KNOWLEDGE MANAGEMENT IN EDAPHOLOGY USING SELF ORGANIZING MAP (SOM)

[1]A. Meenakshi and [2]V. Mohan

[1]Associate Professor, Department of Computer Science and Engineering
K.L.N.College of Information Technology
Sivagangai District, Tamil Nadu, India
ameenakshiphd@gmail.com
[2]Professor and Head, Department of Mathematics,
Thiagarajar College of Engineering,
Madurai, Tamil Nadu, India

ABSTRACT

In this paper, we propose a proficient method for knowledge management in Edaphology using self organizing map (SOM). The method will assist the edaphologists and those related with agriculture in a big way by finding out the plants apt for the input query. The method has three phases namely dataset processing, neuron training and testing phase. The input data is first converted and normalized in the data processing phase. The SOM is constructed from the processed dataset after the neuron training. The plant name is outputted in response to the input user query in the testing phase. We have added the screen shots of the proposed method in the result section and also evaluated the method with use of evaluation metric values of number of plants retrieved, time of computation and memory usage. The experimental results portrayed that the knowledge engineering approach achieved persistent and compact data storage and faster and knowledge retrieval even for the unknown variables.

KEYWORDS

Knowledge management, Knowledge Retrieval, Soil, Edaphology, SOM, Neural network.

1. INTRODUCTION

Information retrieval (IR) is an area, concern with searching for documents, for information within documents, and for metadata about documents, as well as that of searching structured storage, relational databases, and the World Wide Web [1]. Here, we can observe a rising need for advanced tools that direct us to the kind of information looking for as retrieval results [2]. This directly leads making decisions with uncertain and incomplete data [3]. Examples sources of uncertainty include measurement errors, data staleness and multiple repeated measurements with uncertainty the value of a data item is often representation not by one single value but by multiple values forming probability distributions. Another most encountering problem is dealing with missing values in data set with Homogenous attributes [4,5].

Knowledge Management (KM) is an intelligent process by which the raw data is gathered and is transformed into information elements. These information elements are then accumulate and organized into context-relevant structures [6]. KM is intended to approve ongoing business success all the way through a formal, structured initiative to brighten the creation, distribution, or use of knowledge in an organization [7]. In information sciences to illustrate different levels of abstraction in human centered information processing the data-information-knowledge-wisdom hierarchy is used. For the management of each of them, computer systems can be designed. Data

DOI: 10.5121/ijdms.2012.4507

Retrieval Systems (DRS), such as database management systems, are well appropriate for the storage and retrieval of structured data [8]. Decision support system (DSS) includes knowledge based system well developed to extract useful information from raw data to identify and solve problems and make effective decisions[6]. DSS is implemented by using SOM which is an artificial intelligence discipline. Initially, the dataset is data type converted and normalized so as to make the data fit for further processing. In SOM, data sets are trained using unsupervised learning mechanism by which maps are built to automatically classify a new input vector given as test vector.

To design an efficient intelligent system by converting the existing database into a new knowledge base, we have mainly concentrated only in a particular domain i.e. Soil Analysis System (EDAPHOLOGY). Edaphology is a domain which is concerned with the influence of soils on living things, particularly plants. The term is also applied to the study of how soil influences, man's use of land for plant growth as well as man's overall use of the land. An agricultural soil science explores soil's physical and chemical properties to find the plants appropriate for cultivation [9, 10]. The knowledge management of critical areas within soil observation is still inefficient and in no more than an early stage of development. The most common pitfall information is the lack of standardization of the nomenclature and of the data-acquisition procedures. The user is not required to know fully the model to interact with the system. The retrieval of a large amount of the same type of data is very efficient, even though the user need not know completely the DB schema to formulate the queries. Here the SOM is used in arriving at decisions on what kind of plants can be that grown in soil, based on the domain information (Edaphology-soil characteristics) given by the user. This effectively supports the decision making problems.

2. LITERATURE REVIEW

This section gives the brief review of papers in literature relating to knowledge management and Edaphology. Krista Lagus et al.[1] used WEBSOM for automatic organization of full-text document collections using the self-organizing map (SOM) algorithm. The document collection was ordered onto a map in unsupervised manner utilizing statistical information in case of short word contexts. The resulting ordered map, where similar documents lie near each other thus presented a general view of the document space. With the aid of suitable interface, documents in interesting areas of the map can be browsed and the browsing can also be interactively extended to related topics which appear in nearby areas on the map.

Amarasiri.R et al.[2] presented random weight adaptation scheme which was capable of simulating the effect of presenting the inputs in a random order to self-organizing map algorithms. The resulting effect enabled the inputs to be presented in sequential order and still achieved results similar to that of presenting the inputs in random order. This capability enabled efficient processing of massive datasets. The random weight adaptation was implemented on a growing variant of the self organizing map algorithm called the high dimensional growing self organizing map (HDGSOM). Beaton et al. [3] came up with TurSOM, which stands for Turing self-organizing map which introduced new concepts, responsibilities and mechanisms to the traditional SOM algorithm. It drew its inspiration from Turing unorganized machines, competitive learning techniques, and SOM algorithms. Turing's unorganized machines (TUM) were one of the first computational concepts of modeling the cortex. Turing also described these machines as having self-organizing behaviors. The primary difference between Turing's self-organization description and more traditional models was the fact that it gave importance to connections, rather than neurons, self-organize. TurSOM adheres to unsupervised, competitive learning technique, where in all neurons and connections between them are self-organized and competing.

Chen et al.[4] presented a WordNet-VSM (W-VSM) model for Web services representation which not only enriched the conventional VSM feature vectors' semantic information but also reduced their dimension and sparsity. Then, a set of kernel cosine similarity measures were proposed to well estimate the similarity of the Web services. Le Li et al.[5] proposed a neural network algorithm called uncertain self-organizing map (USOM) which combined fuzzy distance function and self-organizing map to mine and visualize the uncertain data. The self-organizing map assigned the high dimensional data to the corresponding neurons and projected them on a low-dimensional grid which consisted of the neurons. Each neuron was viewed as a small cluster which is a collection of the uncertain data. They merged the neurons in the low-dimensional grid to form the bigger clusters by minimal spanning tree.

3. PROPOSED METHOD FOR KNOWLEDGE MANAGEMENT BASED ON SELF ORGANIZING MAP (SOM)

In this paper, we propose an efficient knowledge management system in Edaphology which is based on construction of Self Organizing Map. The method can also effectively retrieve plant names based on the input query. The method consists of three phases: Dataset processing, Neuron Training phase and Testing Phase. The proposed method will largely help edaphologists and those associated with agriculture in a big way. The block diagram of the proposed method is given in figure 1.

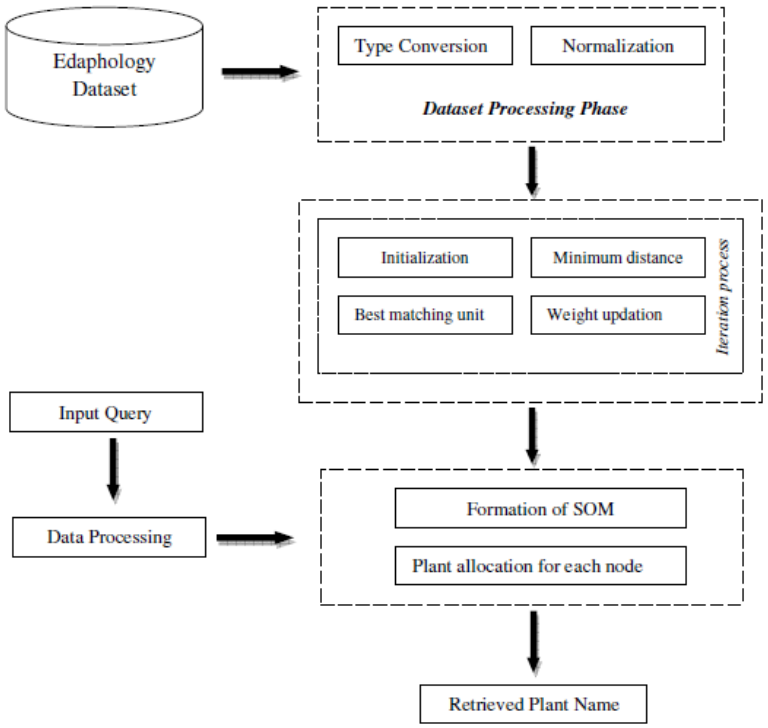

Figure 1. Block diagram of the proposed method

a) Dataset Processing

The dataset of the soil, which consists of the soil characteristics and plant details, is obtained from the information collected by edaphologists. The dataset obtained has many attributes and is bulky; hence we have to reduce it before presenting them to a neural network. This will help in reducing the size of the network, and thereby directly reducing the learning time. Another reason is that it will also reduce the tendency of over-fitting the training data and will reduce or remove the

correlation between attributes that they might slow down the training and reduce the performance of the network. For reducing the dataset, we carry out two processes of Type conversion and Normalization.

1. Type Conversion

The dataset that we have used consists of mixed data having both categorical and numerical data. Initially, we use histogram density function to convert the string to numeric values. The histogram is also made use to efficiently calculate the mean and standard deviation of very large data sets. Histogram samples having the same value are classified as a single group which allows the statistics to calculate values based on few groups rather than a large number of individual samples. The number of data is taken as total area of the histogram. A histogram can also be normalized by making use of relative frequencies which is a function that counts the number of observations that fall into each of the disjoint categories. In type conversion, each column of categorical variable from the dataset is read to find the unique string values. Let W_{US} be the unique string word, F_{US} be the frequency of the unique string word and N_{CV} be the number of categorical value, then the probability of unique string value P_{US} is obtained using the formula

$$P_{US} = \frac{F_{US}}{N_{CV}}$$

2. Normalization

In normalizing the type converted data, the goal is to ensure that the statistical distribution of values for each net input and output is roughly uniform. In addition, the values should be scaled to match the range of the input neurons. This means that along with any other transformations performed on network inputs, each input should be normalized as well. Normalization of data using z-scores overcomes objections of relativism which can be applied to methods that allocate points pro data. It converts all indicators to a common scale with an average of zero and standard deviation of one. The average of zero means that it avoids introducing aggregation distortions stemming from differences in indicators' means. The scaling factor is the standard deviation of the indicator across the converted values which are being ranked. Thus, an indicator with extreme values will have intrinsically a greater effect on the composite indicator. The z-scores can then be converted to indices, so all scores for all rankings are within the same range. Let the values of the i^{th} column be represented by C_{ij} , $0 < j \le N_{ic}$ where N_{ic} is the total number of values in the i^{th} column. Hence the mean μ_i and standard deviation σ_i of the i^{th} column can be equated from:

$$\mu_i = \frac{\sum_{1}^{N_{ic}} C_{ij}}{N_{ic}}, \quad {}_i = \frac{\sum_{1}^{N_{ic}} C_{ij}{}^2}{N_{ic}}$$

$$\sigma_i = \sqrt{{}_i - (\mu_i)^2}$$

Hence, the normalized values N_{ij} can be calculated by subtracting mean of the column from the original value and dividing by the standard deviation of the column. The equation for finding the normalized value is given by:

$$N_{ij} = \frac{C_{ij} - \mu_i}{\sigma_i}$$

Hence, we have type converted and normalized the input dataset to improve a network's performance.

b) Neuron Training Phase

The artificial neural network that we have used is the Self Organizing Map (SOM) which is also called as Kohonen Map. In SOM, training using unsupervised learning mechanism is carried out in which the map is built using input examples which is known as vector quantization. Three different types of network initializations are there which includes random initialization, initialization using initial samples and linear initialization. The structure of the SOM is given in figure 2.

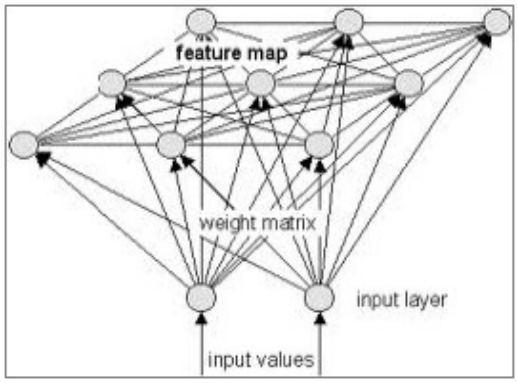

Figure 2. Structure of Self-Organizing Map

In our paper, we make of random initialization, in which random values are assigned to vectors. This case comes in handy as nothing or little is known about the input data at the time of the initialization. Let the input vector be represented by V_i, $for\ 0 < i < N_V$ where N_V is the dimension of the vector input. Training is an iterative process and consists of drawing sample vectors from the input data set and "teaching" them to the SOM. Here let the initial weight vector be represented as W. The teaching consists of choosing a winner unit by the means of a similarity measure and updating the values of vectors in the neighborhood of the winner unit. The process is carried for N_L number of iterations. In one training step, one sample vector is drawn randomly from the input data set. This vector is fed to all units in the network and a similarity measure is calculated between the input data sample and all the vectors. For this, traverse through each node and calculate the minimum distance between the weight vector and input vector is calculated. Considering i^{th} node, minimum distance d_i between the weight vector W and input vector V is calculated as:

$$d(W,V) = \sqrt{(W_1 - V_1)^2 + (W_2 - V_2)^2 + ... + (W_{N_V} - V_{N_V})^2}$$

The best-matching unit (BMU) is chosen to be the vector with greatest similarity with the input sample which is defined by means of a distance measure. After finding the best-matching unit, units in the SOM are updated. During the update procedure, the best-matching unit is updated to be a little closer to the sample vector in the input space.

$$W_v(t+1) = W_v(t) + \varphi(v,t).\alpha(t).(D(t) - W_v(t))$$

Where, t denotes current iteration, $W_v(t)$ is the current weight vector, $D(t)$ is the target input, $\varphi(v,t)$ is the neighborhood function and $\alpha(t)$ is the learning restraint due to time. The topological

neighbors of the best-matching unit are also similarly updated. The update procedure stretches the BMU and its topological neighbors towards the sample vector. Once the training of the SOM is finished, the plant is set for each of the nodes based on the values.

C) Testing Phase

In this phase, the user gives a query for which the corresponding plant name is outputted. Here, the input query is initially type converted and normalized and then checked in the trained SOM. The system in response to the processed input query returns the plant names which is similar to input query which is done by the trained SOM. It also shows the similarities for target plant and input vector which is based on cosine similarity. For any unknown query, the system returns plant which is closest to the target plant based on similarities. That is, if the input sample data is not matched with best matching unit, then the input sample data is checked for next similarity with neighbors. Hence, in short when a user enters a query, it outputs the plant names which best matches the input query.

4. RESULTS AND DISCUSSION

This section presents the results and discussions of our proposed method for SOM based knowledge retrieval in Edaphology. Initially, SOM is trained and then the test query is given to output the plant names. Here, we measure the number of plants outputted for the query, time of computation and the memory used.

4.1 Experimental Set Up and Dataset Description

The proposed method is implemented in JAVA on a system having 4 GB RAM and 2.10 GHz Intel i-5 processor. Initially, the domain knowledge collected from edaphologists is modelled into a knowledge base, which acts as the input data set. The input dataset is type converted and normalized before training and testing processes.

4.2 Screen Shots of the Proposed Method

In this section, we give the screen shots of the proposed method. Figure 3 shows the home screen, figure 4 shows the dataset connection, figure 5 displays the details, figure 6 shows the type converted data, figure 7 shows the normalized data, figure 8 shows the SOM values formed , figure 9 shows the plant assignment to the node and figure 10 shows the user interface for the input test case.

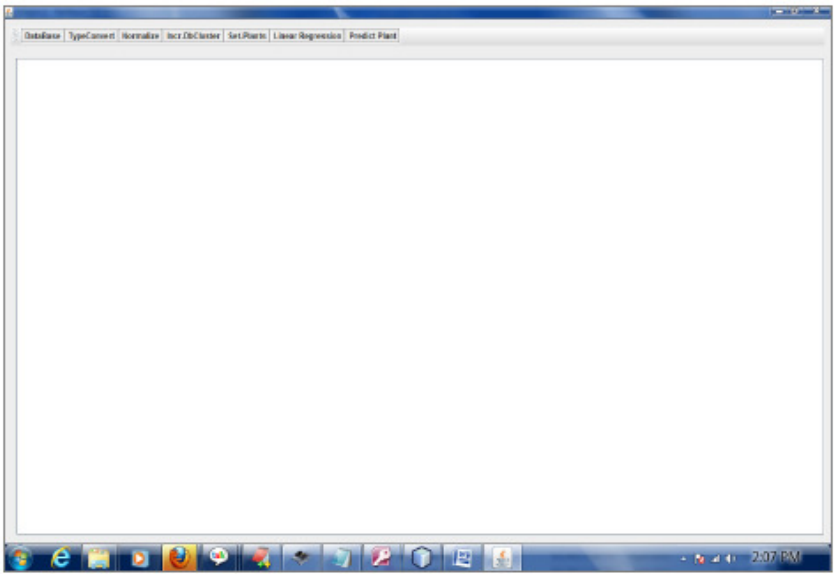

Figure3. Home window

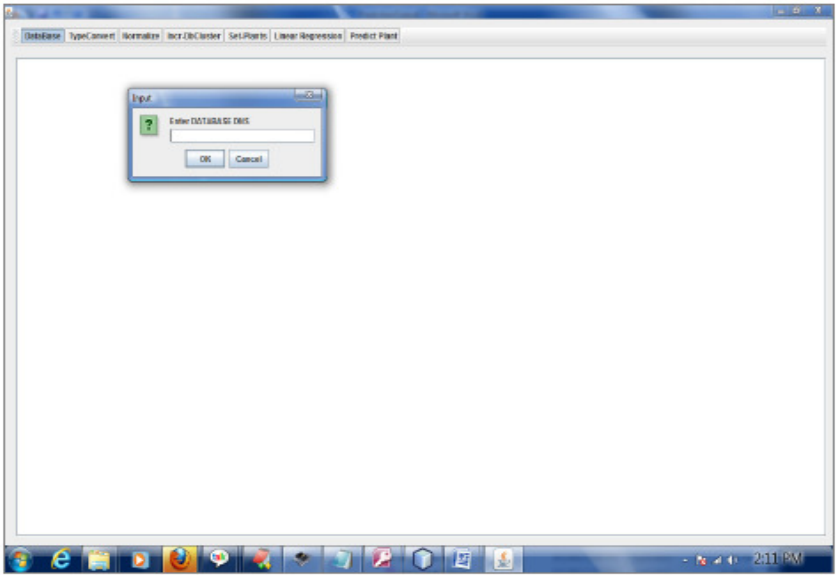

Figure 4. Dataset Connection

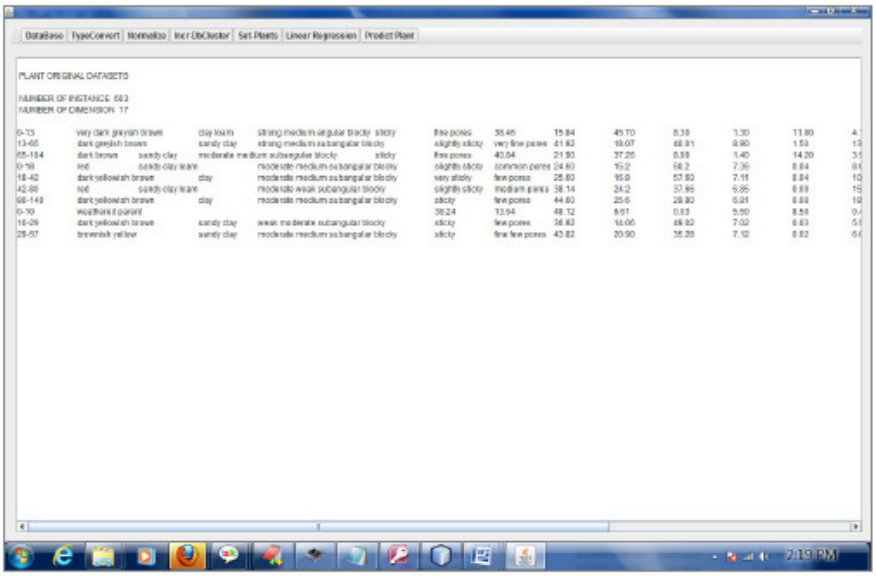

Figure 5. Showing the plant and soil details of the dataset

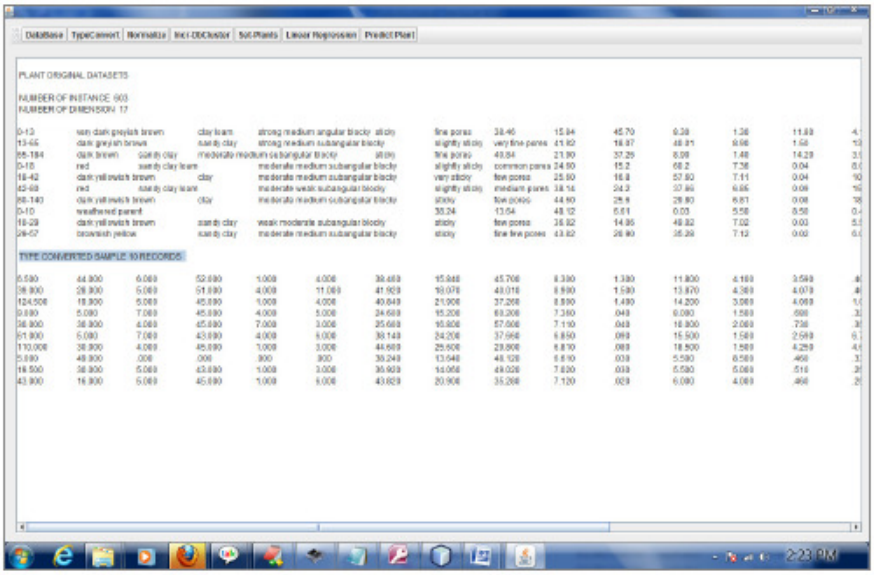

Figure 6. Showing the type converted data

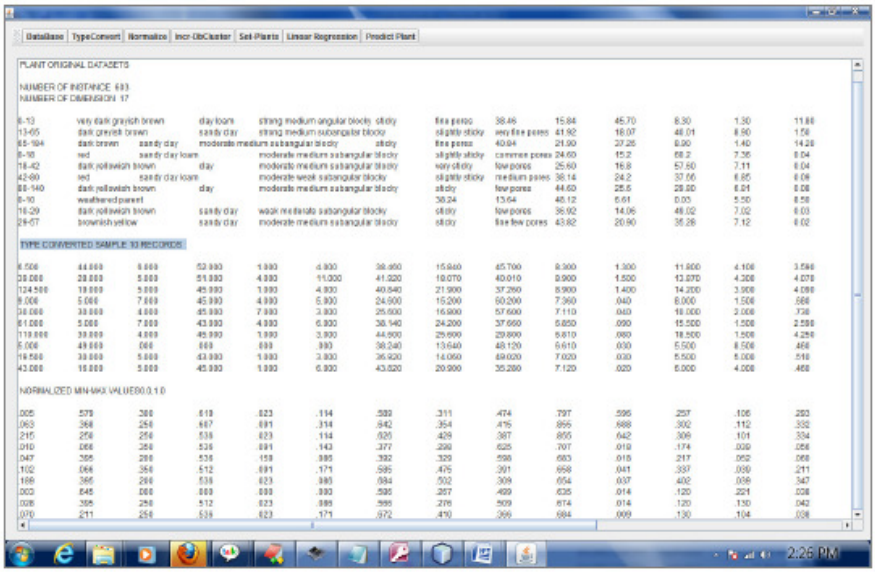

Figure7. Showing the Normalized data

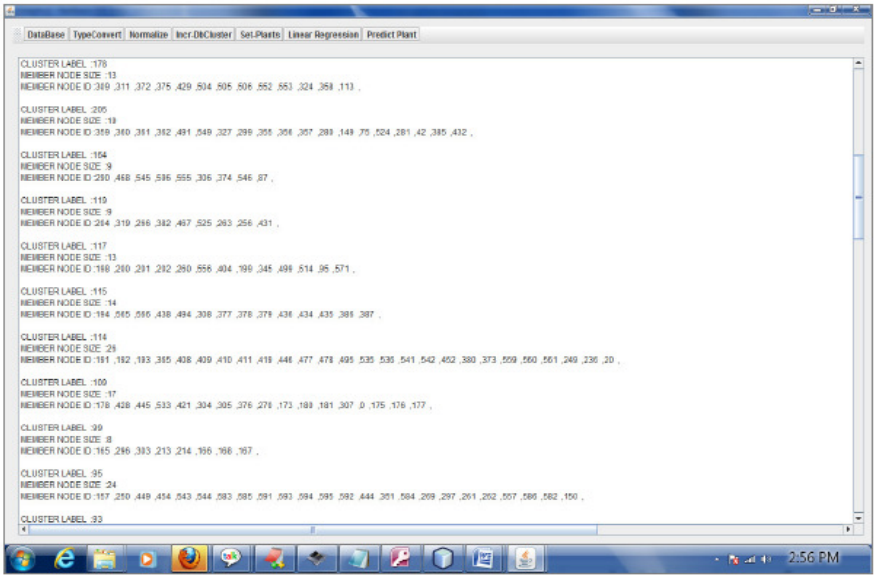

Figure 8. Formation OF SOM values

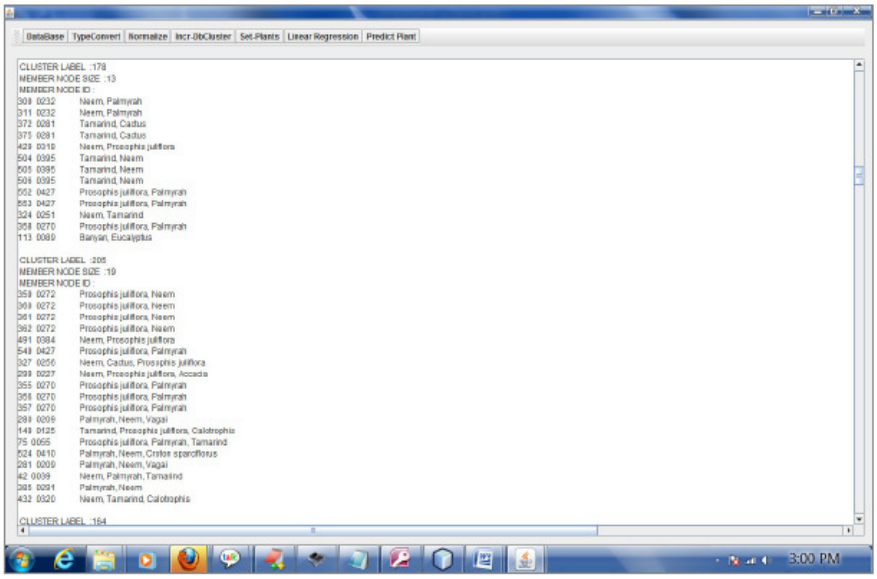

Figure 9. Assigned plant to the nodes

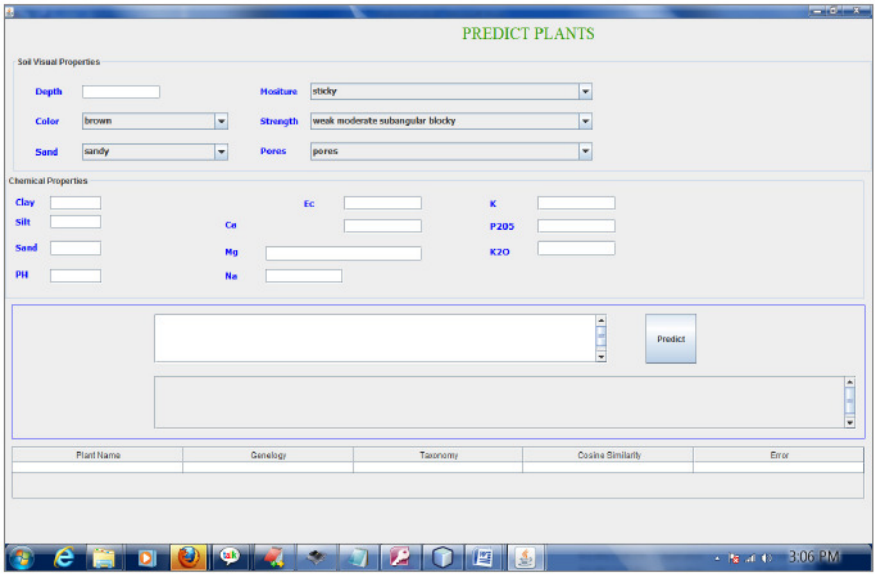

Figure 10. User interface for test case input query

4.3 Performance Metrics

In order to find the performance and to evaluate our proposed method, we make use of certain parameters that constitute to the performance metrics. Selection of performance metrics parameters is of high importance as it should give a clear cut idea of how well the method works and also should be able to validate the effectiveness of the method. Here in this paper, we make use of three parameters that form the evaluation metrics.

Number of plants retrieved: The input to the method will be a user query which will have the soil characteristics and the output will be the plant list which will have the names of plants that satisfy the input user query. As the number of plants retrieved increases, the effectiveness of the plant retrieval method also increase.

Computation time: Computation time refers to the time incurred between the input query and the output list. Reduction of the computation time show better and faster processing of the query.

Memory usage: The amount of memory used up while executing the query is known as the memory usage. Having a lesser memory usage will validate the effectiveness of the method.

Figure 11 and figure 12 shows the two outputs obtained for two different input queries. From the figure we can see the number of plants retrieved for query 1 is four and that for query 2 is two.

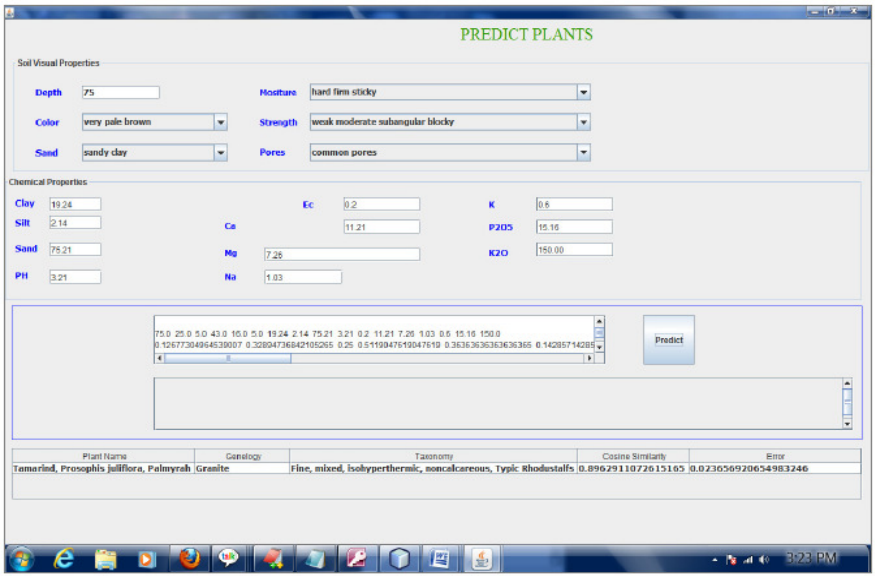

Figure 11. Result obtained for test case input query 1

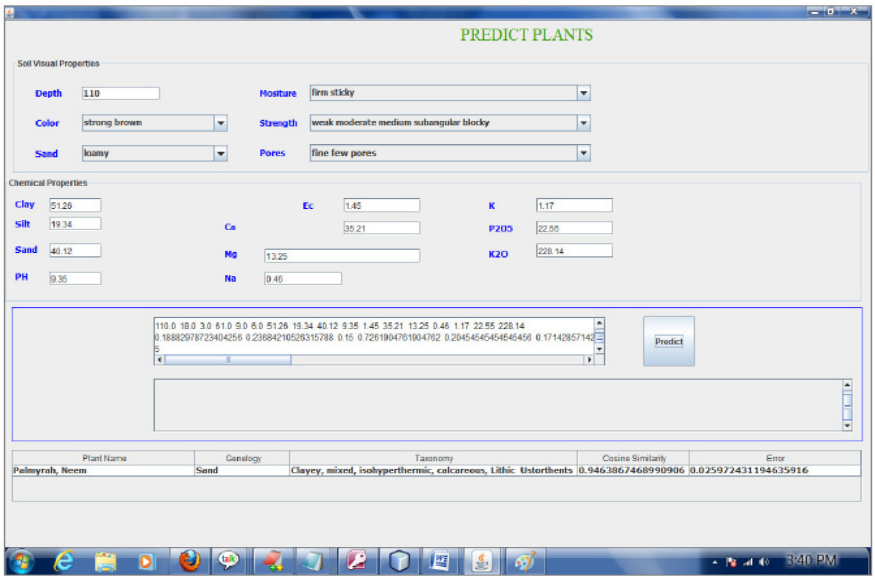

Figure 12. Result obtained for test case input query 2

The time taken for execution for query 1 came about 0.303 s and that of query 2 came about 0.309 s. The memory usage of query 1 came about 8987 kB and for query 2 it came about 9678 kB. The values obtained the fact that our proposed method have performed well and performs effective retrieval based on the input query.

5. CONCLUSION

We have proposed an efficient knowledge management system based on Self Organizing Map for handling the information which is in the form of knowledge. The Edaphology information that we have used in the method was collected from the edaphologists. Here, it has three phases namely dataset processing, neuron training and testing phase. The SOM is constructed from the initial two phases and the plant name is outputted based on the input query in the testing phase which is based on the earlier constructed SOM. The experimental results portrayed that the knowledge engineering approach achieved persistent and compact data storage and faster and knowledge retrieval even for the unknown variables.

REFERENCE

[1] Y. Li and N. Zhong "Web Mining Model and its Applications for Information Gathering", *Knowledge-Based Systems*, Vol.17, pp.207-217, 2004.

[2] Bjoern Koester, " Conceptual Knowledge Retrieval with FooCA: Improving Web Search Engine Results withContexts and Concept Hierarchies", *Advances in Data Mining,* pp. 176-190, 2006.

[3] Brian D. Newman, Kurt W. Conrad," A Framework for Characterizing Knowledge Management Methods, Practices, and Technologies" *In Proc. of the Third Int. Conf. on Practical Aspects of Knowledge Management*, pp.30-32, 2000.

[4] Y. Li and Y. Y. Yao, "User profile model: a view from Artificial Intelligence", *In proceedings of 3rd International Conference on Rough Sets and Current Trends in Computing*, pp. 493-496, 2002.

[5] Xiaohui Tao, Yuefeng Li, and Richi Nayak, " A Knowledge Retrieval Model Using Ontology Mining and User Profiling", *Integrated Computer-Aided Engineering* **,** Vol.15, No.4, pp. 1-24, 2008

[6] Apistola, M., L. Mommers, and A. Lodder (2001), A Knowledge Management Exercise in the domain of Sentencing: towards an XML Specification, *In: Proceedings of the Second International Workshop on Legal Ontologies, Amsterdam*, The Netherlands: December 13, 2001, pp. 49-57.

[7] Denning, S., "The role of ICT's in knowledge management for development", *The Courier ACP-EU*, No.192, pp. 58 - 61, 2002.

[8] Rizwana Irfan and Maqbool-uddin-Shaikh , "Enhance Knowledge Management Process for Group Decision Making", *In Proceedings of World Academy of Science, Engineering and Technology*, 2009.

[9] Bui, Henderson, and Viergever, "Knowledge discovery from models of soil properties", *Ecol. Model*, Vol.191, pp.431–446. 2006.

[10] Bui, "Soil survey as a knowledge system", *Geoderma*, Vol.120, pp.17–26, 2004.

[11] Krista Lagus, Timo Honkela, Samuel Kaski and Teuvo Kohonen, "Self Organizing Maps of Document Collections: A New Approach to Interactive Exploration", 1996.

[12] Amarasiri.R, Alahakoon.D, Premarathne.M, Monash Univ and Clayton, " The Effect of Random Weight Updation in Dynamic Self Organizing Maps", *In proceedings of International Conference on Information and Automation*, 2006.

[13] Beaton, Valova, MacLean, "TurSOM: A Turing inspired Self-Organizing Map" Neural Networks, *In proceedings of IJCNN International Joint Conference*, 2009.

[14] Chen, Lei Yang, Zhang, Yingzhou, Zhengyu Chen, "Web services clustering using SOM based on kernel cosine similarity measure", *In 2nd International Conference of Information Science and Engineering (ICISE),* 2010.

[15] Le Li, Xiaohang Zhang, Zhiwen Yu, Zijian Feng, Ruiping Wei, "USOM: Mining and visualizing uncertain data based on self-organizing maps", *In proceedings of International Conference on Machine Learning and Cybernetics (ICMLC)*, 2011.