DATA EXPLORATION USING HYBRID SYSTEMS IN INFORMATION RICH DOMAIN

E. Subramanian[#1]

[#1] Department of Computer Science and Engineering,

CARE School of Engineering, Tiruchirappalli, Tamil Nadu. India

esmani84@gmail.com

***Abstract* - Effective scanning and quickly identifying key information are two of the most crucial skills for the users in today's information intensive domain. The users need the methods to understand their information domain as well as to integrate the understanding into their planning and decision making processes. The term Hybrid System (HS) refers to the methods to achieve this integrative understanding. This Hybrid System is made up of three components, i) a Hierarchical Hyperbolic Self Organizing Map (H^2SOM) for structuring the information domain and visualizing the information with respect to identified topics, ii) a spreading activation network for the selection of the most relevant information sources with respect to an already existing knowledge infrastructure and iii) measures of interestingness for association rules as well as statistical testing that facilitates the monitoring of already identified topics. These three modes can support the active organization, exploration and selection of information that matches the needs of decision makers in all stages of information gathering process.**

Keywords – Self Organizing Map, Hybrid System, Hierarchical Hyperbolic Self Organizing Map.

I. INTRODUCTION

The term Hybrid System comprises of the Finding phase, Growing Phase and the Monitoring phase and it collectively encourages an integrative understanding. The information seeking process of HS can be split into three stages as follows,

- Finding the new information should be largely unguided by a priori knowledge of the user, in order to prevent any restriction to the already known areas in the information domain.
- During the growing of already existing knowledge, the user follows the well defined information interests and focuses more on particular aspects that are considered to be worth investigating in more detail.
- In the observing stage there is a shift from detection to monitoring of update in the information domain and consequently the source is updated to the changes in already explored areas of the environment.

In the **Finding phase** the Hierarchical Hyperbolic Self Organizing Map (H^2SOM) [3] is used for structuring the information and Visualization can be done. The H^2SOM methodology provides a similar interactive fish eye interface as like in the SOM tree browser but it also allows the users able to learn the hierarchical structure of an unstructured information source in a completely unsupervised way. Therefore, it allows the user to have an initial overview of a large document collection, which can be interactively explored.

In the **Growing phase** the network can be spread out of already existing knowledge. For that Information Foraging Theory (IFT) [5] is used which is a imperfect perception of the value, cost, or access path of information is labeled as information scent, which can be modeled by means of a spreading activation network. Information foraging theory is an approach to understanding how strategies and technologies for information seeking, gathering, and consumption are adapted to the flux of information in the domain. The theory assumes that user, when possible, will modify their strategies or the structure of the domain to maximize their rate of gaining valuable information.

In the **Observation Phase** the measures of interestingness for association rules as well as statistical testing that facilitates the monitoring of already identified topics. An already identified topic is commonly described by a set of relevant terms. If these terms are co occurring in several documents, the set of terms makes up a frequent item set. Monitoring changes of patterns in the course of time has been of interest to support a variety of different decision making tasks.

The foremost necessitate of this system is the data in the information rich domain plays an important role to find the relevant information in the profit time. The need is to minimize the time consumption of the user and also to

provide the visual understanding to the user in order to integrate the gathered knowledge. The gathered knowledge can be used in the decision making process of the user. Though there are several algorithms are used by the researchers in before each lack in one the criteria of the Hybrid Systems. So for overcoming of the demerits of that algorithms and providing integrative understanding Hybrid System has been chosen.

A. *Related Work*

The Hybrid System can be simplifying the visualization to the user as well as the growth according to the user perspective. The changes in the data's can also be monitored and the changes if any changes can also be reflect back to the original data.

The work [1] K. Carley dealt with the problems that have been faced by organizations in complex domains in terms of the number of coexisting issues competing for attention and resources in the growing step. Identification of key economic, social, and technology related issues affecting the organization, their life cycle stages, and their features relative to each other would help allocate attention and resources to them. Decision makers clearly need robust information technologies capable of helping them assess issues in an accurate, timely, and efficient way. So, growing has become a fundamental, early step in developing strategies that help an organization adapt to its domain. Online Computerized Data Bases (OCDB) has not found their way into the growing process. The most attractive features of OCDB's are "keyword" search protocols to longitudinally track several issues at once over several time intervals. OCDB's are accessed using keywords. A data base search protocol can be readily designed to, not only produce information about how frequently one of these issues has been cited across a broad cross section of media at a point in time, but also geographically where, what sources cited it and whether that issue was even associated in the same media source with any other issues of interest. The data's used for decision making in an organization can be stored and cited in different locations. So the organizing and accessing the database stored in the data warehouse is a tough process and the visualizing those data's is also a relative difficult one.

The work [2] Farid dealt with the modern data warehouses in which the data is stored distributed across multiple data centers and maintained by different database servers. Identifying similar or semantically equivalent attributes over different databases is essential toward defining data integrity constraints. There are two issues with making use of remote data sources. Firstly the discovery of relevant data sources and performing the proper joins between the local data source and the relevant remote databases. Both can be solved if one can effectively identify semantically related attributes between the local data source and the available remote data sources. Visualization can be done based on Self Organizing Maps (SOM). SOM benefits from fast training algorithms and is easy to visualize due to its low dimensional regular grid layout. SOM has been applied in numerous works in visualization of text corpus. Data is from general databases in addition to text data there is also a numerical data to handle. Self Organizing Map (SOM) can be used to meet the challenge of extracting Coincident Meaning from Heterogeneous textual and numerical data types, by extending the traditional text analysis from information retrieval. For that the author proposes a new algorithm, named Common Item set Based Classifier (CIBC) that enhances the results obtained from SOM's trained map. It improves the precision and heterogeneity of clusters and helps differentiating visually between heterogeneous and homogenous data clusters. SOM based visualization technique can allow the end users to perform data

integration and schema mapping in a semi automated fashion. This allows the users to have a clear understanding of the relationship among attributes from different and unfamiliar databases. The numerical data is preprocessed differently from the textual one.

II. METHODOLOGY

Existing automatic HS systems lack in at least one of the three modules like Finding, Growing and Observing. With respect to the HS process of three stages, the system does not support for the finding stage, since a pre - specified set of event topics and their respective properties have to be defined beforehand by the user. Moreover, the system does not present the results in any visual context. So the Hierarchical Hyperbolic Self Organizing Map (H^2SOM) can be introduced which supports the all three phases of Hybrid Systems.

B. Hierarchal Hyperbolic Self Organizing Map (H^2SOM) Method

In H^2SOM [2] the nodes of the H^2SOM are embed in hyperbolic space. The H^2SOM uses the SOM kohonen's Algorithm for finding the relevant items. The steps to arrive this H^2SOM Network Architecture has can be given as in fig 2.1,

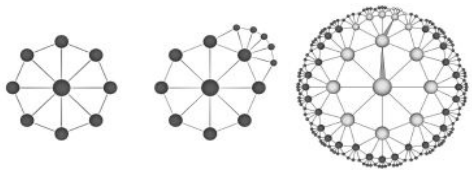

Fig. 2. 1. H^2SOM Network Architecture

1. *Initialization:* A ring of n_b equilateral hyperbolic triangles centered at the origin of Information Hierarchy. The $n_b + 1$ vertices of the triangles form the network nodes of the first level of the hierarchy.
2. *Growth Step:* During a growth step, nodes in the periphery of the existing network are extended by the vertices of additional n_b - 2 equilateral triangles. The branching factor n_b determines how many nodes are generated at each level and how fast the network is reaching out into the hyperbolic space.

C. Foraging Method

Information Foraging Theory (IFT) [5] can be used in the Growing phase where the network can be spread out of already existing knowledge. The basic model of IFT is based on the Patch and Prey Model of Optimal Foraging Theory, where foraging theory is the biological term used in an assuming that information is structured in patches and the user has to decide which patches should be visited, for an amount of time. The user uses proximal cues to assess profitability and the prevalence of information sources and their representations by means of documents contents.

This imperfect perception of the value, cost, or access path of information is labeled as information scent, which can be modeled by means of a spreading activation network. Information Diet specifies which is the best content of all the document content based on the ranking, given by the number of users visited or grown the document label. Diet is very much useful for drilling down the information gathered by the user and to make the plan accordingly. Information foraging theory is an approach to understanding how strategies and technologies for information seeking, gathering, and consumption are adapted to the flux of information in the environment. The theory assumes that user, when possible, will modify their strategies or the structure of the environment to maximize their rate of gaining valuable information. Information Foraging Theory consists of three methodologies,

1. Information Scent

Instead of gaining a thorough understanding of the semantic content, users typically engage simple heuristics by means of using single terms and phrases without their semantic context as cues for the appraisal of the distal information inherent in a document. This helps to predicting a path's success and giving answers to the following questions,

- Does the page navigation signal to the user that they have reached, or are nearing their goal?
- Does the destination page meet the expectations set by the navigation?

2. Information Diet

The construction of the optimal selection of relevant documents, denoted as information diet in IFT terminology. IFT states that the behavior of information users follows the maximization of the rate of gain of valuable information per time unit cost is given in Equation 2.1,

$$\text{Max } R = G / B+T \qquad (2.1)$$

Where R referring to the ratio between the total net amount of valuable information gained, G, and the total time spent on searching this information, e.g., the cumulative time spent on switching from one information source to the next, B, and the total time spent on extracting and handling the relevant information from the information representations, T. In a partitioned information environment, the average ratio of gain of information of each information patch h can be given in an Equation 2.2,

$$R(I_h) = \frac{\sum_{i=1}^{Ih} \lambda h \,.ghi}{1+\sum_{i=1}^{Ih} \lambda h \,.thi} \quad \text{with } T = \sum_{h=1}^{H} \sum_{i=1}^{Ih} thi \qquad (2.2)$$

Where the assessed gain g_{hi} of document d_{hi} £ I_h is extracted in time t_{hi}. Actually, t_{hi} is the time needed to read and comprise document d_{hi} in its entirety. The prevalence of relevant documents in information patch h is denoted by the encounter rate $\lambda_h = 1 / b_h$, which is derived from the average time needed to encounter a document. It should be noted that the gain of documents might differ from environment to environment, because then documents might differ with respect to length, structure, and so forth, and, most importantly, relevance. The profitability p_{hi} of each document d_{hi} is given by $p_{hi} = ghi / t_{hi}$.

3. Diet Selection Rule

Starting from the heuristic assessment of the profitability p_{hi}, the following algorithm can be used to determine the rate maximizing subset of information sources given in information environment h that should be selected and extracted by the user,

a) Rank the information sources by their information profitability. For simplicity of presentation, let the index i be ordered such that $p_{hi} > p_{hi(i+1)}$.

b) Add further information sources to the diet until the rate of gain of information for a diet of the top N_h information sources is greater than the profitability of the $(N_h + 1)$ information source. The condition checking is given in Equation 2.3 as,

$$R(N_h) = \frac{\sum_{i=1}^{Nh} \lambda h \,.ghi}{1+\sum_{i=1}^{Nh} \lambda h \,.thi} > \frac{gh(Nh+1)}{th(Nh+1)} = Ph(Nh+1) \qquad (2.3)$$

The above equation provides a criterion for the selection from an ongoing stream of information and therefore it advises how to dedicate the limited attention and time of the user to the documents aligned to one

patch, by recommending which documents should be actually considered.

D. Monitoring the Changes

An already identified topic is commonly described by a set of relevant terms. If these terms are co occurring in several documents, the set of terms makes up a frequent item set. Utilizing rules by dividing the terms into two subsets the rule antecedent and the rule consequent enables a more precise assessment of the topic by quantitative measures of the rule. Such rules fit into the mental schemes of users, whose decisions are more typically based on if - then knowledge. Monitoring changes of patterns in the course of time has been of interest to support a variety of different management tasks.

Mining for new association rules in a document stream to detect upcoming topics or events is vulnerable by the definition of frequent item sets. In early stages, only very few insiders might discuss the sense and meaning of a specific event or development, compare it to alternative events, or exchange ideas on how to make up the event. Due to the very few people involved, the number of documents related to the event is likely to be low, and therefore, the event is likely to be excluded by the algorithms for mining association rules. Moreover, in the early stages of an event when the signal is imprecise and fuzzy and, therefore, the terms used to describe and discuss the events are likely to be in general rather than specific terms, the event is hardly separated from background noise.

So the alternative technique can be attempted by using statistical properties of association rules. For this the document can be divided into partitions called regimes. Each message d is assigned to regime ד to capture changes in the course of time. $M_{ד}$ be the set of all messages assigned to regime ד and $w_{m_{ד}}$ the set of words occurring in a message d. then the support, confidence and lift for the rule A→ C can be calculated as,

The Support is given in Equation2.4,

$$\text{Sup}_r\ (A \rightarrow C) = \frac{|Wmr \in Mr|(A \cup C) \subseteq Wmr}{|Mr|} \qquad (2.4)$$

The Confidence is given in Equation 2.5,

$$\text{Conf}_r\ (A \rightarrow C) = \frac{|Wmr \in Mr|(A \cup C) \subseteq Wmr}{|Wmr \in Mr|\ |A \in Mr|} \qquad (2.5)$$

The Lift is given in Equation 2.6,

$$\text{Lift}_r\ (A \rightarrow C) = \frac{Supr\ (A \rightarrow C)}{Supr\ (A)\ .\ Supr\ (C)} \qquad (2.6)$$

By using these statistical measures the changes in pattern can be detected and it can be updated to the already known information.

E. Overall Design

The design consists of three modules, Discovering and Structuring, Growth and Observing phases. In this the dataset can be get it from the user, which are organized and visualized to the user by using kohonen's algorithm and the visualized information can be grown primarily. The visualized result will affect if there is any changes like if modifying or updating the original dataset. So the changes can be reflecting back to the original dataset as well as to the visualized result. This can be acquired by using some simple association rules. Fig 2.2 represents the design,

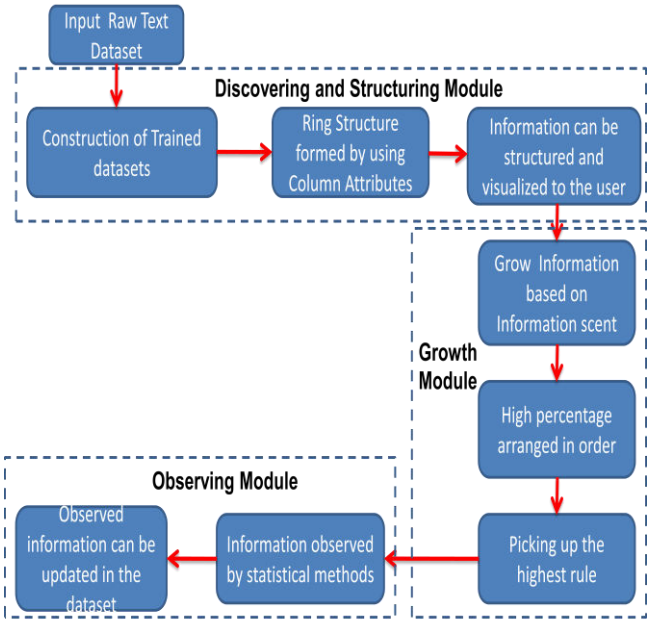

Fig 2.2 Overall Design

In the Discovering and Structuring module, the dataset can be taken as an input and the column attributes of the dataset can be recognized by using the Best Matching Unit (BMU) Concept in Kohonen's Algorithm. The Kohonen's Algorithm is used for visualizing purpose. The hierarchal form structure can be obtained by using H^2SOM. In the algorithm the learning starts by initializing the weight of the node. Then the vector can be chosen in random from the training dataset. Every node is examined to calculate which one's weights are most like to the input vector. The winning node is commonly known as the Best Matching Unit (BMU). Then the neighborhood BMU is also calculated. The recognized data can form a ring node structure and the hierarchal structure obtained by using the H^2SOM.Then the ring node data can be structured by using Common Item Based Classifier (CIBC) Method. The CIBC Method mainly clusters the data labels accordingly based on the input given by the user and main focus is to smooth the clusters got from the kohonen's algorithm outcome.

In the Growth module, initially, the structured data can be grown hierarchal. The prior knowledge of the drilling data is must for the user. The structured data is visualized to the user in the form of the nodes. Instead of displaying the entire data from the taken dataset, only the selected portion of the data can be shown to the user by using Information Diet. For that Gain of all data's can be calculated, from that the gain can be ordered by highest to lowest. The data's are back to back clustered and the portion chosen by the user can be displayed to the user. The highest of all data's from the taken dataset can be extracted by calculating the maximum gain. The maximum gain can be taken out and displayed to the user at first, once the dataset are imported. Then the user perspective grown results can be visualized to the user for the support of the decision making process. User can also narrow the results, if the desired results are obtained.

In the Observing module, the user can get some decision from the visualization acquires from the growth step. The structured information or data can be affected while the original dataset undergone any modification or revision. That can also be reflected to the structured as well as the visualized data. To monitor those changes simple association rules are used. By these rules the data in the taken dataset can be compared at a particular time interval with the structured and visualized information or data. If there is any revision then the changes can be return back to the taken dataset. The visualized information can also be changed accordingly.

This monitoring can be done by the statistical rule which can calculate the Support, Confidence and Lift for the taken dataset. The taken dataset can be monitored in the particular time interval. At that occasion, the statistical measure information can be compared each and every time. This can also deals with the interest based on the number of times the users access the particular page and profitability. The highest profitability information can be calculated by using CARMA algorithm. This algorithm uses the classification and the association. The data from the taken dataset can be classified first and then the data's are associated by the statistical measures.

F. Dataset

The dataset taken for this work is County data from the 2000 presidential election in Florida which contains the nominees and the voting to the nominees in the presidential election. This dataset can be given as input and it can formed as a ring structure by using H^2SOM and initially grown by clicking in the ring structure which can be formed by the column of the dataset. At the start the datasets with four columns can be taken as input. The Dataset can be grown in the order of Country, States, Districts and then finally by the Nominees. From this the Nominees in particular Countries, States and District votes can be inferred and predict the chances in the next presidential election.

III. PERFORMANCE RESULTS

In the Discovering and Structuring phase the hyperbolic tree has high fast search capability in which we taking initial center node as the search root and then following the hierarchical structure in the hyperbolic grid. Starting from this node, we recursively determine the k best matching nodes among its n_b neighbors until we reach the goal. For finding the neighbors log n_b comparisons made so the complexity is O ($\log_{nb}$ N). This will continue for the k best match so that the entire complexity of this phase is O (K $\log_{nb}$ N).

In the Initial Growing phase we growing the activation network based on the information gain. The information gain can be calculated first and that information gain can be sorted into an order. Association strength is also calculated for the growth information. So the complexity for this is O (K N^2) where k is the number of words and N is the number of documents.

The Observing phase has a complexity of O (V. N ד) with V denoting the number of all distinct items included in any of the rules and N ד denoting the number of documents included in regime ד .The most critical part of the system that is responsible for the initial filtering of the data scales very favorably with logarithmic complexity, i.e., even if the data sources grow very strong, the H^2SOM clustering will be capable of dealing with that.

IV. CONCLUSION AND FUTURE ENHANCEMENTS

Using this Hybrid Systems integrative understanding can be easily achieved. By Hybrid Systems the time and efforts spend by the users for searching is minimized. By the visualization the users can also got a clear picture of what they inferred from the previous data and the present data. Initially gathered information can be grown by the users according to their perspective by implemented foraging concepts. The changes can also be reflecting back to the user's if there is any revision in the data when we compared to the original dataset. In collection, using Hybrid Systems, the visualization of data which provides understandings, initial exploration of data and the active changes can be achieved. Support and Confidence in the Association Rule Mining helps to obtain a rank for the visited information or data and the highest rank is considered as a prioritized one for the user to minimize the effort of finding the information.

This work can do the visualization for the offline data and there is problem in fish eye browser if the user chooses the large number of the dataset. The visualization can't be obtained properly for the huge datasets such as online data. The Hyperbolic Multi – Dimensional System can be used in future to overcome the visualization problem as well as for the online data. Instead of using Hierarchical Hyperbolic Self Organizing Map the Hyperbolic Self Organizing Map can be used to overcome the visualization problem. In hyperbolic the visualization can be done be at positive and negative edges. So that if the user can increases the dataset the data can fully visualized to the user and from that the user can easily obtains the results.

V. REFERENCES

[1]. Farid Bourennani, Ken Q. Pu, Ying Zhu, "Visualization and Integration of Databases using Self – Organizing Map", First International Conferences on Advances in Databases, Knowledge, and Data Applications, pages 77 – 88, 2009.

[2]. J. Walter, J. Ontrup, D. Wessling, and H. Ritter, "Interactive Visualization and Navigation in Large Data Collections Using the Hyperbolic Space," Third IEEE International Conference on Data Mining, pages 123 – 131, 2003.

[3]. J. Ontrup and H. Ritter, "Concept-Based Clustering of Textual Documents Using HSOM," IEEE Transactions on Neural Networks, vol. 19, no. 6, pp. 751-761, 2006.

[4]. Yang Yang, Yanning Zheng, "A Review of Information Acquisition Based on Information Foraging Theory", Second International Symposium on Knowledge Acquisition and Modeling, pages 409 – 419, 2009.

[5]. J. Sue Warcup, Don Zimmermon, D. Wessling, "Geo – Historical Context Support for Information Foraging," Third IEEE International Conference on Data Mining, 2009.

[6]. Kantardzic M.M, Pandit A. A, "New Hierarchical Model using Self Organizing Map improves accuracy of classifiers for Multi class Datasets", International Conference on Information, Communication and Automation Technologies, pages 1-6, 2011.

[7]. Bauer H. U, Villmann T, "Growing a hyper cubical output space in a Self Organizing Map", IEEE Transactions on Neural Networks, vol. 8, issue no. 2, pp. 218-226, 1997.

[8]. Chihli Hung, Jian – Je Huang, "Mining rules from One Dimensional Self Organizing Map", International Conference on Innovations in Intelligent System and Applications, pages 292-295, 2011.

[9]. Zijian Feng, Xiaohang Zhang, Le Li, "USOM: Mining and visualizing uncertain data based on Self Organizing Map", International Conference on Machine Learning and Cybernetics, pages 804-809, 2011.