ISSN: 2277-5528
Impact Factor: 5.085

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & MANAGEMENT

STUDY OF DOMAIN CLASSIFICATION OF RANDOM OFFLINE AND ONLINE DATA

Raj Kumar, Ms Puja Trivedi
Department of Computer Science Engineering
RKDF School of Engineering, Indore (M.P.)

ABSTRACT

The search engine like google provides record of results that shows a list of ranked output. The ranking does not consider the subject of the file. The results of search engine are not in a well-defined group. This may also be frustrating, as the users have to scroll through many inappropriate results. This could come up when the user is a beginner or has superficial capabilities about the domain of interest, however more as a rule it is due to the question being brief and ambiguous. One answer is to organize search results through categorization, in specific, the classification. A goal of testing is to test designed on a controlled data set, which shows that classification-bounded search could enhance the person's search expertise in terms of the numbers of results the person would must check out earlier than pleasing his/her query.

This work uses the naive bayes classifier, which is a simple and effective method for establishing classifiers. The proposed model for finding domain, related to user query based on document index matrix. The proposed implementation combine the both approach simultaneously which is term based and phrased based. Document index matrix used term, phrased based document matrix in such a manner that it is compare with training data, and put them into relatively domain. The naïve bayes algorithm used to find maximum probability occurrence from both the matrix. The output comes in the form of suggestion domains list. user easily retrieve the data with minimum time.

INTRODUCTION

Search engines are web portals for finding information on the Web. Search engines index a large Section of the web and store the information in databases. Unlike a web directory, the search engine carries operations automatically. The underlying technique of search engines is information retrieval. The accessible question in information retrieval is how to find the relevant documents for each user query. This is the main issue that try to address.

Knowledge mining is a procedure of mining information from the raw information. Extraction of the similar text from a raw set of text is the generation of text data mining. Clearly, text data belongs to an unstructured method and labelling of information is tricky undertaking, therefore, many of the applications are utilizing the classification approaches for categorizing knowledge.

Text classification captures the relevant result for each user query, Naïve Bayes classifier is a simple and effective approach to classify text document, which uses probabilistic classification technique. Naïve Bayes classifier using Bayes' theorem for classifying unknown retrieved data from google search engine and few modifications are there to increase performance of classifier.

The use of search engines increases rapidly, users of search engine faces problems related to number of replies according to their query. Google search engine returns ranked list of results, which are not relevant to users they are scrolling and finding specific result. The motivation of our study is that queries submitted to a search engine

ISSN: 2277-5528
Impact Factor: 5.085

may have multiple meanings. Example, depending on the user, the query "apple" may refer to a fruit, the company Apple Computer or the name of a person, and so forth. Thus, providing categorized query (user, interested in "apple" as a fruit get invitations about fruit, while users interested in "apple" as a company get suggestions about the company's products) certainly, it helps user's formulate the more effective queries according to their needs. The underlying idea of our proposed technique focuses on concepts and their relations extracted from the submitted user queries, the web content and training set.

TECHNIQUES OF TEXT MINING

This section introduces the distinct types of approaches and the text sample analysis with mining. Probably this discusses the used major systems.

Information Extraction

Preliminary factor for pcs to analyse unstructured textual content is to make use of expertise extraction. Expertise extraction application recognizes key phrase and associations inside the textual content. This uses the predefined sequences in text data that process alls the pattern matching. The software concludes the relationships among all the identified objects to provide the user relevance information from text data sources. This technique can also be very useful when a giant volume of text information is required to analyse. Classical data mining accepts that the information to be "mined" is already in the form of a relational database [7].

Topic Tracking

A topic tracking works by keeping user interest as profiles and, according to their interest, they document the user views. Yahoo makes it possible for users to select key words when information concerning those themes. Topic tracking technology does have limitations, however. Example, if a consumer creates an alert for "text mining" user will seize news stories, and issues, which can be sincerely required. Number of better text mining tools let user choose particular domain of interest or the software automatically can the user's interests according to browsing history and click information [8].

Summarization

Text summarization is most helpful for making efforts to draw whether or not an extensive document encounters the user's needs for further information. Massive amount of text, text summarization technique helps to processes and summarizes the document in a paragraph. The most important goal of summarization is to diminish the amount of text data and element of a document and emphasise its predominant features, which produce the total meaning of the report. The predominant assignment is that, despite the fact that desktops are ready to establish objects in files.

Categorization

Categorization techniques involve the identification of the main subjects of a document by providing the document into a pre-defined set of subjects. When, categorization take location to a report, than it will deal with the file as a group of words. Categorization does not attempt to procedure the common understanding as knowledge extraction does. On the other hand, categorization only calculates the found words and, using this calculations, the main topics identifies. Categorization usually depends on dictionary-based topics and associations are recognize by observing for frequent terms, synonyms, and interrelated terms. Categorization

ISSN: 2277-5528
Impact Factor: 5.085

techniques commonly have a method for ranking the documents by which documents have the most content on a particular subject [6].

Clustering

Clustering is a method to group similar kind of documents, but that is different approach from categorization process. Categorization process creates group of documents based on the predefined domains or subjects. The main benefit of clustering is that documents can represent in multiple subtopics, therefore, confirming that a useful document will not be obtained from search results. A basic clustering technique creates a list of topics for each document and calculates the weights for how accurately a document fit into a group. Clustering methodology can be helpful in the organization of information management applications, which may include a thousand of documents [7].

Concept Linkage

Concept linkage systems associate similar documents by recognizing their commonly distributed concepts and help to the users for finding information that they have not found using classical search approach. Concept Linkage supports information browsing reasonable than searching. Concept linkage is an appreciable technique in text mining, particularly in the biomedical domain where performs too many research and it is impossible for scholars and researchers to read all the documents and create relationship among each other. Preferably, concept-linking techniques can find connectivity among diseases and treatments with any human effort.

Information Visualization

Visible text mining, or information visualization, is process for giant textual sources to represent in a visible hierarchy. Visual hierarchy presents searching capabilities, moreover for making simple search DocMiner is a gadget that demonstrates visually for enormous quantities of textual content to analyse the content. User can interact with the document hierarchy by scaling, zooming, and creating sub-hierarchy. Information visualization is useful technique if a user requires reducing a large range of documents and wanting to explore related subjects.

Question Answering

It is software of traditional language processing is natural language queries, or question answering. This software handles the issues to search out the quality reply of a given question. Various websites that uses the question answering methodology. That allows end users to ask a question to the computer and to provide an appropriate answer. This technique can consume multiple text mining methods.

Association Rule Mining

Association rule mining is a way used to find association between a large set of variables in available data set. Association rule Mining uses in variety of industry application and disciplines but not widely used in the social sciences. Association rule mining denotes to the discovery of associations between a massive set of variables, that is, available in a database, each includes two or more attributes and their particular values. Association rule mining regulates variable-value permutations that repeatedly occur which is much similar to the concept of correlation analysis, in which associations between two attributes are not covered. Association rule mining is also used for discover fluctuating connections, but each relationship may include two or more attributes.

ISSN: 2277-5528
Impact Factor: 5.085

This section supplies the overview of text mining approaches that text information becomes classifiable in subsequent we discuss the not too long ago development procedures and trends in textual content mining.

PROPOSED MODEL AND ALGORITHM

We provides the detailed study about the traditional Agglomerative clustering and the improved semi-supervised agglomerative algorithm, which used for developing the proposed semantic text clustering technique. Agglomerative algorithm is a kind of hierarchical algorithm and frequently used for clustering textual contents with no previous class and semantics knowledge.

Figure 1 shows the proposed model, which describes overall steps in system. First, user gives an input a search query to search engine that returns a list of ranked results then classification applied to that result and user gets categorized results.

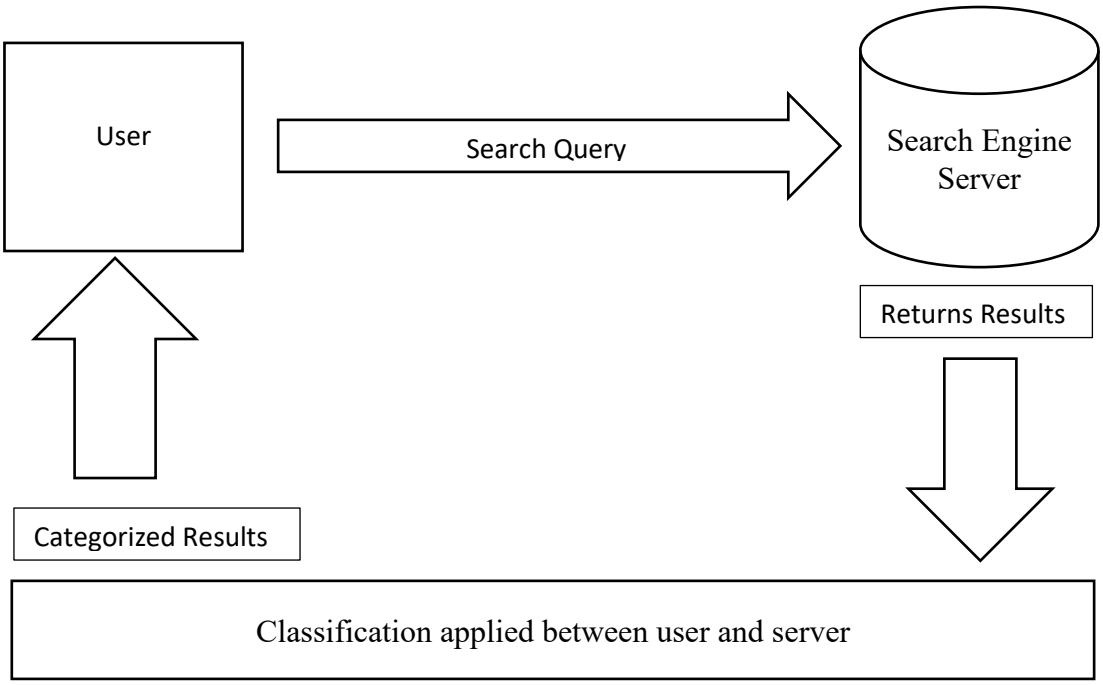

Figure 1 Model

Document matrix stores the record of term and phrased value of different domain in proposed work.

Document matrix is of following two types

1) Term wise document matrix:
2) Phrased wise document matrix:

Term wise document matrix: In this matrix,we collect all the terms from the given text and put into respected trained domains.

Figure 2 shows flow chart of proposed algorithm

ISSN: 2277-5528
Impact Factor: 5.085

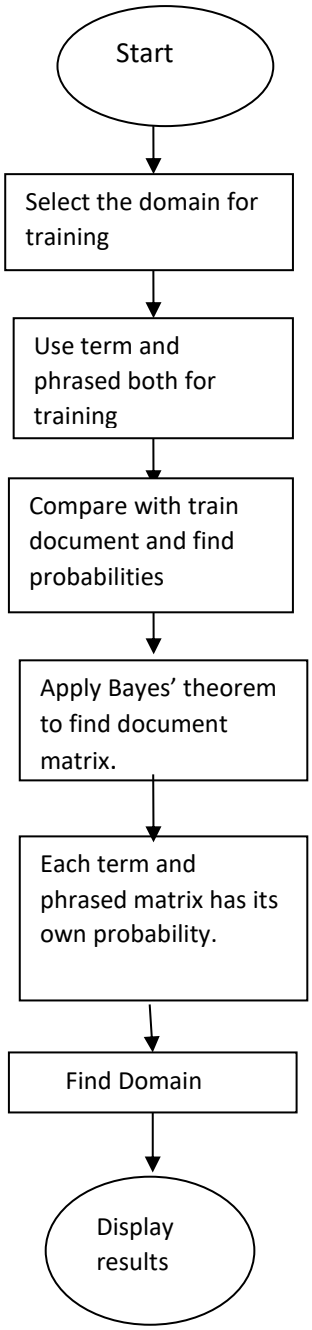

Figure 2 Flow Chart of Proposed algorithm.

Proposed algorithm shows the overall process of getting domain from the unknown dataset.

Step 1: Input unknown web text, Output Domain finder.

Training Phase:

Step 2: Select the domain to train.

Step 3: Give the related words and sentences (phrased) to particular domain.

Testing Phase

Step 4: Calculate the frequency of term and phrased respectively of given web document.

ISSN: 2277-5528
Impact Factor: 5.085

Step 5: Apply Naive Bayes to get probability of occurrence of different term and phrased.

Step 6: Compare only those words, which is having high probability of occurrence with training domain.

Step 7: Display classified domain.

IMPLEMENTATION AND RESULT

The proposed model classifies the domain of google search results, in this chapter the implementation of the proposed system. Therefore, this describes the required tools and techniques with implemented scripts in detail. Different UML diagrams (such as use-case diagram, system context diagram and state chart diagram) describe the desired software (proposed classifier) features in detail. These diagrams are intended to explain the software in satisfactory detail that programmers may build up the software with minimum extra effort.

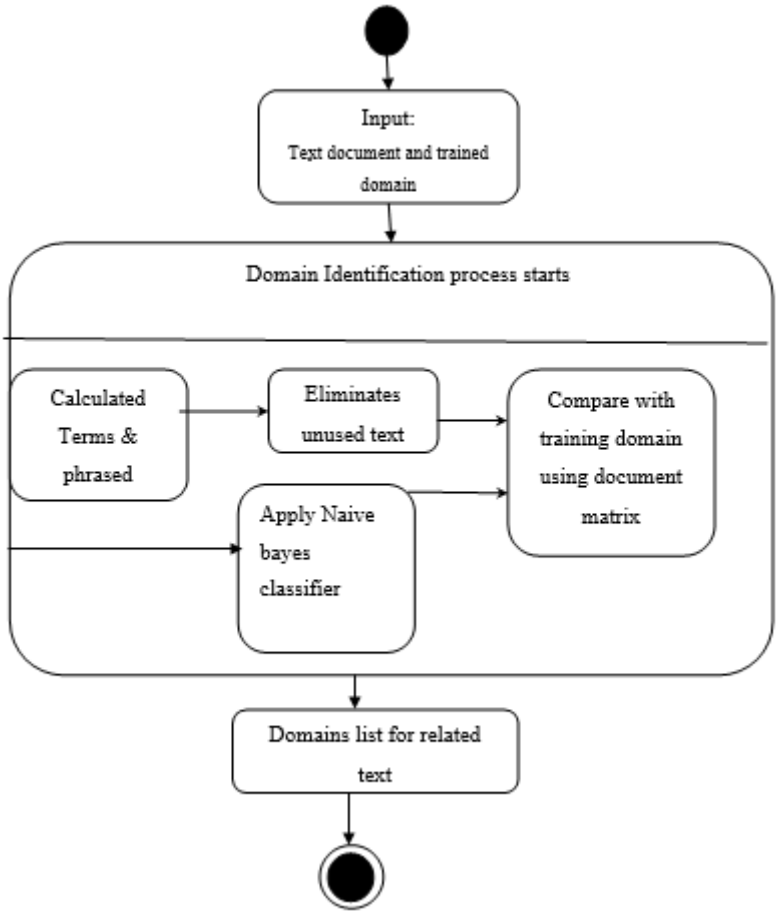

Figure 2 The State Chart Diagram of Proposed Domain identifier.

RESULT ANALYSIS

This section discusses the different types of graph and table related to our result. This shows the comparatively results of simple search and proposed search. The X-axis contains the categories and the Y-axis contains time consumed in terms of milliseconds in these diagrams. According to the comparative results analysis the performance of the proposed technique shows the less time consumption as compared to the traditional technique.

ISSN: 2277-5528
Impact Factor: 5.085

- **Search Result For Java**

Table 1 gives time consumption for query java, the category wise search result related to java query. There are five categories are found related to java keywords which is hardware, internet, multimedia, networks and security. This table gives the various comparisons performed between simple search and using modified Bayes search. The result shows the improved search is better than simple search. Improved search uses minimum time in all aspects.

Table 1Time consumption for query java

Categories	Simple Search	Improved Search
Hardware	380	270
Internet	370	310
Multimedia	280	230
Networks	210	190
Security	340	260

The time consumption of the proposed and tradition algorithms for search query java. The graph shows various attributes related to the maximum number of occurrence search result related to java query

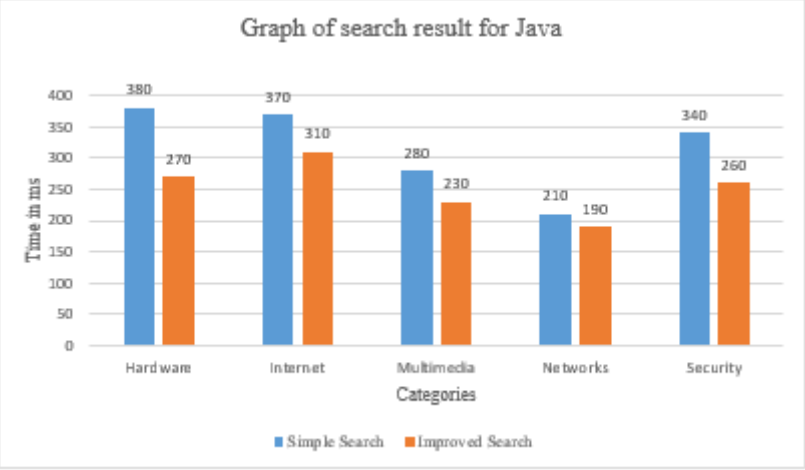

Figure 3 Time consumption of the proposed and tradition algorithms for search query java

- **Search Result For Package**

Table 5.2 gives time consumption for query package, which have categories wise search result related to package query. There are five categories are found related to java keywords which is hardware, book, Companies, Auto Racing and Stores. This table gives the various comparisons performed between simple searches and using

ISSN: 2277-5528
Impact Factor: 5.085

modified Bayes search. The result shows the improved search is better than simple search. Improved search uses minimum time in all aspects.

Table 2 Time consumption for query package

Categories	Simple Search	Improved Search
Hardware	300	270
Book	310	230
Companies	320	240
Auto Racing	420	210
Stores	230	200

Figure 4 shows time consumption of the proposed and tradition algorithms for search query package. The graph shows various attributes related to the maximum number of occurrence search result related to package query.

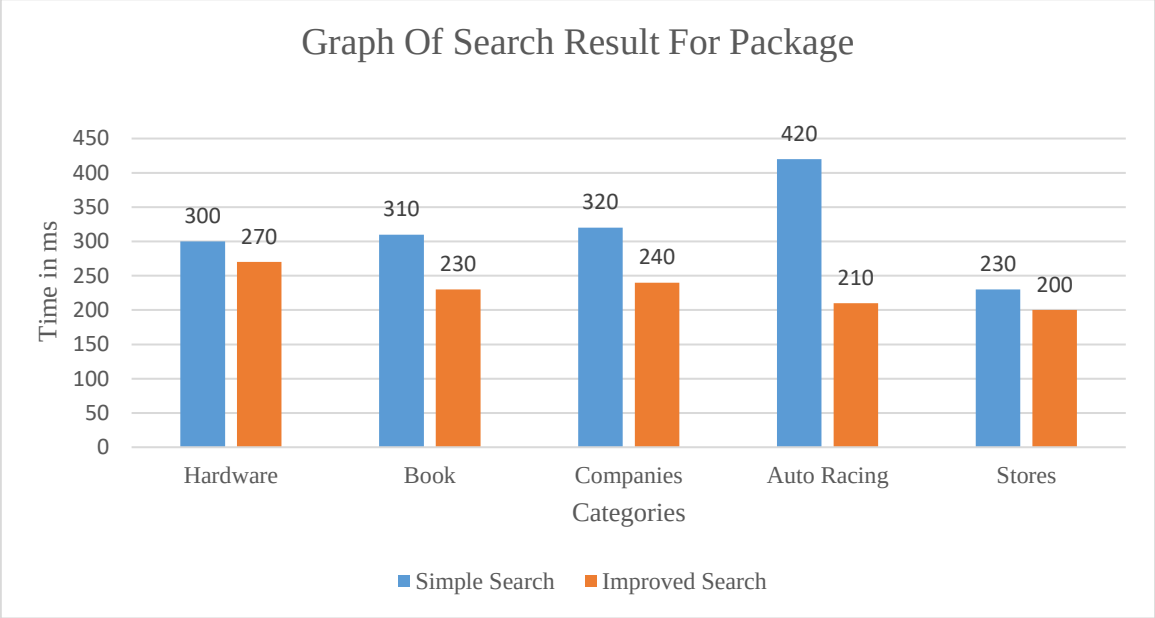

Figure 5Time consumption of the proposed and tradition algorithms for search query package.

- **Search Result For Games**

The table 5.3 shows the categories wise search result related to Games query. There are five categories are found related to Games keywords which is Games, Software, Companies, Family Kids and Stores. This table gives the various comparisons performed between simple search and using modified Bayes search. The result shows the improved search is better than simple search. Improved search uses minimum time in all aspects.

Table 5.3 Time consumption for query Games

ISSN: 2277-5528
Impact Factor: 5.085

Categories	Simple Search	Improved Search
Games	380	350
Software	300	270
Companies	290	250
Family Kids	330	310
Stores	240	230

Figure 6 shows the time consumption of the proposed and tradition algorithms for search query games. The graph shows various attributes related to the maximum number of occurrence search result related to Games query.

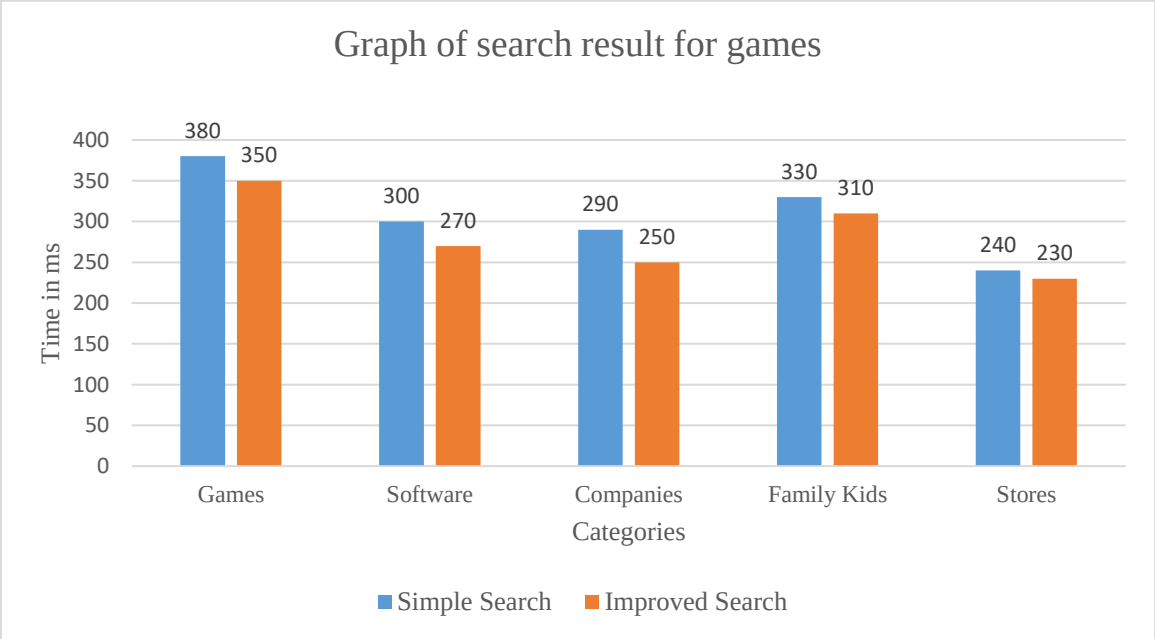

Figure 6 Time consumption of the proposed and tradition algorithms for search query games.

We calculate the result related to its domain. When user gives the query, proposed implementation search by comparison the term and phrased used in this document. Related domain find after followed some procedure.

CONCLUSION

The proposed work explores the data mining approach, which uses to classify the google search results according to their higher probabilities of specific data; therefore, the proposed study focuses on analysing the text mining and the classification algorithms. This chapter provides the summary of entire work performed for domain classification of google search. Additionally promising future extensions are also included in this work.

Domain name classification is a great challenge in a web mining. Thousands of domains are there, find the right one from them is quite difficult process. Searching process takes too much time and many filtrations to search out

the right one from them. Naive Bayes classifier comes to solve the problem of domain classification. Proposed work uses term and phrased based approach simultaneously to get the accurate result from the training domain. A new query result returns document, when that document enters, Naive Bayes uses to find the probability of highest term and phrased available there using matrix. A modified Naive Bayes algorithm exists to deal with that filter result. Modified Naive Bayes works on selected resulted whose frequency of occurrence is high, so the modified Naive Bayes performance gets higher in terms of timing. Experimental results show the presented work outperforms far better than the previous one.

REFERENCES

I. Samuel Ieong, Nina Mishra, Eldar Sadikov, Li Zhang, "Domain Bias in Web Search" , ACM New York, NY, USA ©2012.
II. Bo Geng, Linjun Yang, Chao Xu, Xian-Sheng Hua, "Ranking Model Adaptation for Domain-Specific Search", IEEE Transactions on Systemsknowledge and data engineering vol. 24, no. 4, April 2012.
III. Junghoon Chae, Dennis Thom, Yun Jang, Sung Ye Kim, Thomas Ertl, David S. Ebert, "Public behaviour response analysis in disaster events utilizing visual analytics of microblog data",Elsevier Ltd. All rights reserved 2013.
IV. Umajancy. S, Dr. Antony Selvadoss Thanamani, "An Analysis on Text Mining –Text Retrieval and Text Extraction", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 8, August 2013.
V. Xia Hu, Lei Tang, Jiliang Tang, Huan Liu, "Exploiting Social Relations for Sentiment Analysis in Microblogging", WSDM '13, February 4–8, 2013, Rome, Italy, ACM 978-1-4503-1869-3/02/ 2013.
VI. Rahul A. Patil, Prashant G. Ahire, Pramod. D. Patil, Avinash L. Golande, "A Modified Approach to Construct Decision Tree in Data Mining Classification", International Journal of Engineering and Innovative Technology (IJEIT), Volume 2, Issue 1, July 2012
VII. R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," Proc. ACM SIGMOD, 1993.
VIII. Vishal Gupta, Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, August 2009
IX. G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," Information Processing and Management:An Int'l J., vol. 24, no. 5, pp. 513-523, 1988.
X. H. Ahonen, O. Heinonen, M. Klemettinen, and A.I.
XI. Verkamo, "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document
XII. Collections," Proc. IEEE Int'l Forum on Research and Technology Advances in Digital Libraries (ADL '98),