The HCARD Model using an Agent for Knowledge Discovery

Bobby D. Gerardo[a], Jae-Wan Lee[a], and Su-Chong Joo[b]

[a] *School of Electronic and Information Engineering, Kunsan National University*
68 Miryong-dong, Kunsan, Chonbuk 573-701, Korea
Tel: +82-63-4694696, Fax: +82-63-4694699, Email: {bgerardo, jwlee}@kunsan.ac.kr

[b] *School of Electrical Electronic and Information Engineering, Wonkwang University*
344-2 Shinyong-dong, Iksan, Chonbuk 570-749, Korea
Email: scjoo@wonkwang.ac.kr

Abstract

In this study, we will employ a multi-agent for the search and extraction of data in a distributed environment. We will use an Integrator Agent in the proposed model on the Hierarchical Clustering and Association Rule Discovery (HCARD). The HCARD will address the inadequacy of other data mining tools in processing performance and efficiency when use for knowledge discovery. The Integrator Agent was developed based on CORBA architecture for search and extraction of data from heterogeneous servers in the distributed environment. Our experiment shows that the HCARD generated essential association rules which can be practically explained for decision making purposes. Shorter processing time had been noted in computing for clusters using the HCARD and implying ideal processing period than computing the rules without HCARD.

Keywords:

Data mining, clustering, association rule, agent

1. Introduction

The problem of finding information on the Web or distributed networks is well known and the Web is the largest and most used distributed hypermedia systems known to date, and certainly, in terms of users. The fact that information systems that both distributed and heterogeneous in nature continue to become available on the Web complicates the problem even more [1]. As a result, attempting to locate, integrate and organize related information has become a major challenge. The "information overload" is becoming commonplace and many research activities are being done in an attempt to curb the Web and the information contained within it. The role of agents for distributed information management, which included resource discovery, information integrity and navigation assistance is perceived to be important.

On another perspective, data mining had been viewed or identified as valuable and essential steps towards knowledge discovery because it could be applied to extraction or mining of knowledge from vast amount of data. Consequently, various people treat data mining as Knowledge Discovery in Databases, or KDD. The data mining and discovery of such information often yields important insights into business and its client may lead to unlocking hidden potentials by devising innovative strategies.

Several data mining tools had been developed and innovated to address major applications in the academic, business or industrial purposes. Some examples of these tools are used for concept description, association analysis, classification, prediction, and cluster analysis. The present trends show that vendors of data management software are becoming aware of the need for integration of data mining capabilities into database engines, and some companies are already allowing for integration of database and the data mining software.

Clustering can be utilized to group data into clusters so that the degree of association is strong between members of the same cluster and weak between members of different clusters [2]. Thus, each cluster describes the class to which its members belong. For that reason, cluster analysis can reveal similarities in data which may have been otherwise impossible to find.

Some other constraints that most researchers observed in the data mining tasks were computing speed, reliability of the approach for computation, heterogeneity of database, and vast amount of data to compute [3], [4], [5]. Most of the time these are limitations that defeat typical and

popular mining approach. This study investigates the formulation of the cluster analysis technique as integrated component of the proposed model on Hierarchical Clustering and Association Rule Discovery (HCARD) in order to partition the original data prior to implementation of other data mining tools. The model that we proposed uses the hierarchical nearest neighbor clustering method and apriori algorithm for knowledge discovery purposes to be implemented on business or transactional databases. In addition, we will develop a multi-agents based on CORBA architecture for data search and extraction in the distributed environment. Two agents such as the user interface agent and the facilitator agent will be developed for interface, search, extraction, integration and management.

2. Preliminary concepts

Data mining uses various data analysis tools in order to discover patterns and relationships in dataset that can be used to establish association rules and make effective predictions. Among the essential mechanisms of data mining is association rule discovery rendered on from simple database repositories to complex database in a distributed system. Association rule mining tasks are finding frequent patterns, associations, or causal structures among sets of items or objects in transactional databases, relational databases, and other information repositories.

2.1. The CORBA Architecture

The Common Object Request Broker Architecture (CORBA) was developed by an industry consortium known as the Object Management Group (OMG), an architecture that enables pieces of programs, called objects, to communicate with one another regardless of what programming language they were written in or what operating system they are running on [6], [7].

Because of the easy way that CORBA integrates machines from so many vendors, with sizes ranging from mainframes through minis and desktops to hand-held and embedded systems, it is the middleware of choice for large (and even not-so-large) enterprises. One of its most important, as well most frequent, uses is in servers that must handle large number of clients, at high hit rates, with high reliability [7]. Because of CORBA's popularity, it has several implementations. The most widely used are being IBM's SOM (System Object Model) and DSOM (Distributed System Object Model) architectures. CORBA has also been adopted by Netscape as part of its ONE (Open Network Environment) platform. Two competing models are Microsoft's COM (Common Object Model) and DCOM (Distributed Common Object Model) and Sun Microsystems' RMI (Remote Method Invocation).

The multi-agent system can be used for the construction of open, distributed, heterogeneous and flexible architectures, capable of offering services for collective works. A multi-agent system is composed of a group of agents that are autonomous or semiautonomous and which interact or work together, to perform some tasks or achieve some goals. In addition, the agents in such systems may either be homogeneous or heterogeneous and they may have common goals or goals that are distinct with each other [1]. We proposed a multi-agent system based on CORBA architecture, which will be intended for interface, search, extraction, and management. These agents will be used in conjunction with the proposed HCARD model.

2.2. Cluster Analysis

The goal of cluster analysis is categorization of attributes like consumer products, objects or events into clusters or groups, so that the degree of correlation is strong between members of the same cluster and weak between members of different clusters [8], [9]. Each group describes the class in terms of the data collected to which its members belong. Cluster analysis is also used as a tool for knowledge discovery. It may show structure and associations in data, although not previously evident, but are sensible and useful once discovered. The results of cluster analysis may contribute to the definition of a formal classification scheme, such as in taxonomy for related animals, insects or plants; suggest statistical models with which to describe populations; indicate rules for assigning new cases to classes for identification and diagnostic purposes; provide measures of definition, size and change in what previously were only broad concepts; or find patterns to represent classes [8].

2.3. Data Mining

Numerous data mining algorithms have been introduced that can perform summarization, classification, deviation detection, and other forms of data characterization and interpretation. There are varieties of data mining algorithms that have been recently developed to facilitate the processing and interpretation of large databases. One example is the association rule algorithm, which discovers correlations between items in transactional databases. Apriori algorithm is used to find candidate patterns and those candidates that receive sufficient support from the database are considered for transformation into a rule. This type of algorithm works well for complete data with discrete values. Some limitations of association rule algorithms, such as the Apriori is that only database entries that exactly match the candidate patterns may contribute to the support of that candidate pattern. Other research goals are to develop association rule algorithms that accept

partial support from data.

3. Architecture of the HCARD Model

In this study, the researchers developed the proposed architecture for the data mining system which we named HCARD model as shown in Figure 1 and will be presented in refined view in the subsequent sections.

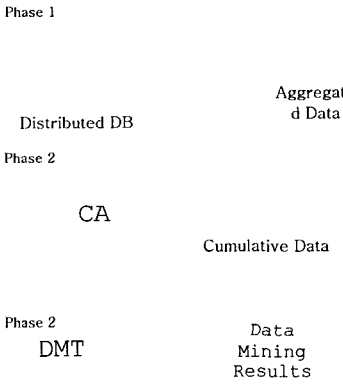


Figure 1 - The Proposed HCARD Model in a distributed environment

Figure 1 shows the proposed three phase implementation architecture for the HCARD model. The first phase includes data extraction, transformation, loading and refreshing. This will result to an aggregated data as illustrated in the same figure. Phase 2 shows the implementation of the hierarchical clustering algorithm (CA) while Phase 3 is the implementation of data mining technique (DMT) for knowledge discovery and analysis.

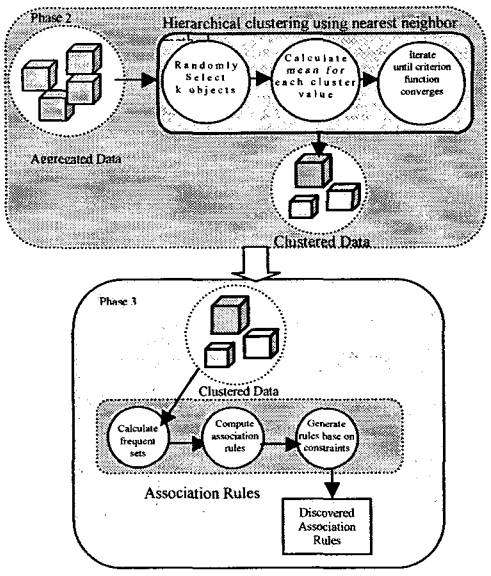


Figure 2 - Refinements of the HCARD at Phase 2 and 3

Figure 2 shows the implementation of the cluster analysis using the hierarchical nearest neighbor clustering algorithm and the implementation of association rule discovery method. It also shows the refined view of Phase 3. In this illustration, association rule algorithm is used as part of the data mining process. The successions of transforms for association rule algorithm which are represented by bubbles are shown in the shaded rectangle.

Multi-agent will be used for data search and extraction in a distributed environment. We proposed a Facilitator Agent (FA) to attempt to integrate some of the information that can assist users in locating information they need either within their local information systems or on the Web.

In our study, the term Facilitator Agent will also refer to the Integrator Agent (IA). Mainly, it supports routing of information, discovery of information, delayed and persistent notification of messages and communication management. The detailed view of the distributed architecture is shown in Figure 3 while the architecture showing the interface, facilitator, and user interface agents is presented in Figure 4.

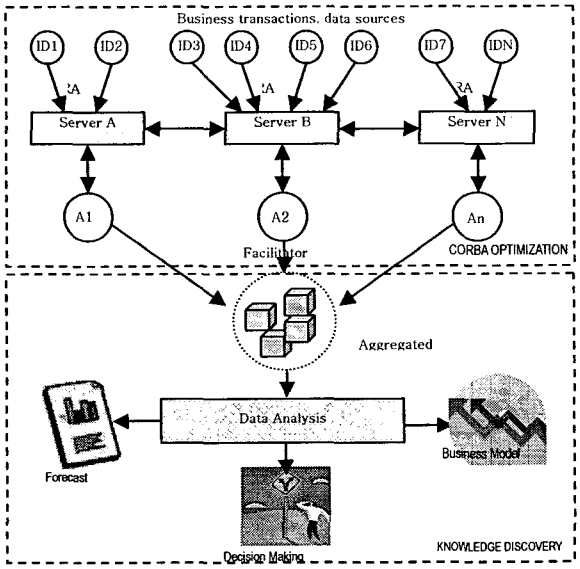


Figure 3 - Detailed view of the distributed architecture

The functions of User Interface Agents (UIA) are to (1) provide the user with an interface whereby he/she can make use of the data/services available in the system through transparent communication with agents; (2) keeps information entered by users or data gathering devices; (3) provides other agents within the system with information about the user and his/her activities which can be used by other agents to achieve a number of tasks; and (4) in cases

where a user enters a query, the agent is responsible for integrating the responses returned from various agents and presenting them to the user.

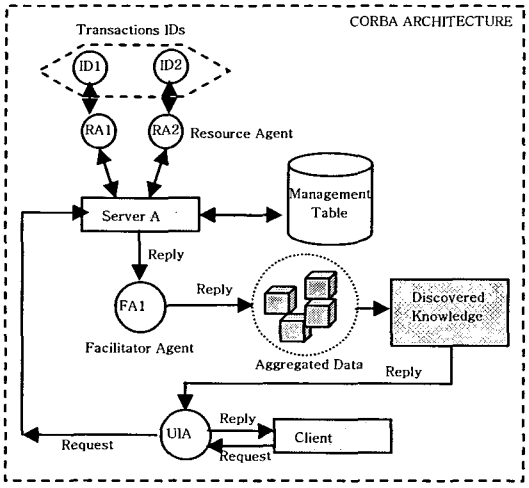


Figure 4 - Multi-agents system architecture

4. The Clustering and the HCARD Model

4.1. Nearest-Neighbor clustering

One of the simplest agglomerative hierarchical clustering methods is single linkage which is also known as the nearest neighbor technique. The defining feature of the method is that distance between groups is defined as the distance between the closest pair of objects, where only pairs consisting of one object from each group are considered. Supposed that we have two points given by, $X = (x_1, x_2, x_3, ..., x_n)$ and $Y = (y_1, y_2, y_3, ..., y_n)$, then the distance between the two points can be computed using equation 1 below.

$$d(X,Y) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} \qquad (1)$$

In another case, the mean nearest neighbor distance can be calculated using equation 2. Here, N is the number of points and d_i is the nearest neighbor distance for point i.

$$\overline{d} = \left(\sum_{i=1}^{N} d_i \right) / N \qquad (2)$$

An agglomerative hierarchical clustering procedure produces a series of partitions of the data, *Cn, Cn-1* until *C1*. The first *Cn* consists of n single object clusters, while the last *C1* consists of single group containing all *n* cases. At each particular stage the method joins together the two clusters which are closest together or are most similar. Figure 5 shows the algorithm for the nearest neighbor clustering.

Our evaluation using the HCARD model will involve calculation for the clustered attributes at a separate phase as indicated in the system architecture in section 3. Such output shall then be processed by implementing the extended steps in data mining tool which uses the association discovery process.

```
Given:
     A set X of objects {x1, ..., xn},  A
distance function dis(c1, c2)
    1.  for i = 1 to n
            ci = {xi}
        end for
    2.  C = {c1, ..., cb}
    3.  l = n+1
    4.  while C.size > 1 do
          a) (cmin1, cmin2) = minimum dis(ci, cj)
        for all ci, cj in C
          b) remove cmin1 and cmin2 from C
          c) add {cmin1, cmin2} to C
          d) l = l + 1
        end while
```

Figure 5 - The nearest neighbor clustering algorithm

4.2. The HCARD for association rule discovery

In our study we used the Apriori algorithm for the association discovery. This algorithm has an important property called Apriori property which is used to improve the efficiency of the level-wise generation of frequent itemsets. There are two steps in the implementation of Apriori property, namely the join step which will find *Lk*, a set of candidate k-itemsets by joining *Lk-1* with itself. The next step is the prune step in which *Ck* is generated as a superset of *Lk*, that is, its members may or may not be frequent, but all of the frequent k-itemsets are included in *Ck*. The Apriori property implies that any *(k-1)*-itemset that is not frequent cannot be a subset of a frequent k-itemset; hence, the candidate can be removed.

5. Experimental Evaluations

The experiment was performed on the database containing 30 attributes comprising of six (6) major dimensions and a total of 1,000 tuples of e-commerce and transactional types of data. The evaluation platforms utilized in the study were IBM compatible computers, Windows XP, Sun Solaris 9, Borland Visibroker, Java, and Python.

The abbreviated notations for the attributes stand as follows: An= books and its corresponding subcategories, Bn = Electronics, Cn = Entertainment, Dn= Gifts, En = Foods, and Fn = Health. Furthermore, Book attribute is consist of subcategories like A1= Science, A2=social, A3=math, A4=computer, A5=technology, A6=religion, and

A7=children books. Other dimensions are written with notations similar to that of An.

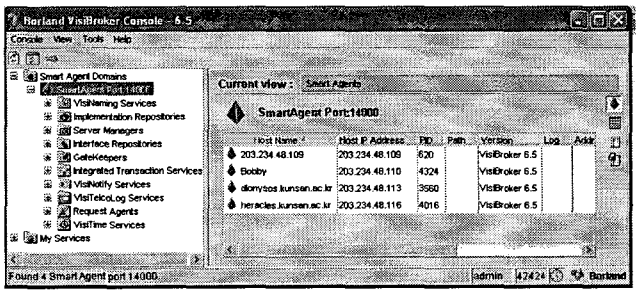
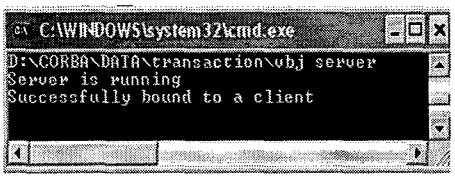


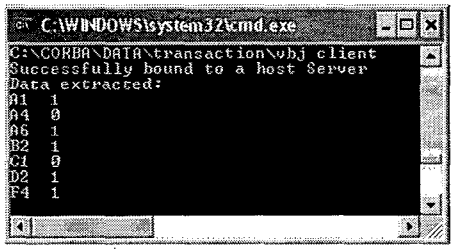
Figure 6 - Smart agent console

The smart agent console is presented in Figure 6. In this illustration, four host computers were considered for experiment purposes. Search and extraction of data were done through the active agents running on port 14000. The Sun Solaris Server is indicated by the PID 620 while the PID 4324, 3560, and 4016, respectively, are hosts running on Windows XP operating systems.

Figure 7 shows the CORBA implementation to extract information from a host computer. Figure 7(a) shows that the server (host) is active and is ready for communications with clients. The data obtained are in the form of online transactions. The notations A1, A4, A6, B2, and etc. in Figure 7(b) correspond to the items (attributes) that had been extracted from the host computer.



(a) Host terminal, storing transaction data



(b) Client terminal running facilitator agent

Figure 7 - Illustration of CORBA implementation to extract information from a host

5.1. Hierarchical clustering results

This procedure attempts to identify relatively homogeneous groups of cases based on selected characteristics. A total of 4 clusters had been identified and the group membership of each case is partially indicated in the Table 1. In summary, cluster 1 has a total of 433 cases (43.3%), cluster 2 has 235 cases (23.5%), cluster 3 has 165 (16.5%) and cluster 4 has 167 cases (16.7%).

5.2. Data mining Results after Clustering

After implementing the clustering, we then employed the association rule algorithm (Apriori property). The result is shown in Table 1. The use of such algorithm is for discovering association rules that can be divided into two steps: (1) find all itemsets (sets of items appearing together in a transaction) whose support is greater than the specified threshold. Itemsets that meet the minimum support threshold are called frequent itemsets, and (2) generate association rules from the frequent itemsets. All rules that meet the confidence threshold are reported as discoveries of the algorithm.

The result only shows the two clusters and the first five rules generated. The support threshold that we set prior to the experiment was 0.90. In the original dataset, those who buy A6 (books on religion) will most likely buy A2 (books on social science) and F4 (Health supplement) with support of 0.935 and confidence of 0.942 (94.2% probability of buying). The same fashion of explanation and analysis could be done to other rules.

Table - Comparison of the Discovered Rules

Models	Discovered Rules, (showing 2 clusters and the first 5 rules generated)	Sup.	Conf.
Without HCARD (1,758 rules)	A6=Buy -> A2=Buy F4=Buy	0.935	0.942
	A6=Buy -> A2=Buy A3=Buy	0.927	0.934
	A6=Buy -> A2=Buy C4=Buy F4=Buy	0.916	0.922
	A6=Buy -> A2=Buy D2=Buy F4=Buy	0.915	0.921
	A6=Buy -> A2=Buy A3=Buy C4=Buy	0.910	0.916
HCARD Clusters 1(1154rules)	A6=Buy -> A2=Buy F4=Buy	0.924	0.939
	A6=Buy -> A2=Buy A3=Buy	0.919	0.934
	A6=Buy -> A2=Buy D2=Buy F4=Buy	0.905	0.920
	A6=Buy -> A2=Buy C4=Buy F4=Buy	0.903	0.918
	A6=Buy -> A3=Buy F4=Buy	0.901	0.915
2 (708 rules)	A6=Buy -> A2=Buy F4=Buy	0.912	0.923
	A6=Buy -> A2=Buy A3=Buy	0.910	0.921
	A6=Buy -> A2=Buy	0.963	0.974
	A6=Buy -> A2=Buy C4=Buy	0.942	0.953
	A6=Buy -> A2=Buy D2=Buy	0.940	0.951

In cluster 1, those who buy A6 (religion) will most likely buy A2 (social science) and F4 (Health) with support of 0.924 and confidence of 0.939 (93.9%). Similar approach of analysis could be made for other rules in this cluster. And a similar fashion of explanation could also be done for other rules discovered such as in clusters 2, 3 and 4, respectively.

One good implication for our findings is that we reduced

the task of making generalization on the rules obtained for the entire dataset. If we will think of increasing the number of dataset entries, it will follow that the number of rules should be exponentially increasing and it is impractical to that effect to handle such output. We had noted some convenience in analyzing the rules generated since we based it on the rules obtain per cluster.

5.3. Processing performance

The difference of computing performance is shown by the graph below, showing the comparison between HCARD (clustered) and without HCARD. Shorter processing time had been observed to compute for smaller clusters of attributes implying faster and ideal processing period than processing the entire dataset. The result of the time comparison is shown in Figure 8.

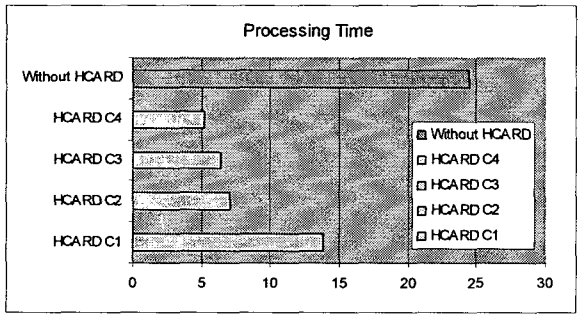


Figure 8 - Processing time between HCARD and without HCARD

The blending of cluster analysis and association rule generation in HCARD model specifically isolate groups of correlated cases using the hierarchical nearest neighbor clustering and then using of the data mining like the algorithm for association rule generation. The HCARD identify relatively homogeneous groups of cases based on selected characteristics and then employed the Apriori algorithm to calculate for association rules. This resulted to some partitions where we could conveniently analyze specific associations among clusters of attributes. This further explains that the generated rules were discovered on clusters indicating highly correlated cases which will eventually implies simplification of analyses for the data mining results.

6. Conclusions and Recommendations

Although our implementation uses heuristic approach, the experiment shows that efficiency in terms of convenience and practicality in analyzing the results based on the discovered rules had been observed. The Facilitator Agent (Integrator) based on CORBA architecture was used for search and extraction of data from heterogeneous servers in the distributed environment. Shorter processing time had been noted in computing for smaller clusters implying faster and ideal processing period than dealing with the entire dataset. We have provided examples, performed experiments and generated rules but more rigorous treatment maybe needed if dealing with more complex databases.

The HCARD model reveals clusters that have high correlation according to predetermined characteristics and generated isolated but imperative association rules based on clustered data which in return could be practically explained for decision making purposes. Our future tasks involve implementation of the model on real databases since we used only synthetic and randomized data in our experiment. Furthermore, we would explore the use of fuzzy neural network to improve the search and extraction capability of the proposed integrator agents.

7. References

[1] S. El-Beltagy, D. C. DeRoure, and W. Hall, "A Multiagent system for Navigation Assistance and Information Finding", In Proceedings of the Fourth International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology, pages 281-295, 1999.

[2] B. Chen, P. Haas, and P. Scheuermann, "A new two-phase sampling based algorithm for discovering association rules" Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2002.

[3] J.L. Hellerstein, S. Ma, and C. S. Perng, "Discovering actionable patterns in event data" IBM Systems Journal, Vol. 41, No. 3, 2002.

[4] J.M.W. Lam, "Multi-Dimensional Constrained Gradient Mining" Published Master's Thesis, Simon Fraser University, 2001.

[5] H. Bronnimann, B. Chen, M. Dash, P. Hass, Y. Qiao, & P. Scheuermann, "Efficient Data-Reduction Methods for On-Line Association Rule Discovery", In Data Mining: Next Generation Challenges & Future Directions, 2004.

[6] J. Siegel, "CORBA Fundamentals and Programming", USA: John Wiley & Sons, Inc., 1996.

[7] Object Management Group, "Common Object Resource Broker Architecture (CORBA)", available at http://www.omg.org, 2005.

[8] E. H. Han, G. Karypis, V. Kumar, & B. Mobasher, "Clustering in a high-dimensional space using hypergraph models", available at www- users. cs. umn. edu/ ~karypis/publications/.../cluster-hyper-dim.pdf.

[9] C. Siourbas, "Determining the Number of Clusters", Available at: http://cgm.cs.mcgill.ca/ ~soss/ cs644/ projects/ siourbas, 2005.