A survey on Predictive data mining techniques for disaster prediction

Arjun Singh

Computer Science department,

Rajiv Gandhi Proudyogiki Vishwavidyalaya

Address-290/2, Shakti nagar near bhagat singh chowk
Bhopal (M.P.),INDIA

Abstract—*the world is unpredictable and random in nature.Some of events are human generated and some of them are nature inspired. Among these events the natural events are change the face of leaving and also impact on the human life. In such events the disasters are most of the time affecting human and their daily life. In this paper a survey on the disasters and their effects are prepared first. Then after a technique is introduced to perform the text and news analysis by which the location of disaster can be predictable. The proposed data model is used to analyse the text and HTML documents for making the prediction. That technique is not used to predict the disaster before it occurred, it helps to discover the locations of disasters for improving recovery operations.*

Keywords— *Disaster management, data mining, prediction, text and news analysis, recovery*

I. Introduction

The India is a developing country and huge amount of changes are occurred in a small units of time. Digitization and technology development is rapidly growing the country. A number of efforts are placed in the direction of development of new technology and information processing but too few working performing to maintain these technologies and their generated data during the natural unwanted events. Some disasters are human generated such as terrorist attack and accidents but these events not much affecting the digital infrastructures. On the other hand the natural events such as flood, earth quacks and others are not only affects the human life it also harms the digital infrastructure of country.

In order to improve the digital infrastructure and their ability to prevent the recovery operations a new kind of system is required. Therefore the proposed survey is conducted for developing the effective and recoverable computing. The following objectives are placed:

1. **Study of disaster management:**in this phase the different disasters and the involved management techniques are investigated.
2. **Study of current digital infrastructure:**in this phase the different digital infrastructures those are involved in our daily life is investigated.
3. **Design a data analysis module for improving the recovery operations:** in this phase a recovery management system module for the existing infrastructure is proposed and implemented.
4. **Performance analysis of implemented module:** in order to justify the proposed solution a performance analysis of the system is conducted on different parameters.

This section provides the basics of the proposed investigation the next section provides a brief detail on the study of disaster and their management phases.

II. Disaster Management

Human life is much sensitive and adopts the changes frequently, some changes in human life are occurred due to circumstances made by human nature and some of them are occurred due to natural events. Natural events such as flood, earthquake and others are some unwanted natural events that changes human life. In this presented work the natural events and their effect in human life is investigated.

Disasters are not defined by static events, or irreversible relationships, but by social concepts and these are liable to change [1]. Most of the impacts of the disasters are much complicated and also much painful. There are various methods by which a disaster event can define some of them are human dependent and some of them are natural disasters.

Due to their unpredictable nature of occurrence the disasters are attracting researches among them computer sc. and technology places more effective contributions. In computer science the various applications are developed for predicting weather or predicting the rainfall but prediction is an approximation process of data evaluation. Therefore during disaster computational domain helps by organizing the disaster data, tracking, monitoring and information retrieval.

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 :

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.

III. CURRENT DIGITAL INFRASTRUCTURE

This section includes the study about the current digital infrastructure and the different applications that providing ease in our daily life.

There are a number of applications and the digital infrastructures that are providing ease in various domains, such as banking, online shopping, telecommunication and others. In these applications all of them are having it’s own importance and techniques of infrastructure management and computational abilities. But among them the banking is a critical and sensitive application and infrastructure. Therefore in this presented work the banking applications are their management technique is investigated to improve during the disasters.

The given figure 1 shows the current computing technique of banking servers. In this system a centralized server is used which accept each entries of the customer and their transections. If the primary system is affected than the entire banking and management is affected.

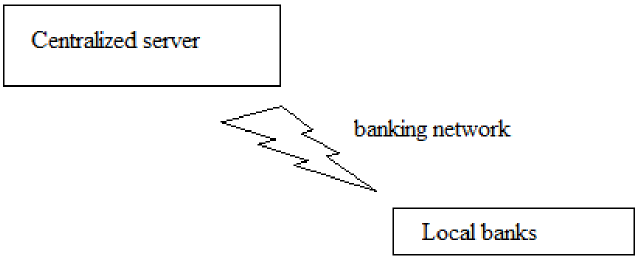

Figure 1 current banking

In order to reduce the effect of disasters the computational manner can be designed in distributed manner as given in figure 2.

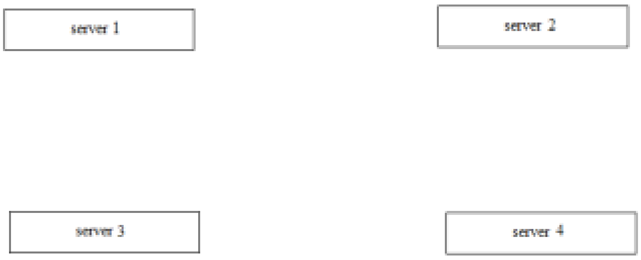

Figure 2 distributed server

In this computational system the centralized server is distributed in four different zones. And each zone servers are equipped with similar processing capabilities and storage units. The key advantage of these servers is to having the entire records in all the places. Similarly in one server is failed then other three servers can serve. Therefore this computing technique is effective and promising.

In addition of that for incorporating the disaster management and data recovery during the unwanted events a new process model for finding the affected area automatically a new framework is introduced in this paper. The desired framework is discussed in the next section.

IV. PROPOSED WORK

Nature helps us and provides us various resources for our daily needs and our leaving. But sometimes it becomes crucial and damages most of the surrounding. Therefore the nature is unpredictable but due to new technology human can predict it and can be use it to stop losses of human being, data and money. On the basis of this concept an accurate technique is required to develop which analyse the predictive NEWS contents and provide the accurate location information where the problems can be occurred. The proposed system is desired to develop for a real time information system which provides the support to manage and preserve the sensitive and essential data during large accidents and disasters.

The proposed solution leads to solve the problem of data management during unwanted natural events. Therefore the following proposal is provided.

1. **Prepare the local and global data storage management:** in this phase the local branches of banks and the zone wise data is stored for banking transactions.
2. **Predictive news analysis:** in this phase the News content analysis is performed for finding the predictive natural events in all the places where the bank branches are established.
3. **Improving the relevancy of data analysis and content extraction:** during content analysis that is required to improve the relevancy and accuracy of data search.

Thus the proposed work intended to design the following system model as given in figure 3.

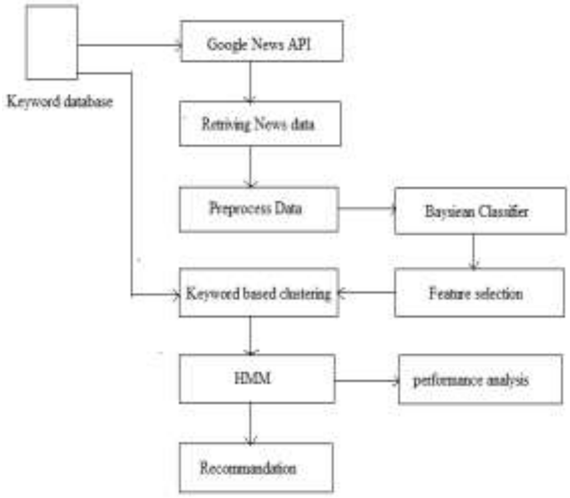

Figure 3 proposed recommandation system

The proposed news analysis technique includes the search terms in a key word database such as disaster, earth quack and others which are taken as input to the Google search API. The API returns a number of search results which is produced as HTML. The retrieved news data is preprocessed using the pre-processing techniques. HTML contents are parser and then tokenization, stop word removal and other operationsare performed to filter the data. In further a Bayesian classifier is activated which classify the contents and helps to select the appropriate features from the available pre-processed data. Using the keyword database the directed K-mean algorithm is used to prepare the group of similar news data. These groups are based on locations and keywords basis. Using these groups the hidden Markov model processes the data which works as recommendation engine to predict the place where the disaster or other event can occurred.

This section provides a brief description of the proposed data backup and recovery management system. the next section provides the detailed understanding about the involved algorithms.

V. REQUIRED ALGORITHMS

This section includes the different algorithms that are used to develop the proposed recommendation model for disaster recovery management.

A. K-Means clustering

The K-Means clustering algorithm is a partition-based cluster analysis method [3]. According to the algorithm we initially select k objects as initial cluster centers, then compute the distance between each object and each cluster center and allocate it to the nearest cluster, renew the averages of all clusters, replicate this process until the criterion function converged. Square error criterion for clustering

$$E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left\|x_{ij} - m_i\right\|^{2}$$

x_{ij} is the sample j of i-class, m_i is the center of i-class, n_ii is the number of samples of i-class. K-means clustering algorithm is simply described as

Input: N objects to be cluster $x_1, x_2 \dots, x_n$, number of clusters k;

Output: k clusters and the sum of dissimilarity between each object and its nearby cluster center is the small;

1. Arbitrarily select k objects as initial cluster centers$(m_1, m_2, \dots, m_k)$;

2. compute the distance between each object Xi and each cluster center, then allocate each object to the nearby cluster, formula for calculating distance as:

$$d(x_i, m_i) = \sqrt{\sum_{j=1}^{d}(x_i - m_{j1})^2}\,, i = 1 \dots N, j = 1 \dots k$$

$d(x_i, m_i)$is the distance between data i and cluster j.

3. Compute the mean of objects in each cluster as the fresh cluster centers,

$$m_i = \frac{1}{N}\sum_{j-1}^{n_i} x_{ij}\,, i = 1,2, \dots, K$$

$N_i$is the number of samples of current cluster i;

4. Repeat 2) 3) until the principle function E converged, return $(m_1, m_2, \dots, m_k)$ Algorithm terminates.

B. Hidden Markov model

An HMM is a double implanted stochastic process with two hierarchy levels. It can be used to model much more complex stochastic processes as compared to a traditional Markov model. In a specific state, an observation can be generated according to an associated probability distribution. It is only the observation and not the state that is visible to an external observer. An HMM can be characterized by the following [4]:

1. N is the number of states in the model. We denote the set of states' S = {S_1; S_2;..., S_N}, where S_i, i= 1;2;...;N is an individual state. The state at time instant t is denoted by qt.

2. M is the number of distinct observation symbols per state. We denote the set of symbols

V = {V_1; V_2; ...V_M}, where V_i, I = 1; 2; ...; M is an individual symbol.

3. The state transition probability matrix A = [a_{ij}], where

$$a_{ij} = P\left(q_{t+1} = S_j \middle| q_t = S_i\right), 1 \le i \le N, 1 \le j \le N; t = 1,2 \ldots$$

Herea_{ij}> 0 for all i, j. Also,

$$\sum_{j=1}^{N} a_{ij} = 1 , 1 \le i \le N$$

4. The remark symbol probability matrix B = {$b_j(k)$}, where

$$b_j(k) = P\left(V_k \middle| S_j\right), 1 \le j \le N, 1 \le k \le M \; and$$

$$\sum_{k=1}^{M} b_j(k) = 1 , 1 \le j \le N$$

5. The initial state probability vector r= πi , where

$$\pi_i = P(q_1 = S_i), 1 \le i \le N$$

Such that

$$\sum_{i=1}^{N} \pi_i = 1$$

6. The remark sequence O = O_1; O_2; O_3; ...O_R, where each remark O_t is one of the symbols from V, and R is the number of remarks in the sequence.

It is manifest that a complete specification of an HMM needs the approximation of two model parameters, N and M, and three possibility distributions A, B, andπ. We use the notation ʎ = (A; B; π) to specify the complete set of parameters of the model, where A, B implicitly contain N and M.

An observation sequence O, as mentioned above, can be generated by many possible state sequences. Consider one such particular sequence Q = q_1; q_2; ...; q_R; where q1 is the initial state. The probability that O is generated from this state sequence is given by

$$P(O|Q,\lambda) = \prod_{t=1}^{R} P(O_t|q_t,\lambda)$$

Where statistical independence of observations is assumed Above Equation can be expanded as

$$P(O|Q,\lambda) = b_{q1}(O_1) b_{q2}(O_2) \ldots \ldots b_{qR}(O_R)$$

The probability of the state sequence Q is given as

$$P(Q|\lambda) = \pi_{q1} a_{q1q2} a_{q2q3} \ldots \ldots a_{qR-1qR}$$

Thus, the probability of generation of the observation sequence O by the HMM specified by can be written as follows:

$$P(O|\lambda) = \sum_{all\ Q} P(O|Q,\lambda)\, P(Q|\lambda)$$

Deriving the value of $P(O|\lambda)$ using the direct definition of is computationally intensive. Hence, a procedure named as Forward-Backward procedure is used to compute$P(O|\lambda)$.

C. Bayesian classifier

The Naive Bayes classification algorithmic rule is a probabilistic classifier. It is based on probability models that incorporate robust independence assumptions. The independence assumptions usually don't have an effect on reality. So they're thought of as naive. You can derive probability models by using Bayes' theorem. Based on the nature of the probability model, you'll train the Naive Bayes algorithm is a supervised learning technique. In general a naive Bayes classifier assumes that the value of a specific feature is unrelated to the presence or absence of the other feature, given the category variable. There are two types of probability as follows [5]:

1. Posterior Probability [P (H/X)]
2. Prior Probability [P (H)]

Where, X is data tuple and H is some hypothesis. According to Baye's Theorem

$$P\left(\frac{H}{X}\right) = \frac{P\left(\frac{X}{H}\right) P(H)}{P(X)}$$

VI. CONCLUSIONS

The proposed work is motivated from the disaster recovery management systems and need of current digital infrastructure. Therefore this paper provides a survey on the current processes of the disaster management and a new suggestion and assumption is made to develop such a framework to reduce the effect of disaster. The proposed framework is further developed as a recommendation engine which accepts the news data for make analysis and generates the place and kind of disaster as prediction according to the news generated.In near future the proposed model is implemented using the JAVA based technology.

REFERENCES

[1] R. W. Perry, E. L. Quarantelli, *"WHAT IS A DISASTER"*, Copyright © 2005 by International Research Committee on Disasters.

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

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

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

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

.