Labeled Topic Models forInterpretable Text Mining", KDD'11, August 21–24, 2011, San Diego, California, USA.Copyright 2011 ACM 978-1-4503-0813-7/11/08.

[11]Zhong,Ning, Li, Yuefeng, & Wu, Sheng-Tang, "Effective pattern discoveryfor text mining". IEEE Transactions on Knowledge and Data Engineering, (2010)

[12]Anne Sunikka, Johanna Bragge, "Applying text-mining to personalization and customization researchliterature – Who, what and where", 2012 Elsevier Ltd. All rights reserved

[13]Koby Crammer,Mark Dredze,Fernando Pereira, "Confidence-Weighted Linear Classification for Text Categorization", Journal of Machine Learning Research 13 (2012) 1891-1926 Submitted 12/10; Revised 12/11; Published 6/12

[14] Roberto Navigli, Stefano Faralli, AitorSoroa, Oier de Lacalle, EnekoAgirre, "Two Birds with One Stone: Learning Semantic Models for Text Categorization and Word Sense Disambiguation", CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK. Copyright 2011 ACM 978-1-4503-0717-8/11/10

[15]FuzhenZhuang,PingLuo,HuiXiong,QingHe, YuhongXiongand Zhongzhi Shi, "Exploiting Associations between Word Clusters and Document Classesfor Cross-domain Text Categorization",Received 26 April 2010; revised 7 October 2010; accepted 27 October 2010, DOI:10.1002/sam.10099, Published online 30 November 2010 in Wiley Online Library.

[16]Deng Cai and Xiaofei He, "Manifold Adaptive Experimental Design for TextCategorization",Manuscript received 26 Aug. 2009; revised 08 May. 2010; accepted 17 Sep.2010

[17]Luís Moreira-Matias, João Mendes-Moreira, João Gama, and PavelBrazdil, "Text Categorization Using an Ensemble Classifier Based on a Mean Co-association Matrix",P. Perner (Ed.): MLDM 2012, LNAI 7376, pp. 525–539, 2012. © Springer-Verlag Berlin Heidelberg 2012

[18]Jin Ming Koh, Marcus Sak, "Efficient Data Retrieval forLarge-Scale Smart City Applicationsthrough Applied Bayesian Inference", 2015 IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)

[19]JunbeomHur and Kyungtae Kang, "Secure Data Retrieval for DecentralizedDisruption-Tolerant Military Networks", IEEE/ACM TRANSACTIONS ON NETWORKING, VOL 22, NO 1, FEBRUARY 2014

[20]Rui Zhang, H. V. Jagadish, Bing Tian Dai, KotagiriRamamohanarao, "Optimized Algorithms for Predictive Range and KNN Queries on Moving Object", Volume 35, Issue 8, December 2010, Pages 911–932, 2010 Elsevier

