An Algorithm for Market Intelligence Data Collection from Heterogeneous Sources with Similarity-Based Selection Clustering Technique Using Knowledge Maps

Tapati Bandopadhyay, ICFAI Business School, Bangalore, India; E-mail: chatterjee_tee@yahoo.com
Pradeep Kumar, ICFAI Business School, Gurgaon, India; E-mail: pkgarg@ibsdel.org
Arjya Chakravarty, ICFAI Business School, Gurgaon, India; E-mail: arjya@ibsdel.org
Anil Kumar Saini, USMS, GGSIP University, New Delhi, India; E-mail: aksaini@rediffmail.com

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.

Keywords: Market Intelligence, Business Intelligence, Multi-dimensional Scaling

1. INTRODUCTION

1.1 Business Intelligence and Market Intelligence

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. Amongst various Business Intelligence elements, Market Intelligence is one of the most significantly and practically applied concept or tool. (Future-Group 1997 and subsequent reports). However, the volume and heterogeneity of information available in the internal and more prevalently external domains (e.g. the Internet), create a huge information overload. (Bowman et.al 1994). In this paper, an algorithm is proposed for collecting market intelligence associated with its 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

1.2 Existing Technologies for Collecting Market Intelligence

In recent years, business intelligence tools have become important for analysis of information on the Web (Fuld et al 2003). Despite recent Improvements in analysis capability there is still a long way to go to assist qualitative analysis effectively. Due to limited analysis capability, existing tools are weak at summarizing a large number of documents collected from the Web, thus handling the problems of relevance, heterogeneity and volume.

1.3 Information Overload- handling techniques

Few algorithms proposed in text/web mining and document clustering find relevant applications here. Lin (1997) identified various display formats for handling multi-dimensional data e.g. scatter displays (Spence 2001) 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 (TTT) to study the types of data and tasks involved in visual displays of textual information (Shneidermiin 1996), (Wise et al 1995). Most processes of document visualization involve three stages i.e. document analysis, algorithms, and visualization (Spence 2001). He et al. (2001) proposed an unsupervised clustering method that was shown to identify relevant topics effectively. Bharat and Henzinger (1998) augmented a connectivity analysis-based algorithm with content analysis.

1.4 Algorithms

Partitioned clustering, in context of unstructured documents, 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. 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. But, heuristics are needed to find good values to the criterion selected.

1.5 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).*

2. KNOWLEDGE MAPS FOR COLLECTING MARKET INTELLIGENCE

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.

2.1 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). 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.

For example, the word-type information can 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.

2.2 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:

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

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).

3. 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

Figure 1. Market intelligence data collection system using knowledge maps

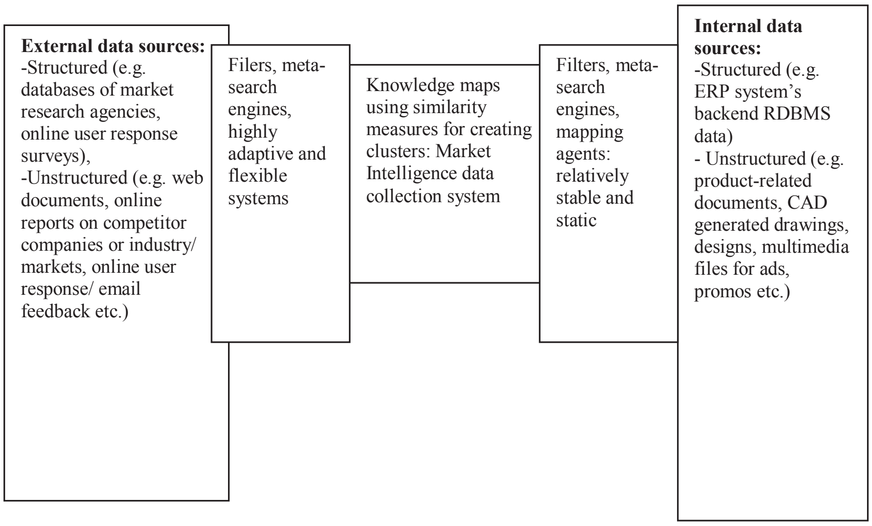
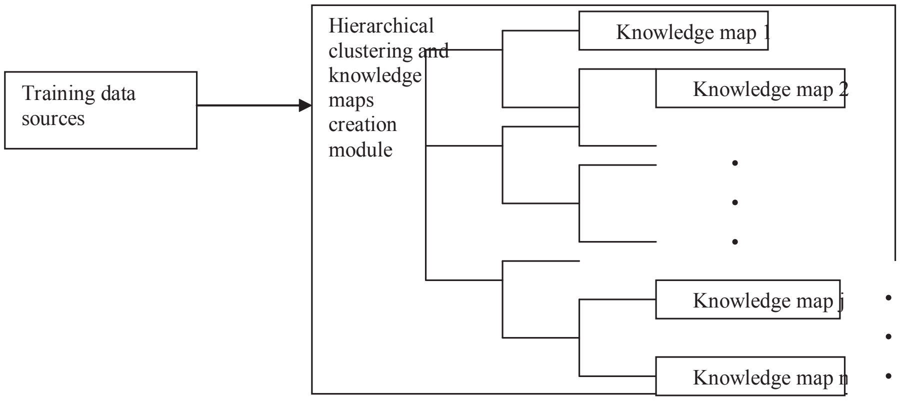

External data sources:
-Structured (e.g. databases of market research agencies, online user response surveys),
-Unstructured (e.g. web documents, online reports on competitor companies or industry/ markets, online user response/ email feedback etc.)
Filers, meta-search engines, highly adaptive and flexible systems
Knowledge maps using similarity measures for creating clusters: Market Intelligence data collection system
Filters, meta-search engines, mapping agents: relatively stable and static
Internal data sources:
-Structured (e.g. ERP system's backend RDBMS data)
- Unstructured (e.g. product-related documents, CAD generated drawings, designs, multimedia files for ads, promos etc.)

Figure 2. Example of a hierarchical graph of knowledge maps



the training phase. Then these clusters can be transformed into two-dimensional knowledge maps using MDS.

Example Run of the Proposed Algorithm:
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.

Torgerson's classical MDS procedure, (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 D by subtracting each element by the maximum value in the original matrix. 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 x 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.

4. 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

Figure 3. Data sources as inputs to the clustering and knowledge maps creation module

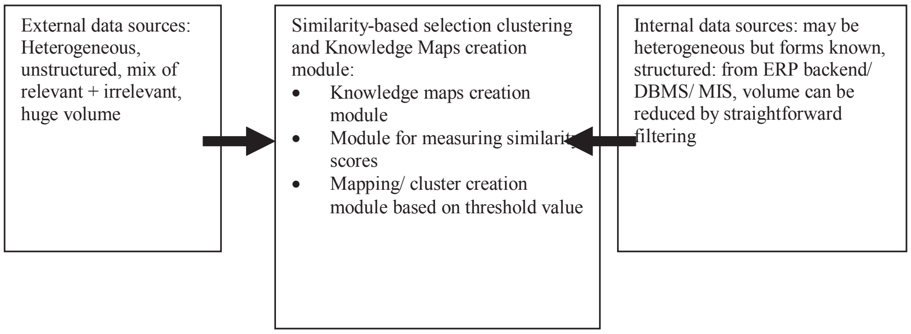
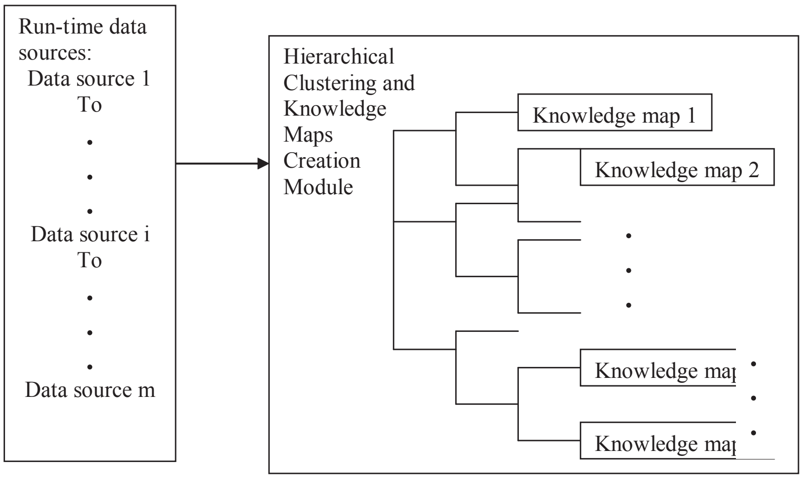

External data sources: Heterogeneous, unstructured, mix of relevant + irrelevant, huge volume
Similarity-based selection clustering and Knowledge Maps creation module:
• Knowledge maps creation module
• Module for measuring similarity scores
• Mapping/ cluster creation module based on threshold value
Internal data sources: may be heterogeneous but forms known, structured: from ERP backend/ DBMS/ MIS, volume can be reduced by straightforward filtering

Figure 4. Run-time collection of data sources



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. 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 phase 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 maps is more than the threshold value). In such situation, a new knowledge map has to be created and put in the approrpasite place in the hierachy of knopwldge map clusters.

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.

Bowman, C.M., Danzig. P.B., Manber. U.; Schwartz, F.(1994) "Scalable Internet resource discovery: Research problems and approaches", *Communication of the ACM. Vol* 8 : 98-107.

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.

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

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.

Related Content