Data Mining based Technique for Natural Event Prediction and Disaster Management

Suraj Singh Chouhan
Vikrant Institute of Technology & Management, Indore
RGPV University, Bhopal, Madhya Pradesh, India

Ravi Khatri
Vikrant Institute of Technology & Management, Indore
RGPV University, Bhopal, Madhya Pradesh, India

ABSTRACT

Due to advancement of technology the computational algorithms and computer based data analysis are used for making large and effective decision. Therefore a significant role on the human life is observed the main aim of developing such kind of technology is to provide ease in various domains and making future planning for managing the risk. Among various issues the disaster is also a critical risk in today's scenario of India. Thus the risk management and disaster management techniques are developed for keep in track the losses in controlled manner. In this presented work a new model using the data mining techniques for predicting the disaster and their place is proposed for development. Therefore various different data mining techniques and methods are included for developing the accurate and effective data model. The proposed work includes the three main contributions for prediction based technique development. First the pre-processing technique development by which the unstructured data is processed and filtered for transform the information into the structured data format. Therefore in this phase the Bay's classifier is used, Secondly development of learning technique for accurate pattern learning of the disasters and their places. Therefore in this phase the k-means clustering and hidden Markov model is employed for performing the training. Finally the prediction and their performance evaluation, in this phase the trained model is used to accept the current scenarios and predict the next event.The implementation of the proposed technique is performed using the JAVA technology and for dataset generation the Google search API is used. After the implementation of the proposed system the performance of the system in terms of accuracy, error rate, time complexity and space complexity is evaluated. The experimental results demonstrate the effective and accurate learning of the system. Thus the proposed data model is adoptive and acceptable for the various real world data analysis and decision making task.

Keywords

Data mining, classification, supervised learning, implementation, performance study.

1. INTRODUCTION

Machine learning and data mining enables us to analyse the huge amount of historical data and extract the patterns which can be used for different applications. The evaluation of the historical data is termed in machine learning as training of the computational model. Based on the experience collected form the historical records the future trends and upcoming events are predicted or approximated. Therefore the data mining techniques supports the classification and prediction based on supervised learning concept for analysis of previous data. There are a number of different kinds of applications of supervised learning is available by which the prediction, classification, pattern recognition and problem optimization becomes feasible.

In this presented work the supervised data mining technique is used to evaluate the historical data and the natural events and predict the upcoming events. This task is performed with the help of data mining algorithms by analysing the news contents. Therefore different processes of the data mining technique are combined together for developing the enhanced scheme for the disaster management and prediction. In order to prepare the proposed data model need to perform the collection of data, pre-processing of data, data model development and finally utilizing the data model for predicting the upcoming events.

A disaster is a serious disturbance of the operational public or a social activities widespread human, material, economic or environmental losses and impacts. In current scenarios, disasters are seen as the significance of unfortunately managed risk. These risks are the product of a combination of both hazards and vulnerability. In order to deal with the issues and circumstances of the natural disasters the palling and management is primary requirement. Therefore the disaster management techniques are developed to reduce the impact of disasters in the social and economic life cycles.

Disaster is a kind of emergency where a large amount of human life and revenue is lost therefore in order to manage or recover the loss from such event a management scheme is required to handle the conditions. According to the effect and management steps the entire management process can be described in four phase process [27]:

1. **Mitigation:** mitigation is a kind of awareness for individuals and families, they are train to avoid risks thus it includes assessment of possible risks for health and personal property, and also train to take steps to reduce the outcomes of a tragedy, or to protect against effects of a disaster.
2. **Preparedness:** Preparedness concerned about preparing techniques and procedures to use when a disaster occurs. These techniques are used to moderate the effect of disaster.
3. **Response:** in this phase the process and team work are involved to providing the relief for the affected individuals and families from the disasters. This process is taken place as the disaster is occurred.
4. **Recovery:** after passing out the disaster and after making the response the social and personal effect of disaster is measured and for improving the social conditions and improving life of effected peoples the additional steps of relief is taken place is termed as the recovery.

In these phases of the disaster management a rich amount of data generated and preparation of this data needs appropriate knowledge management techniques. Thus the proposed work is focused on find the appropriate technique for knowledge management during the disaster management phases. And also to be developing an effective data model by which the approximations of these events are becomes predictable. In this section the basic overview of the proposed work is provided and in the next sections the entire details and proposed data model is provided.

2. PROPOSED WORK

If the upcoming events are predictable then a significant amounts of losses are avoidable by the proper management and preparations. With this aim the proposed work is modelled to find a technique by which the Google data is analysed for finding the news and the relevant events for preparing the predictive data model. Basically the data mining techniques are offered the analysis of the patterns of data and utilize them to develop classification, prediction and pattern recognition data models. In this work the classification technique is investigated which is a kind of supervised learning. Additionally for experimentation and predictive system design the unstructured data sources are utilized. The obtained data is extracted from the real world search engine i.e. Google.com. Therefore to process the unstructured data for extraction of knowledge proposed technique is developed in three major modules. Individual modules are working as subcomponents to develop and design required data model.

2.1 Pre-processing

The pre-processing is a technique for filtering the unwanted data from the input datasets. This process may include the cleaning of data, transformation of data, dimensional transformation of data and data quality assessment and enhancement. Therefore the following process is prepared for the extracting the web data pre-process them utilized for further decision making task. The figure 1 shows the proposed pre-processing technique of the system. The components of the proposed pre-processing model are described as:

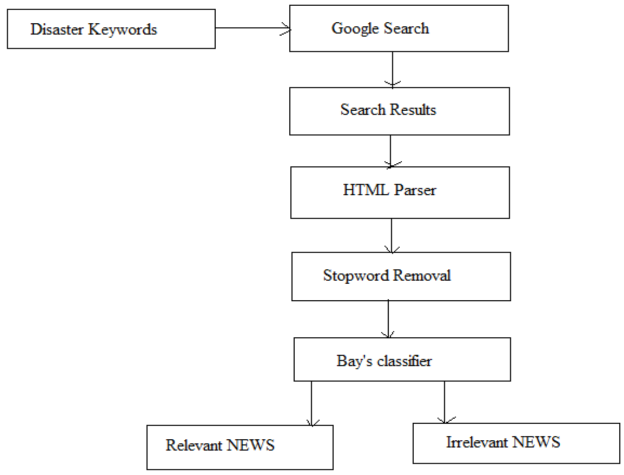

Figure 1. Pre-processing of data

Disaster Keywords: in order to initiate the working of the system some input values are required. Therefore a list of disaster keywords is prepared which are used to find the NEWS relevant to the input disaster keywords. These listed keywords are used as input to the proposed system.

Google search: the Google search API is implemented in this module, therefore the input user query (i.e. disaster keywords) is applied on the Google search function. The Google search the keyword from the web and reported the results.

Search results: the obtained results from the Google search API is collected using the HTML format thus the entire data is found in unstructured and noisy format for utilization with the predictive model.

HTML Parser: the Google search results which are obtained in HTML format is processed here thus in this phase the HTML tags and the keywords are removed through the HTML parser and the extracted content data is used in further processes.

Stop word removal: the extracted HTML contents are processed again for finding reducing the amount of data. Therefore some stop words (i.e. this, that, is, am, are) is removed from the content extracted from the HTML.

Bay's classifier: the reduced contents are evaluated again with respect to the disaster keyword produced as input to the system. Therefore the bay's algorithm is working here as a filter that separate relevant information according to the user keyword.

Relevant NEWS: this is a part of the search results that are relevant to the current input query. Therefore in this part of information can be used for further data model training. Thus the tokenization is performed and three key information is extracted from the relevant NEWS data. That data is preserved for further use thus that is arranged in the temporary data storage. Table 1 demonstrate the organization of filtered data.

Table 1. Filtered Data

Date	Place	Event

Irrelevant NEWS: a part of information extracted from the Google search results is not relevant to the current input query is left that is not used further data modelling.

2.2 Training

After the pre-processing of unstructured data the useful information is used in this phase for performing the training. The figure 2 shows the training model of the proposed predictive data model. The different participating components of the training model are defined as:

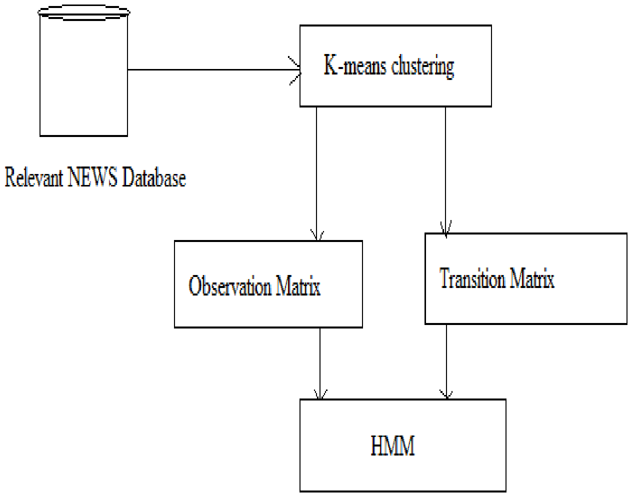

Figure 2. Training Model

Relevant NEWS database: the previously recognized valuable data is used in this phase thus the prepared table 1 is used.

K-means clustering: the table data is grouped in the similar amount of clusters as the amount of keywords are used for performing the search operations. Therefore the search keywords are works in further process as the states of the Hidden Markov Model. On the other hand another cluster is prepared using the different places that are used for prediction for the place. Thus the place wise data grouping is considered here as the observations of the Hidden Markov Model.

Transition Matrix: using the previous phase of data clustering the states of the Markov model is recovered in terms of the keywords used for search thus the matrix is prepared using the states to states and their occurrence probability. To understand the concept the table 2 shows the transition matrix of the system.

Table 2. Example Transition Matrix

	Keyword1	**Keyword2**
Keyword1		
Keyword2		

Observation matrix: in the similar ways the observational matrix is also prepared with the help of keywords searched as states and the places for prediction as the observation. Thus the table 3 demonstrate the example of the observation matrix which is used with the Hidden Markov model.

Table 3. Observation Matrix

	Place1	**Place2**
Keyword1		
Keyword2		

Hidden Markov Model: the two different outcomes of the k-means clustering is used to develop the observation and transition matrix of the hidden Markov model. Thus in this phase the hidden Markov model is used with the input matrix for computing the trained data model for prediction.

2.3 Prediction

The outcome of the previous phase trained data model is used in this phase for finding the prediction of the different places and the upcoming events. Therefore the predictive model is demonstrated using the figure 3.3. in this diagram in first condition the required parameters such as place and the keyword is selected which is previously occurred in that place and the next event is predicted according to their historical patterns. Thus in this phase the system return the two different outcomes performance and predicted event for the specified place.

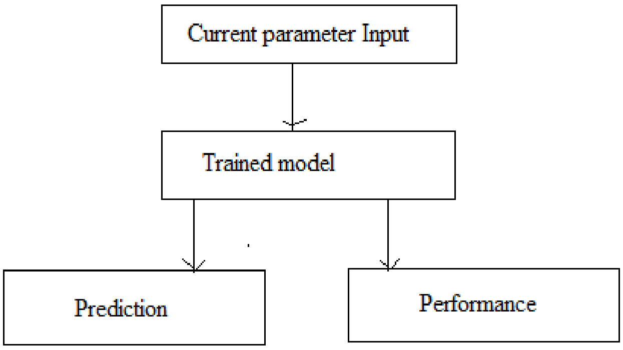

Figure 3. Prediction Model

3. RESULTS ANALYSIS

The given section provides the understanding about the evaluated results and parameters. These parameters are shows the effectiveness of the proposed technique.

3.1 Accuracy

In this presented work the key aspect is to develop a predictive system. The predictive system accepts the current scenarios inputs to the system and returns the approximated upcoming events. Thus for demonstration of such a data model the predictive capability of the system is measured in terms of accuracy. The accuracy of the system provides the accurately recognized events form the system. The accuracy of a predictive system can be defined using the following formula:

$$accuracy = \frac{correctly\ recognized\ patterns}{total\ patterns\ to\ identify} X100$$

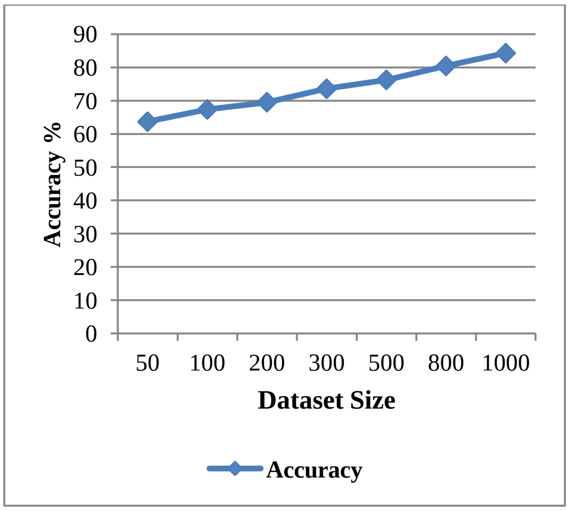

Figure 4 accuracy

The performance of the proposed technique is demonstrated using figure 4. In this diagram the X axis contains the different amount of patterns available for preparing the data model and the Y axis contains the performance in terms of accuracy percentage. According to the given results the performance of the system is increases as the amount of data for training the model is increases. Thus the accuracy of system is dependent of the amount of knowledge stored on the data for evaluation or learning.

3.2 Error Rate

The error rate of the proposed predictive data model is demonstrated in this section. The error rate of the predictive system is the incorrectly recognized patterns from the input samples that are used for prediction. The error rate of the proposed model is given using the figure 5. In this diagram the X axis contains amount of data provided for evaluation and the Y axis shows the error rate of the system in terms of percentage. The estimated error rate shows the model is adoptive and as that is learns with huge data that is simulating the accurate evaluation of the data. Thus model is able to perform good prediction with the increasing amount of learning data.

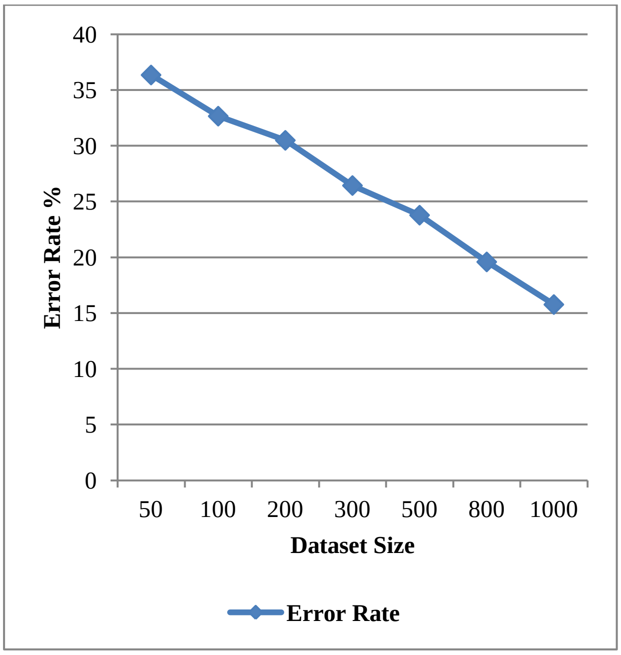

Figure 5. Error Rate

3.3 Memory consumption

Memory consumption or the space complexity of the system shows the amount of main memory required to process the data using the prepared data model.

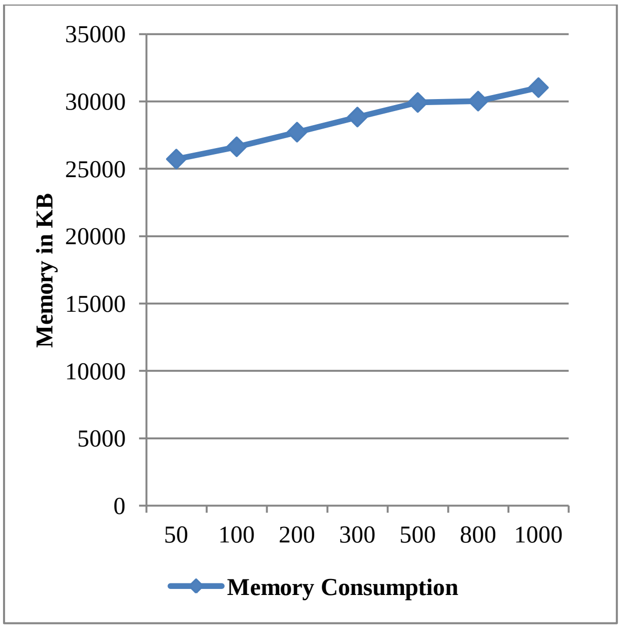

Figure 6. Memory Consumption

The performance of the proposed system in terms of memory utilization is demonstrated using the figure 6. In this diagram the X axis shows the amount of data used for training and the Y axis shows the corresponding memory usages of the system. According to the obtained results the performance of the system is depends on the amount of data produces for the training, thus as the amount of training data is increases the amount of main memory consumption is also increases in the similar ratio.

3.4 Training time

This parameter is also termed as the training time complexity, the training time consumption of the proposed predictive system is given using figure 7. In this diagram the X axis shows the amount of data provided for training and the Y axis shows the amount of time consumed for performing the training using the proposed concept. In this diagram the amount of data (x –axis) is increases the amount of time consumed is also increases in the similar manner. The computed training time of the system is demonstrated here in terms of milliseconds. Finally the results show the performance of the system is depends on the amount of data produced for training.

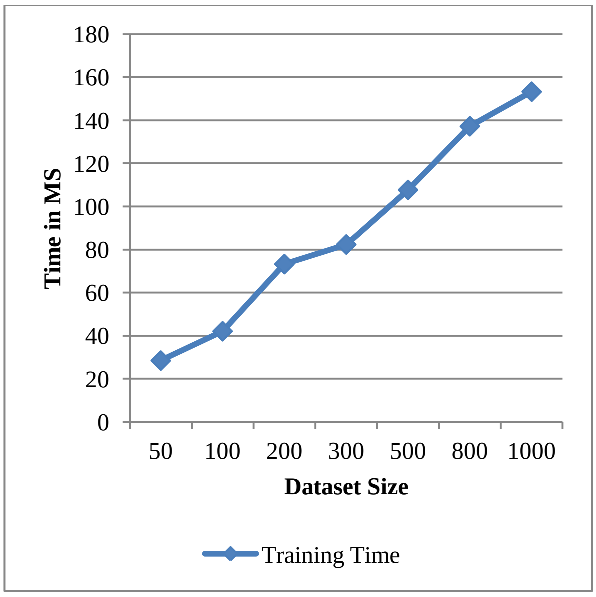

Figure 7. Training time

3.5 Prediction time

After the successfully training of the proposed system the system is used for predicting the upcoming events. This event prediction needs a small amount of time to process the data and provides the conclusion that is termed as the prediction time of the system. The prediction time of the model is reported using the figure 8 basically the prediction time is the complete time which is used for processing and predicting the class label from the data. The given figure demonstrates the performance of the system in terms of prediction time. Therefore in the X axis shows the different sets of experiments performed with the system and the Y axis shows the amount of time required to predict the class label from the input data. The given time is measured in terms of milliseconds. During the evaluation of results that is found that the amount of prediction time is depends on the amount of data to be process but that is prediction time is always less than the training time.

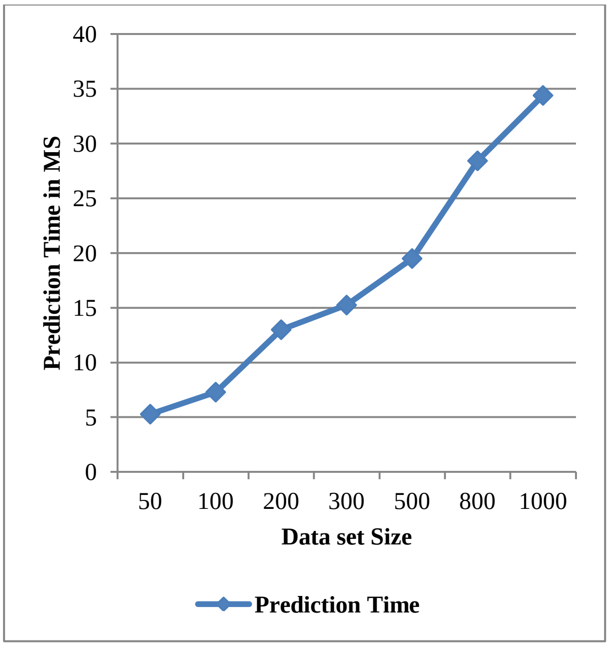

Figure 8. Prediction Time

4. CONCLUSIONS

The proposed work is devoted to find an accurate technique which analyse the historical data using the data mining techniques and used for predicting the disaster events for the predefined or sensitive places. The given section provides the summary of the entire research work performed additionally the future extension of the proposed system is also provided.

4.1 CONCLUSION

Data mining is a technique by which the raw data is processed using the computational algorithms and the valuable patterns are extracted for utilizing them into various real world application developments and decision modeling. Basically the raw data can be available in both the formats structured and unstructured. The structured data is a well-defined organization of data in a structure of table or others. In the similar ways the unstructured data is available in the format of the text or others. The unstructured data is most of the time noisy in nature and a strong pre-processing technique is required to handle them.

Therefore the key issue of the proposed predictive model design is to develop a pre-processing technique that extracts the accurate valuable data from the unstructured data source. Thus first of all the pre-processing technique is developed with the help of the Bay's classification algorithm and data is classified to keyword relevant and irrelevant data. In the next step the pre-processed knowledge is used for performing the training of the system. Thus in next step k-means clustering is applied on the data for finding the observations from the data and the transitions on the data among two consecutive events. Thus the keywords are considered as the states and the places are used as the observations. Using the states and places the transition matrix and observational matrix is prepared and the HMM (Hidden Markov Model) is used for predicting the event and the place. For prediction of the event the HMM need to accept the place and the current state of the natural event.

After the formulation of the proposed system the model is implemented using JAVA technology. Additionally the performance of the system is measured in terms of different performance factors. The summary of the evaluated performance parameters are described in the table 4.

Table 4. Performance Summary

No.	Parameters	Remark
1	Accuracy	The high accurate results are obtained additionally the accuracy of the predictive data model is increases as the amount of data for training is increases
2	Error rate	The system demonstrate the low error rate as the amount of data for training is increases
3	Memory consumption	The amount of memory consumption is moderate but increases slightly as the amount of data for training is increase
4	Training time	A significant amount of time required because the data is fetched at the real time process the entire data and then used with the system for training purpose
5	Prediction time	The prediction time of the proposed algorithm is very low when the model is previously trained with the patterns.

According to the obtained results using the different experimental scenarios and different keywords and place information the testing is performed. During the different experiments the technique is found effective and accurate for the event prediction techniques.

4.2 Future Work

The proposed technique is effective and produces the accurate results according to the data and the information available on the knowledge base. But the system has some improvements for future extension and development on the proposed model. Some essential of them are reported as:

1. The system consumes a significant amount of time during the pre-processing and training of the data model therefore needs to optimize for the training time of the system.

The performance of the system is dependent on the quality of data and the amount of data available for the training purpose. Therefore need to increases the amount of training data before use for accurate predictions..

5. REFERENCES

[1] Li Zheng, Chao Shen, Liang Tang, ChunqiuZeng, Tao Li, Steve Luis, and Shu-Ching Chen, "Data Mining Meets the Needs of Disaster Information Management", IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, VOL. 43, NO. 5, SEPTEMBER 2013

[2] Data mining Concepts and Techniques, Second Edition, Jiawei Han and MichelineKamber, http://akademik.maltepe.edu.tr/~kadirerdem/772s_Data.Mining.Concepts. and .Techniques.2nd.Ed.pdf

[3] Jerzy W. Grzymala-Busse and Ming Hu, "A Comparison of Several Approaches to Missing Attribute Values in

Data Mining", Springer-Verlag Berlin Heidelberg 2001, pp. 378−385

[4] Ritika, "Research on Data Mining Classification", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 4, April 2014

[5] Abbas Jafari, S. S. Patil, "Use Of Data Mining Technique to Design a Driver Assistance System", Proceedings of 7th IRF International Conference, 27th April-2014, Pune, India, ISBN: 978-93-84209-09-4

[6] FabricioVoznika, LeonarDoviana, "Data Mining Classification", http://courses.cs.washington.edu/courses/csep521/07wi/prj/leonardo_fabricio.pdf

[7] A.K. Jain, M.N. Murthy, P. J. Flynn, "Data Clustering: A Review", © 2000 ACM 0360-0300/99/0900–0001

[8] A Comparative Study of Data Clustering Techniques, KhaledHammouda, Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

[9] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, Arthur Zimek, "Density-based Clustering", WIREs Data Mining and Knowledge Discovery 1 (3): 231–240. doi:10.1002/widm.30

[10] 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

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

[12] Andreas Hotho, Andreas Nurnberger, Gerhard Paaß, FraunhoferAiS, "A Brief Survey of Text Mining", Knowledge Discovery Group Sankt Augustin, May 13, 2005s

[13] Hien Nguyen, Eugene Santos, and Jacob Russell, "Evaluation of the Impact of User-Cognitive Styles on the Assessment of Text Summarization", IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, Vol. 41, No. 6, November 2011

[14] Umajancy. S, Dr. Antony SelvadossThanamani, "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

[15] MilošRadovanović, MirjanaIvanović, "Text Mining: Approaches And Applications", Abstract Methods and Applications in Computer Science (no. 144017A), Novi Sad, Serbia, Vol. 38, No. 3, 2008, 227-234

[16] 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

[17] Li Zheng, Chao Shen, Liang Tang, ChunqiuZeng, Tao Li, Steve Luis, and Shu-Ching Chen, "Data Mining Meets the Needs of Disaster Information Management", IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, VOL. 43, NO. 5, SEPTEMBER 2013

[18] B. Merz, H. Kreibich, and U. Lall, "Multi-variate flood damage assessment: a tree-based data-mining approach", Nat. Hazards Earth Syst. Sci., 13, 53–64, 2013

[19] E. Schnebele, G. Cervone, and N. Waters, "Road assessment after flood events using non-authoritative data", Nat. Hazards Earth Syst. Sci., 14, 1007–1015, 2014

[20] Marco Avvenuti, Stefano Cresci, Andrea Marchetti, Carlo Meletti, Maurizio Tesconi, "EARS (Earthquake Alert and Report System): a Real Time Decision Support System for Earthquake Crisis Management", KDD'14, August 24–27, 2014, New York, NY, USA. Copyright 2014 ACM 978-1-4503-2956-9/14/08

[21] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, Patrick Meier, "Extracting Information Nuggets from Disaster-Related Messages in Social Media", Proceedings of the 10th International ISCRAM Conference – Baden-Baden, Germany, May 2013

[22] Xing Wanli, GuoRui, Petakovic Eva, Goggins Sean, "Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory", 2014 Elsevier Ltd. All rights reserved.

[23] D.A. Adeniyi, Z. Wei, Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method", Applied Computing and Informatics (2016) 12, 90–108

[24] RuibinGeng, Indranil Bose, Xi Chen, "Prediction of financial distress: An empirical study of listed Chinese companies using data mining", European Journal of Operational Research, 2014 Elsevier B.V. All rights reserved.

[25] R. Taghizadeh-Mehrjardi, K. Nabiollahi, B. Minasny, J. Triantafilis, "Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran", © 2015 Elsevier B.V. All rights reserved

[26] Sasan Barak, Mohammad Modarres, "Developing an approach to evaluate stocks by forecasting effective features with data mining methods", Expert Systems with Applications 2014 Elsevier Ltd All rights reserved.

[27] M. E. Baird, "The "Phases" of Emergency Management", Vanderbilt Center for Transportation Research (VECTOR), January 2010

[28] RoshaniChoudhary, JagdishRaikwal, "An Ensemble Approach to Enhance Performance of Webpage Classification", International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5614-5619

[29] Juntao Wang, Xiaolong Su, "An improved K-Means clustering algorithm", 978-1-61284-486-2/111$26.00 ©2011 IEEE

[30] ShwetaJaiswal, Atish Mishra, Praveen Bhanodia," Grid Host Load Prediction Using GridSim Simulation and Hidden Markov Model", International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 7, July 2014)