Market Intelligence Data Collection from Heterogeneous Sources with Similarity-Based Selection Clustering Technique Using Knowledge Maps: A Heuristic Approach

Tapati Bandopadhyay

ICFAI Business School
IDPL Complex,Old Delhi Jaipur Road,Gurgaon-122016,India
91-124-2455485 Fax: 91-124-2455484
tapati@ibsdel.org, pkgarg@ibsdel.org

Pradeep Kumar

ICFAI Business School
IDPL Complex,Old Delhi Jaipur Road,Gurgaon-122016,India
91-124-2455485 Fax: 91-124-2455484
tapati@ibsdel.org, pkgarg@ibsdel.org

ABSTRACT

Business Intelligence (BI) has emerged as one of software solutions that have maximum allocated investments by many organizations for the year 2005. Among various forms and application-based business intelligence, market intelligence (MI) is viewed as a crucial factor for a company to succeed both operationally and strategically in today's' competitive environment. Capturing market intelligence data has apparently become easy, especially with the proliferation of the Web. But, this has made data collection more difficult in reality from the system's point of view, as data sources on the web are voluminous, heterogeneous in terms of structures and semantics, and some part of it may be irrelevant to a specific organizations' marketing decision-making context, which is the primary premises of market intelligence systems. To address these three specific problems, an algorithm based on similarity measures and multi-dimensional scaling (MDS), which produces hierarchical clusters of knowledge maps from a training data-source set for collecting inputs from heterogeneous sources for capturing market intelligence, is proposed in this paper. The paper illustrates that this algorithm can reduce irrelevant or highly similar data sources for inclusion in the selected data-source repository – represented in the form of clusters of knowledge maps. Therefore, it acts as a similarity-based selection and filtering tool also, with the specific purpose of data collection for MI. Incorporating more advanced techniques for Knowledge maps creation e.g. the Genetic Algorithm-based approaches can further expand this work.

INTRODUCTION

Business Intelligence and Market Intelligence

A definition on Competitive Intelligence says that, business intelligence is a systematic and ethical program for gathering, analyzing, and managing external information that can affect a company's plans, decisions, and operations (Nucleus Report 2005). It is also defined as the result of "acquisition, interpretation, collation, assessment, and exploitation of information" (Future-Group 1997) in the business domain. Business intelligence (BI) was earlier viewed as an exclusive domain of large companies with skilled data analyzers and large data warehouses. Now it is seen as an equal or even more valuable tool for small and midsize companies that are growing fast and experiencing the accompanying headaches. With ERP vendors clamoring for customers, adding value to their platforms will be the carrot that they hope lures in new customers. According to the report of Nucleus, a market research firm on IT, in their research about Top 10 IT predictions for 2005, (Nucleus Report 2005) BI has emerged as the first among the maximum sought-after solutions. BI is being viewed as a solution for wrestling data out of ERP systems with the ERP vendors trying to reclaim lost ground by offering expanded BI capabilities to provide one stop shopping. The result is a shake-up with the current pure play vendors, and some interesting dynamics as BI vendors and their ERP partners learn some new steps in the competition process. Amongst various Business Intelligence elements, Market Intelligence is one of the most significantly and practically applied concept or tool.

There are various challenges in gathering optimal amount of market intelligence data where most of the data sources are external and heterogeneous. The other challenges that remain are: Enterprise executives, while gathering and processing Market Intelligence data, often suffer from Information Overloading. Because of the huge volume of the external information available on market intelligence and the high rate of its growth, there is a great demand from enterprises for automated MI management systems. Many computing devices serve as the repository of the Internet posing challenges of effective knowledge discovery from voluminous information. Since the Internet is one of the top five sources of business information (Future-Group 1997), information overload on the Web is likely to hinder business analysis. In a typical analysis scenario, a business analyst in the database technology field might ask the questions like sources of relevant data, grouping, filtering, clustering data for analytical, visualization and decision-support purposes, handling unstructured data and so on. In this paper, we are addressing these challenging issues by proposing a method for collecting external business intelligence using Knowledge Maps as the Knowledge extraction and description mechanism. In contrast with traditional knowledge portal methods where document-level technologies are quite popular, the proposed design in this paper uses the Knowledge Map method for extraction and collection of Market Intelligence data, based on the concept developed and presented by Chen et al(2001). Consequently, we present the process of extracting market intelligence from voluminous, mixed and heterogeneous sources using knowledge maps, which is generated by an information synthesis process and can provide semantic services through various application interfaces and analytical or filter or enterprise-data search engines. This process, ultimately developed and presented in this paper as an algorithm, can integrate various text collections, apply data-mining and dissemination functions on the collections with a defined process flow, and present a personalized browsing and searching interface. This approach can be facilitated or extended with the use of existing Web mining, clustering, and visualization techniques to support effective exploration of market intelligence.

An algorithm is proposed for collecting market intelligence associated with it's three primary problems like relevance, volume and heterogeneity.

The algorithm in specific, and the process, in general, use:

- Knowledge maps for identifying a relevant source of data → addressing the problem of relevance
- Knowledge maps as a selection clustering tool : not for classification or grouping, but for selecting and filtering the data → addressing the problem of volumes

- and then again knowledge maps for transforming all the relevant and filtered data from various heterogeneous systems to a homogenous platform so that various analytical tools can be applied to the resultant data-set → addressing the problem of heterogeneity

Existing technologies for collecting Market Intelligence

During the past decade, many efforts have been made in this field. Generally, MI research and system development efforts have focused on storage and data mining technologies. Data warehousing and on-line analytical processing (OLAP) have typically been used to solve data extraction, transformation, data cleaning, storage, and mining issues. Previous efforts have used document-based technologies and supported document-level functions such as full text search, document classification, and so on. Business practitioners have developed automated tools to support better understanding and processing of information. In recent years, business intelligence tools have become important for analysis of information on the Web (Fuld et al 2003). Researchers have also developed advanced analysis and visualization techniques to summarize and present vast amount of information.

Generally, business intelligence tools are meant for enabling organizations to understand their internal and external environments through the systematic acquisition, collation, analysis, interpretation, and exploitation of information. Fuld et al. (2003) found that the global interest in intelligence technology has increased significantly during the years of early twenty-first century. Automated search capability in many tools has been shown to lead to information overload. (Future-Group 1997) Despite recent Improvements in analysis capability (Fuld et al 2003), there is still a long way to go to assist qualitative analysis effectively. Most tools that claim to do analysis simply provide different views of collection of information {e.g.. comparison between different products or companies). Due to limited analysis capability, these tools are weak at summarizing a large number of documents collected from the Web, thus handling the problems of relevance, heterogeneity and volume. Although search engines may help, their linear list display of numerous results may lead to information overload.

Information Overload

Several frameworks and techniques have been proposed to deal with information overload and the lack of analytical capability of search engines and business intelligence tools to effectively collect and filter BI data. Traditional result-list display of hypertext belongs to the one-dimensional data type. In contrast, data types such as two-dimensional data, tree data, and network data allow more browsing tasks to be done and support analytical capabilities more effectively. Lin (1997) identified various display formats for handling multi-dimensional data e.g. hierarchical displays- an effective information access tool for browsing, network displays, scatter displays (Spence 2001) for reflecting the underlying structures of data, and map displays- to provide a view of the entire collection of items at a distance (Lin 1997). Shneiderman proposed a task by data type taxonomy to study the types of data and tasks involved in visual displays of textual information (Shneidermiin 1996). Other visualization techniques have been developed to enable better understanding of documents. Regarding document visualization, it primarily concerns the task of getting insight into information obtained from one or more documents without users' have read those documents (Wise et al 1995). Most processes of document visualization involve three stages i.e. document analysis, algorithms, and visualization (Spence 2001). Regarding searching documents on the web, Web mining techniques have been applied to analysis of these unstructured, heterogeneous documents. Web content mining treats a web document as a vector of weights of key terms (Bowman et.al 1994). Web structure mining treats the Web as a graph, where nodes (Web pages) are connected to each other through directed edges (hyperlinks). Researchers have tried to combine Web content mining and Web structure mining to improve the quality of analysis. He et al. (2001) proposed an unsupervised clustering method that was shown to identify relevant topics effectively. The clustering method employed a graph-partitioning method based on a normalized cut criterion. Bharat and Henzinger (1998) augmented a connectivity analysis-based algorithm with content analysis. Meta-searching technique also has been shown to be a highly effective method of resource discovery and collection on the Web. By sending queries to multiple search engines and collating the set of top-ranked results from each search engine, meta-search engines can greatly reduce bias in search results and improve the coverage. Mowshowitz and Kawaguchi (2002) concluded from their study that the only realistic way to counter the adverse effects of search engine bias is to perform meta-searching. Chen et al, (2001) showed that the approach of integrating meta-searching with textual clustering tools achieved high precision in searching the Web.

Algorithms

Algorithms have been used to cluster and project a high-dimensional structure onto a two- or three-dimensional space. For example, cluster algorithms and multidimensional scaling algorithms are frequently used in visualization. Cluster algorithms classify objects into disjoint subsets or partitions based on their semantic dissimilarities. Two categories of cluster algorithms are used in previous research: hierarchical and partitioned MR, Hierarchical clustering is a procedure for transforming a proximity matrix into a sequence of nested partitions. Partitioned clustering assigns objects into groups such that objects in a cluster are more similar to each other than to objects in different clusters. Typically, a clustering criterion is adopted to guide the search for optimal grouping. A graph-theoretic criterion, called normalized cut, treats clustering as graph partitioning and computes the normalized cost of cutting a graph. Using this criterion in image segmentation (Shi and Malik 2000) and Web page clustering (He et al 2001) has been shown to achieve high performance. Although partitioned clustering tries to achieve optimal results, it is usually difficult to evaluate all partitions because the number of possible partitions is extremely large. Therefore, heuristics are needed to find good values to the criterion selected. Examples of such heuristics include genetic algorithms, taboo search, scatter search, and simulated annealing (He et al 2002).

Multidimensional Scaling

Multidimensional scaling (MDS) algorithms consist of a family of techniques that portray a data structure in a spatial fashion, where the coordinates of data points x_{ia} are calculated by a dimensionality reduction procedure (Torgerson 1952). The distances (d_{ij}) among data sources can be calculated as follows

$$d_{ij} = [\sum \{x_{ia} - x_{ja}\}^{p}]^{1/p} \quad (p >= 1),\ x_{ia} <> x_{ja}$$

p is referred to as the Minkowski exponent and may take any value not less than 1. r is the coordinate of point on dimension a, and J is an r-element row vector from the i^{th} row of a/i-by-r matrix containing all *n* points on all r dimensions. The MDS procedure constructs a geometric representation of the data (such as a similarity matrix), usually in a Euclidean space of low dimensionality (i.e.. *p = 2).* MDS has been applied in many different domains. Kealy (2001) applied MDS to studying changes in knowledge maps of groups over time to determine the influence of a computer-based collaborative learning environment on conceptual understanding.

Although much has been done to visualize relationships of objects in different domains using MDS, no attempts to apply it to discovering business intelligence/ market intelligence is prevalently seen. In addition, no existing search engine uses MDS to facilitate Web mining.

Knowledge Maps for collecting Market Intelligence

Apart from the document level operations, an effective Market Inelegance collection system should combine extraction technology with semantic web activities, and should generate a semantic network structure to store knowledge. In this section, we present these requirements of an effective market intelligence collection system as shown in Figure 1, which depicts the problems addressed in this paper, namely relevance, volume and heterogeneity of information.

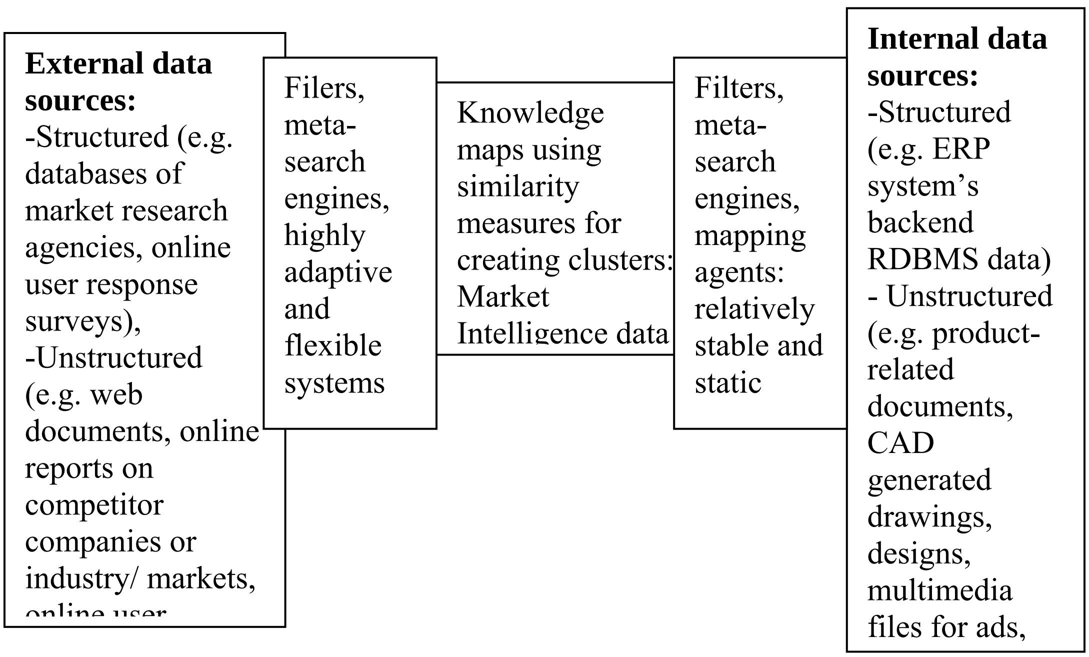

Figure 1: Market Intelligence data collection system using Knowledge Maps

Collection of Data Sources

From Figure 1, it can be seen that there are two major data sources: internal and external. Both these major sources have mixed type of data elements in them i.e. structured (e.g. from RDBMS, data warehouses, ERP backend databases, MIS, spreadsheets etc,) or unstructured (e.g. text, hypertext, multimedia, binary files and so on). Even though both of these major sources have structured and unstructured elements in them, handling internal data sources is relatively easy because the forms in which they exist is known to the organization. Primary problem therefore is to deal with external data sources that exist in various forms unknown to the organization and in various degrees of unstructured-ness. Techniques like meta-searching and automatic parsing and indexing are commonly used for such data collection problems.

Meta-Searching

Major search engines on the web, which is the primary source of external data for MI in an organization are: AltaVista, Google, Infoseek, Lycos, Dogpile, AlltheWeb, Yaboo, MSN, LookSmart, Teoma, Kartoo, Wisenut etc. Kartoo among these engines is a new meta-search engine that presents results in a map format.

Automatic Parsing and Indexing

Since the Web contains documents in various forms like textual content and HTML tag-based hypertext information etc., parsing is necessary to facilitate further analysis, which can automatically extract key words and hyperlinks from the Web data sources. The word-type information is to be used in the co-occurrence analysis. Each key word or noun phrase for example can be treated as subject descriptor type. Based on a revised automatic indexing technique (Bowman et.al 1994), the term's level of importance can be measured by term frequency and inverse data-source frequency. Term frequency reflects how often a particular term occurs in a document. Inverse data-source frequency can indicate the specificity of the term and allows terms to acquire different strengths or levels of importance based on their specificity. A term could be a one-, two-, or three-word phrase.

Co-occurrence Analysis

Co-occurrence analysis can convert data indices and weights obtained from inputs of parameters and various data sources into a matrix that shows the similarity between every pair of such sources. The similarity between every pair of data sources contains its content and structural (connectivity) information. He et al. (2001) designed an algorithm for computing the similarity between every pair of Web documents by a combination of hyperlink structure, textual information, and co-citation. This algorithm has been used in this paper to compute the similarity between data sources, as follows:

Similarity between data source i and data source j is

$$W_{ij} = \alpha \{A_{ij} / |A|_2\} + \beta S_{ij} / |S|_2 + (1- \alpha - \beta) C_{ij} / |C|_2$$

$$0< \alpha, \beta <1, 0 <= \alpha + \beta <=1,$$

where A, S, and C are matrices for A_{ij}, S_{ij}, and C_{ij} respectively. Values for A_{ij} will be 1 if data source i has a direct link to data source j, else 0. S is the asymmetric similarity score between data sources i and j, and is calculated as follows:

$$S_{ij} = \text{sim}(D_i, D_j) = \left[\left[\sum_{k=1}^{p} d_{ki}\, d_{kj}\right] / \left[\sum_{k=1}^{n} d^2_{di}\right]\right] \text{ X } S_{ji} = \text{sim}(D_j, D_i)$$

where

1. n is total number of terms in D_i, m is total number of terms in D_j, p is total number of terms that appear in both D_i, and D_j.
2. d_{ij} = (Number of occurrence of term j in data source i) X log((N/d_{fj})) X w_j) X (Term type factor)
3. d_{fj} is number of data sources containing term j
4. w_j is number of words in term j
5. Term type factor = 1 + ((10-2 X $type_j$ / 10), where $type_j$ = minm 1 if term j appears in title, 2 if it appears in heading, 3 if it appears in context text etc.)

6. C_{ij} is number of data sources pointing to both source i and source j (cocitation matrix).

CREATING THE KNOWLEDGE MAPS

The data sources for Market Intelligence, be it structured or unstructured i.e. text/ binary objects/ documents, can be represented in the form of a graph consisting of nodes as the data sources and edges as the similarities between data sources. Using hierarchical and partitioned clusters simultaneously, a hierarchy of similarity clusters of data sources based on their parameters or properties, can be created in the training phase. Then these clusters can be transformed into two-dimensional knowledge maps using MDS. Let us consider an example where we have n data sources as training data set for training the selection clusters. These training data sets will be used to create a hierarchical graph of clusters transformed into knowledge maps, as shown in Figure 2 below. Partitioning of a graph, say G, can be done in various ways, for example, by using similarity measures as below:

Normalized Cut on graph G = {cut between (A, B)/ assoc(A, V)} + {cut between (A, B)/ assoc (B,V)}

where, Cut between (A,B) = $\sum_{i \in A,\, j \in B} W_{ij}$, W_{ij} is similarity between nodes i and j of the graph. A cut on a graph G = (V, E) is defined as removal of a set of edges such that the graph is split into disconnected sub-graphs, thereby can be converted into a hierarchy of knowledge map.

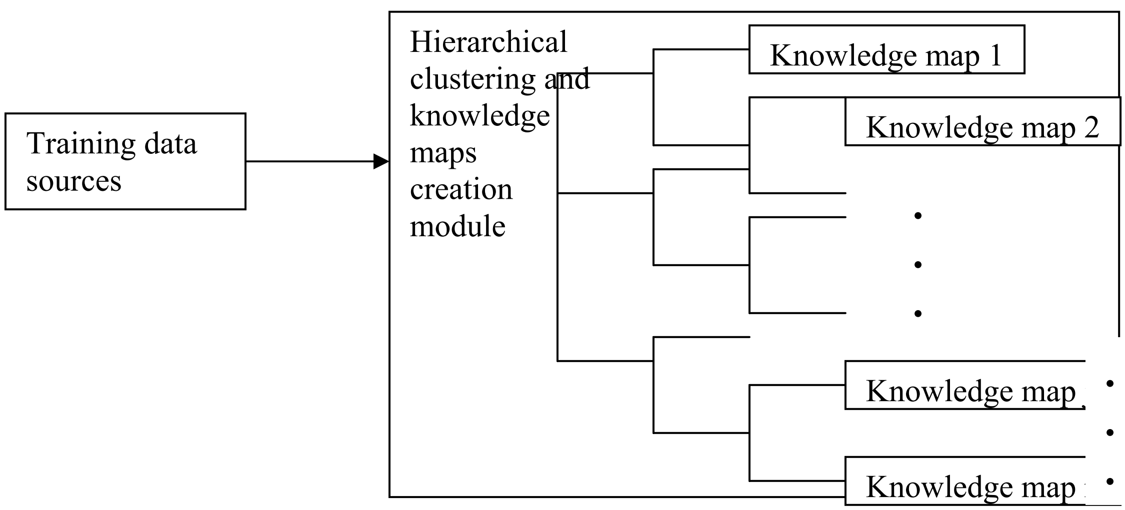

Figure 2: Example of a Hierarchical Graph of Knowledge Maps

Knowledge maps can be created in a simple and less resource-hungry process by using simple MDS. Multi-dimensional scaling or MDS, as explained above, can be used to transform a high-dimension similarity matrix into a two-dimensional representation of points and for displaying them on a map. Torgerson's classical MDS procedure, which work well with non-Euclidean distance matrixes(Young 1987) by giving approximation of the coordinates and does not require iterative improvement (Torgerson 1952), can be used here for it's simplicity and ease of implementation. The MDS procedure can be implemented using the following steps. First, Similarity matrix is to be converted into a dissimilarity matrix by subtracting each element by the maximum value in the original matrix. This matrix can be called as dissimilarity matrix D. Then matrix B which is a scalar product is to be calculated, by using the cosine law. Each element in B is given by:

$$b_{ij} = -1/2\,[\,d_{ij}^2 - 1/n \sum_{k=1}^{n} d_{ik}^2 - 1/n \sum_{k=1}^{n} d_{kj}^2 + 1/n^2 \sum_{g=1}^{n} \sum_{h=1}^{n} d_{gh}^2\,]$$

where d_{ij} is an element in D, n= number of nodes in the data-source graph

After calculating B, singular value decomposition is performed using the formula as below:

$B= UxVxU'$, $X = U\ X\ V^{1/2}$

where U has eigenvectors in its columns and V has eigenvectors on its diagonal.

Therefore, $B = X \times X'$.

The first two column vectors of X thus calculated now can be used to obtain the two-dimensional coordinates of points, which can be used to place the data sources onto knowledge maps.

USING KNOWLEDGE MAPS FOR CREATING CLUSTERS OF COLLECTED DATA

Creation of knowledge maps from a graphical representation of various data sources, based on their similarities or, more specifically and logically their degree of dissimilarities, is shown earlier. Basically, by segregating the graph representing the data sources, we get a hierarchical cluster of various knowledge maps where these knowledge maps can be seen as representing similar data sources. This is what is to be done in the training phase of the clustering and knowledge maps creation module, as explained in Figure 2 in terms of training the modules with the data sources and Figure 3 in terms of the various data sources themselves in the context of Market Intelligence requirements of an organization.

After the use of training data sources a set of hierarchical clusters with knowledge maps have been created and thereafter the run-time data collection has to start. During the run-time, two events can take place. Say, a data source i is being input to the module as shown in Figure 4 below. Now the similarity score of this data source will be calculated by the appropriate sub-module in respect to the existing Knowledge Maps that are already trained into the module. The thershold value of this score will have to be given by the user. It can be given one-time, or it can be exectuion envrionment/ run-time specific depending on the degree of filerting/ reduction requirements.

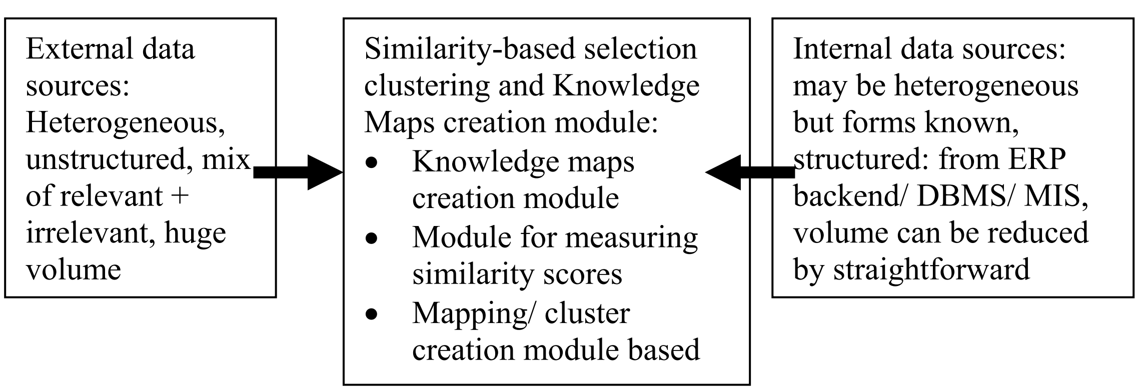

Figure 3: Data Sources as Inputs to the Clustering and Knowledge Maps Creation Module

If the similarity score of data source i is found to be closer(i.e. lesser that the thershold value given) to any of the existing knowldeg maps in the hierarchical cluster, then it is included in that knowledge map. Here a possibility is that the closely-matching knowledge map can be futher fine-tuned with the data-source i input's parameters and properties. In that case, it will be like a fine-tuning training pahse going in tandem with the run-time phase. But here the data source colelciton will be more enriched , more representative and inclusive.

The other possibility is that the data source i does not have a close proximity to any of the existing knowledge maps in terms of it's similarity value and the threshold(i.e. the similarity score in terms of all existing knolwdge amps is more than the threshodl value). In such situation, a new knowledge amp has to be created and put in the approrpasite place in the hierachy of knopwldge map clusters.

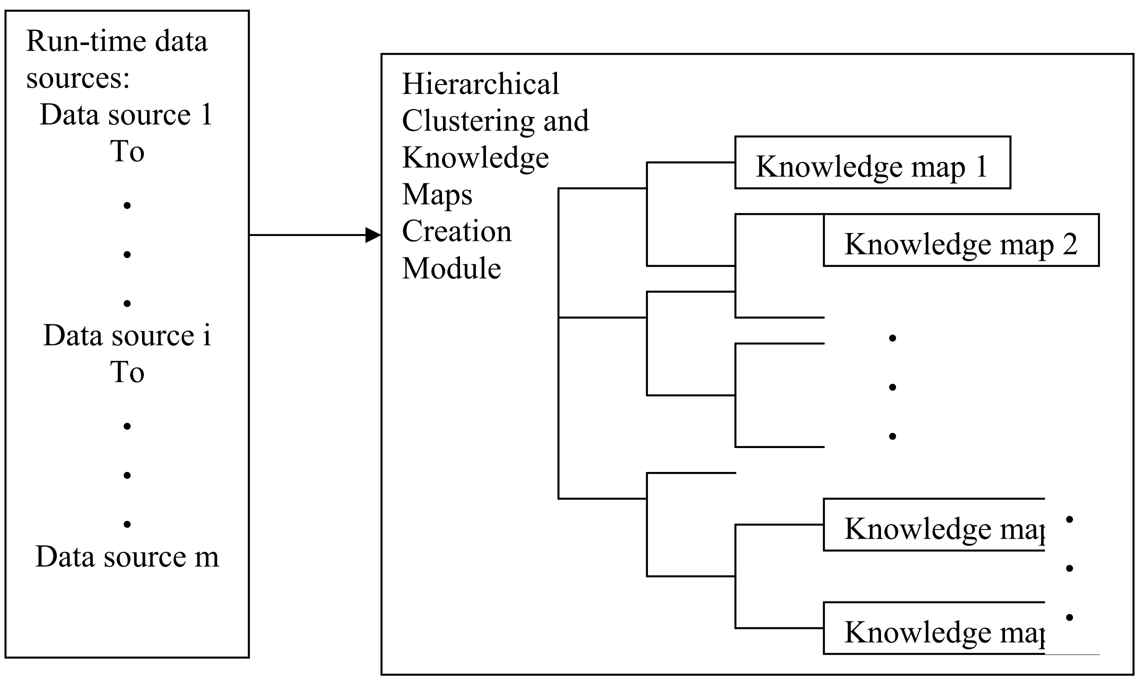

Figure 4: Run-time Collection of Data Sources

As explained above, the steps in Figure 4 can be explained as given below:

Steps:

1. One data source i arrives for feed into the module which has sub-modules like Knowledge Map creation module and similarity scoring module
2. Similarity scoring module measures the similarity score of the data source (for i = 1, S_{ij} = 0, during the training phase) and a KM for data source i, say KM_i is created.

3. The similarity score is calculated and compared against all existing KMs, i.e. KM_1 ti KM_n
4. If the similarity score is < threshold value given for data reduction for any existing KM say KM_j, then the KM_i gets mapped or included into KM_j and KM_j learns for similarity patterns from KM_i and refines itself.
5. if similarity score is > threshold value, KM_i creates another cluster of it's own.
6. Go to step 1.

Using this algorithm, the primary three problems that were introduced in the previous sections, gets addressed.

1. First, by using training data sources, the trained Knowledge Map clusters have the patterns identified only for relevant data which has been included in the training data. So the problem of relevance i.e. eliminating/ reducing irrelevant data collection is achieved to a limited scope depending on the choice and exhaustibility of the training data source-sets.
2. Second, the problem of volume is addressed by using Knowledge maps and similarity-based clustering where similar data sources are not repeatedly included in the collected repository of KM-represented data.
3. Third, the problem of heterogeneity is addressed as all the heterogeneous structured or unstructured data sources are finally being represented in the form of Knowledge Maps, which can then be used as a homogenous input to the analytical modules of the MI systems.

CONCLUSION

This proposed algorithm has been shown to handle the three primary problems of data collection for market intelligence in an organization. Further extensions may include exploring various other knowledge map creation mechanisms including the Genetic Algorithm approaches and extrapolating the Knowledge maps into the analytical systems required for analyzing and visualizing the Market intelligence data.

REFERENCES

Bharat, K., Henzinger. M.R.(1998), “Improved algorithms for topic distillation in hyperlinked environments”, Proceedings of the Twenty-First International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press: 104-111.

Chen, H.; Fan. H.; Chau. M.; and Zeng, D.(2001) “Meta Spider: Meta searching and categorization on the Web”, *Journal of the American Society for Information Science and Technology. 52,* 13: 1134-1147.

Fuld, L.M.: Singh. A.: Rothwell. K.; and Kim, J. (2003) *Intelligence Software Report™ 2003: Leveraging the Web.* Cambridge. MA: Fuld & Company.

Futures-Group Ostriches & Eagles.(1997) The Futures Group Articles, Washington, DC, (available at www.futuresgroup.com).

He. X.; Ding. C; Zha. H.; and Simon, H.(2001)” Automatic topic identification using Webpage clustering”, In X. Wu. N. Cercone, TY. Lin, J- Gehrke. C. Clifton. R. Kotagiri. N. Zhong. and X. Hu (eds,). *Proceedings of the 2001 IEEE International Conference on Data Mining.* Los Alamitos. CA: IEEE Computer Society Press. 2(X)I. : 195-202.

He, Y, and Hui. S.C.(2002) “Mining a Web citation database for author co-citation analysis”, *Information Processing and Management. 38.* 4: 491-508.

Kealy, W.A.(2001) “Knowledge maps and their use in computer-based collaborative learning”, *Journal of Educational Computing Research, 25.* 4: 325-349.

Lin, X. (1997) “Map displays for information retrieval”,*Journal of the American Society for Information Science. 4H.* 1: 40-54.

Mowshowitz. A., and Kawaguchi. A.(2002) “By Online Web” ,*Communications of the ACM, 45.* 9: 56-60.

Nucleus Report (2005) *Top 10 IT Spending for 2005: Survey of CIOs in MNCs: Survey Report March 2005 by Nucleus Research*, http://www.nucleus.com/surveys/2005

Shi. J., and Malik. J.(2000) “Normalized cuts and image segmentation”, *IEEE Transactions on Pattern Analysis and Machine Intelligence. 22.* S (2(X)0):, 8S8-905.

Shneidermiin, B.(1996) “The eyes have it: A task by data type taxonomy for information visualizations”, *Proceedings of IEEE Symposium on Visual languages.* Los Alamitos, CA: IEEE Computer Society Press : 336-343.

Spence, R.(2001), *Information Visualization.* New York: ACM Press, 2001.

Torgerson, W.S. (1952)”Multidimensional Scaling: Theory and method”, *P.sychometrika. 17.4*, 401.19..

Wise, J.A.; Thoma, J.J.: Pennock. K.; Lantrip, *D.:* Pottier, M.; Schur, A.; and Crow, V. (1995)“Visualizing the non-visual: Spatial analysis and interaction with information from text documents”,*Proceedings of IEEE Symposium on Infonnation Visualization.* Los Alamitos, CA: IEEE Computer Society Press: 51-58.

Young, F.W. (1987),*Multidimensional Scaling: History, Theory, and Applications,* ed. R.M. Hamer. Hillsdale, NJ: Lawrence Erlbaum.