International journal of Advancements in Computing Technology
Volume 2, Number 3, August, 2010

Integration of Data Mining and Data Warehousing: A Practical Methodology

Muhammad Usman and Russel Pears
School of Computing and Mathematical Sciences,
Auckland University of Technology, New Zealand
E-mail:muhammad.usman@aut.ac.nz, russel.pears@aut.ac.nz
doi: 10.4156/ijact.vol2.issue3.4

Abstract

The ever growing repository of data in all fields poses new challenges to the modern analytical systems. Real-world datasets, with mixed numeric and nominal variables, are difficult to analyze and require effective visual exploration that conveys semantic relationships of data. Traditional data mining techniques such as clustering clusters only the numeric data. Little research has been carried out in tackling the problem of clustering high cardinality nominal variables to get better insight of underlying dataset. Several works in the literature proved the likelihood of integrating data mining with warehousing to discover knowledge from data. For the seamless integration, the mined data has to be modeled in form of a data warehouse schema. Schema generation process is complex manual task and requires domain and warehousing familiarity. Automated techniques are required to generate warehouse schema to overcome the existing dependencies. To fulfill the growing analytical needs and to overcome the existing limitations, we propose a novel methodology in this paper that permits efficient analysis of mixed numeric and nominal data, effective visual data exploration, automatic warehouse schema generation and integration of data mining and warehousing. The proposed methodology is evaluated by performing case study on real-world data set. Results show that multidimensional analysis can be performed in an easier and flexible way to discover meaningful knowledge from large datasets.

Keywords*: Automatic Schema, Clustering, Data Warehouse, Multi-dimensional Analysis*

1. Introduction

The extensive use of computers and information technology has made extensive data collection a routine task in a variety of fields [1]. Continuously increasing data repositories can contribute significantly towards future decision making provided appropriate knowledge discovery mechanisms are applied for extracting hidden, but potentially useful information embedded in the data [2]. One of the main mechanisms of knowledge discovery is the efficient analysis of data using modern analytical systems. A tough barrier to the efficient analysis of data is the presence of mixed numeric and nominal variables in real-world data sets. Abundant algorithms and techniques have been proposed in the literature for the analysis of numeric data but little research has been carried out to tackle the problem of mixed numeric and nominal data analysis. Traditional methodologies assume variables are numeric valued, but as application areas have grown from the scientific and engineering domains to the biological, engineering, and social domains, one has to deal with features, such as country, color, shape, and type of disease, that are nominal valued [1]. In addition to the problem of efficient analysis of mixed data, high cardinality nominal variables with large number of distinct values such as product codes, country names, model types are not only difficult to analyze but also require effective visual exploration [3]. Visualization techniques are becoming increasingly important for the analysis and exploration of large multidimensional data sets [4]. However, the results of many visualization techniques such as parallel coordinates [5, 6] are affected by the order by which attributes are displayed [7]. Moreover, accurate spacing among the attribute values is mandatory to recognize to the semantic relations in the underlying data.

The major focus of this paper is the seamless integration of data mining and data warehousing. Data mining aims at the extraction of synthesized and previously unknown insights from large data sets [8].

It can be viewed as an automated application of algorithms to detect patterns and extract knowledge from the data that is not obvious to the user [9]. Data warehousing is recognized as a key technology for the exploitation of massive amounts of data nowadays available electronically in many organizations [10]. The two disciplines, namely data warehousing and data mining are both mature in their own right but surprisingly little research has been carried out in integration of these two strands of research. The key problem is that for the integration to occur in a seamless manner, the data has to be modeled in a data warehouse schema. Data warehouse modeling is a complex task, which involves knowledge of business processes of the domain of discourse, understanding the structural and behavioral system's conceptual model, and familiarity with data warehouse technologies [11]. There is an obvious need to automate the schema generation process to overcome the schema modeling complexities and domain dependencies. Additionally, the automation process is required not only to improve the analytical powers but also to increase the flexibility and adaptability of the existing data mining and warehousing systems.

To overcome the problems mentioned above and to fulfill the ever growing requirements, we propose a novel methodology for the seamless integration of data mining and data warehousing. We have used the widely researched technique of data mining discipline, known as hierarchical clustering and automated the generation of most commonly used data warehouse schema called the *STAR* schema. Our proposed methodology has a series of steps. In the first step, we perform hierarchical clustering on the dataset using the numeric attributes of the data to get the individual clusters. The extraction of numeric facts is being done in second step. In the third step *Distance-Quantification-Classing (DQC)* technique proposed by [3] is applied on the nominal data present in each cluster for effective visualization. The *(DQC)* technique, pre-process the nominal variables, calculates the distance between the variables, assigns order and spacing among the nominal values in each variable and finally determine which variable values are similar to each other and thus can be grouped together. The results of *DQC* technique helps in the identification and extraction of the possible dimensions and the dimensional hierarchies within each cluster. In the fifth step, the automatic schema generator use the dimensions and the hierarchy along with the measures or facts extracted in earlier steps to generate warehouse schema as a resultant. In the final step, we build the data cube from generated schema to perform OLAP operations on it. Each of the steps involved in will be discussed in detail in Section 3 of this paper.

The proposed methodology has been implemented and case study has been performed using real world data set namely *Automobile* [12]. We have used a few tools to implement the proposed methodology. For Hierarchical clustering we have used *Hierarchical Clustering Explorer* [13], *XmdvTool* [14] for visualization results of *DQC* approach using parallel coordinates technique, *Microsoft Analysis Services* [15] for the construction of data cube from the automatically generated star schema. We have also developed a prototype using *C Sharp* programming language for the automatic schema generation. Experimental results indicate that the proposed methodology allows efficient analysis (nominal and numeric), effective visualization and a novel method for automatic schema generation. Moreover, the resultant schema allows users to perform targeted analysis based on the natural groping of the underlying dataset. To our knowledge, the proposed methodology is significant as there is no other methodology reported in the literature that incorporates efficient data analysis (nominal and numeric), effective data visualization, automation of schema generation process and seamless integration of data mining and warehousing.

The remaining of the paper is organized as follows: Section 2 highlights the literature review. In Section 3, we present the details of the proposed methodology. Section 4 the implementation details are given followed by Section 5, which gives the results and discussion. Finally, In Section 6, conclusion and possible future directions are drawn.

2. Literature Review

We review past work done relating in four major themes that relate most closely to the research that we undertake. These themes are: numeric and nominal analysis; visualization of

multidimensional data; automatic generation of data warehouse schema; and integration of data mining methods into data warehousing design.

2.1. Numeric and nominal data analysis

Real world datasets consist of a mix of numeric and nominal data. Specially, data sets with a large number of nominal variables, including some with large number of distinct values are becoming increasingly common [3]. For the purpose of efficient analysis of mixed data sets, [2] identified the problems associated with the traditional k-mean algorithm as best suited for numeric data only. In order to perform analysis on mixed data, the authors proposed a new algorithm which uses a cost function and distance measure based on co-occurrence of values. The proposed cost function alleviated the short-coming of cost-effectiveness of Huang's [16] cost function. The limitation of the proposed work is that the analysis relies on co-occurrences of data and discretizing of numeric values which leads to loss of information.

For the same purpose, a feature selection algorithm for mixed data containing both continuous and nominal features was introduced by [17]. The authors stressed that feature selection is a crucial step in pattern recognition. Furthermore, a new evaluation criterion has been used to avoid feature type transformation through careful decomposition of feature space. The limitation of the proposed algorithm is that it produced better results on the experiment performed on artificial data as compared to real-world data. In addition to this, the algorithm has not been compared in terms of computational cost. The proposed algorithm is computationally expensive as a first step it decomposes the feature space along values of nominal values and then combines these measures to produce an overall evaluation. In the quest of mixed data analysis, three different distance measure functions for efficient analysis of mixed variables were compared by the authors in [18]. They identified the fact that there is a strong need to develop Mahalanobis-type distances for mixed variables. The reason is that research done in this regard is either heuristic of makes use of nominal data. The strength of their work is the comparison of measures for computing Mahalanobis-type distances between categorical and numeric dimensions. The limitation of the work is that they have used very small data sets having only a few records to perform the validation. Furthermore, in the nominal data they have not targeted variables with a large number of distinct values, often called the high cardinality nominal variables. In real world data sets there exist a large number of such variables and it is important to analyze these variables in order to identify the semantic relationships among the large number of distinct values present in each variable. Additionally, standard data visualization methods do not deal satisfactorily with high cardinality variables. These methods need to be enhanced significantly to deal with such variables.

The authors in [19], focused on the hierarchical clustering of mixed data based on distance hierarchy. The proposed work differs from the above mentioned work as the authors expressed the distance between categorical values through a hierarchical data structure. The strength of the proposed work was the use of distance hierarchy scheme for both categorical and numeric data. Furthermore, the scheme was applied to mixed data clustering to integrate with the hierarchical clustering algorithm. However, they have not adopted any visualization enhancement technique in the proposed approach. Visualization can play a very important part in the interpretation of both numeric and nominal data. For clustering mixed data, it was reported that most of the clustering algorithms operate mostly on numeric data and only a few can support the analysis of mixed numeric and nominal data[20]. The authors extended the existing Orthogonal partitioning clustering (O-Cluster) algorithm [21] to domains with nominal and mixed variables. The algorithm relies on an active sampling method to accomplish scalability with large volumes of data.

Similar to the work of [19], the proposed (O-Cluster) algorithm uses axis-parallel partitioning to build a hierarchy and identifies hyper-rectangular regions in the input feature space. The limitation was the manual representation of the clustering results instead of using any visualization technique for better interpretation of the results. The authors in [22] proposed a fuzzy clustering approach based on probabilistic distance feature. Authors stressed on the fact that clustering mixed feature-type data is major data analysis task. The cluster formed from this approach contains the weighted means and covariance matrices of numeric attributes and weighted frequencies of the nominal attributes categories. The weakness of the proposed work is that the clustering process has been guided to get clusters on the

basis of nominal data only. Experiments were performed only on the artificial data sets and not on any real-world dataset.

The authors in [23] targeted the same area of clustering mixed data. Authors presented an Evidence based Spectral Clustering Algorithm (EBSC) that works well for data containing a mix of both nominal and numeric features. The performance of the proposed algorithm has been evaluated by performing experiments on real data sets. It has been claimed that the measure based on evidence accumulation works well with data described by mixed type features. The major weakness again in the proposed work was the lack of data visualization techniques and the absence of high cardinality nominal variables in the dataset. The authors in [1] demonstrated that the similarity measure proposed by [24] works well with data with mixed nominal and numeric features. An agglomerative algorithm and simple distinctness heuristic is used to extract a partition of the data from the dendogram. Although the algorithm works with mixed data but it is computationally expensive. For the purpose of effective visualization and analysis of high cardinality nominal variables a new technique was proposed by [3], called *Distance-Quantification-Classing (DQC)* approach. The authors investigated an assignment of order and spacing among nominal data with a large number of distinct values to highlight the relationships among the data points.

It is evident from the literature discussed in this section that efficient analysis of mixed nominal and numeric data is a major problem. Literature review reveals the fact that most commonly used data mining techniques used for the analysis of mixed data is clustering [1, 2, 17, 20, 22, 23]. In clustering, the use of hierarchical clustering technique has shown good results [3, 19]. In our proposed methodology we also adopt the hierarchical clustering technique for the analysis of mixed data but our approach to applying hierarchical clustering is different from [3, 19]. Furthermore, we are using the visualization results generated by DQC approach. In the next section we review the effectiveness of some of the major multidimensional data visualization techniques.

2.2. Effective visualization of multi-dimensional data

For effective visualization of data, authors in [4] introduced the similarity clustering of dimensions as an important technique of enhancing the results of a number of different multi-dimensional visualization techniques. Authors presented a number of similarity measures to determine similarities among dimensions to target the dimension arrangement problem. A heuristic solution based on an intelligent ant system has been proposed for enhanced visualization. The major limitation of the proposed work is that it is only applicable to simply three types of visualization techniques namely parallel coordinates, circle segment and recursive pattern. In addition to this, the work only supports the arrangement (order) of dimension and not the spacing among the values of each dimension. In order to extract useful information from the high cardinality nominal variables both effective spacing among the values is required. Meaningful spacing among the values plays a vital role in the interpretation of visualization results and helps in the recognition of meaningful patterns from the underlying data values. Experiments are performed on the data sets having a small number of dimensions. The work does not give any indication when high dimensional data has to be visualized. Visualization of higher dimensional data was not attempted. More specifically, parallel coordinates technique is more suitable for the datasets which have a small number (maximum 10) of dimensions. High dimensional datasets are difficult to visualize with this particular technique because the effectiveness of parallel coordinates display decreases with increase in the number of variables.

For the problem of high dimensional data visualization, [7] proposed a strategy that offers both dimension ordering and dimension reduction. Authors claimed that enhanced data visualization of high dimensional data can be achieved through dimension reduction and attribute ordering. It has been argued that it is possible to considerably improve the quality of visualization by embedding a simple heuristic to handle the axes layout problem. Based on this heuristic, authors introduced a technique called *Similarity-Based Attribute Arrangement (SBAA)* which produces improved visualization. The weakness of the work is that dimension reduction although improving the visualization, also takes away the analytical flexibility from the analyst. There should be more sophisticated techniques provided to the analyst to ensure that dimensionality reduction does not occur merely on the basis of similarity or correlation. In order to reduce dimension in parallel coordinates, [6] introduced

visualization methods for multidimensional data sets which also includes an effective dimension reduction method based on a genetic algorithm.

In the quest for effective visualization, [5] proposed interactive visualization of large multivariate data sets based on a number of novel extensions to the parallel coordinates display technique. Furthermore, a proximity based coloring scheme ensures that data and clusters from similar parts of the hierarchical structure are shown in similar colors. Navigational tools are also embedded to support localization of data and to perform drilling operations on the data. The only limitation of the work is that the effectiveness of the proposed work depends very much on the parallel coordinate technique. However, different types of data sets in the real world require different types of techniques for the effective visualization. Similar to the work of [5, 7, 25] introduced another efficient approach to construct frequency and density plots from parallel coordinates visualization. The new plots proposed by the authors allow users to remove noise and highlight areas with high concentration of data. This permits data clusters to be visually explored and interactively extracted, regardless of their shape and dimensionality. The results of the experiments were not compared with other algorithms for high dimensional clustering which is a major drawback of the work.

Similar to the work discussed in this section, in our proposed methodology we use parallel coordinate visualization technique for effective exploration of nominal data present in each cluster. As mentioned earlier the parallel coordinates technique is more suitable when there is small number of dimensions to display. Our use of this technique is limited to the visualization of only the high cardinality nominal variables present in each data cluster. Furthermore, our aim is to identify and visualize the groupings among the values present in each variable using this visualization technique of parallel coordinates. In addition to this, we intend to use the underlying clustered data from the parallel coordinates to aid in the process of automatic data warehouse schema generation. In the next section, some of the work done on the automatic generation of schema will be presented.

2.3. Automatic generation of DW schema

Authors in [26], identified the fact that most of the research focuses on the automatic derivation of database schemata from conceptual models but neglect the automatic derivation of OLAP metadata in a way that is integrated with database schema. It was emphasized that such integration is extremely important because it allows end-users tools to query the data warehouse accurately and reduces development time and cost. Model-transformation architecture has been proposed in order to facilitate the automatic generation of warehouse schema. The research work has been implemented in an open-source development platform to automatically generate schema from conceptual multidimensional models.

Likewise, [11] suggested an Object-process-based Data Warehouse Construction Method (ODWC) for the construction of data warehouse schema. The method uses a stepwise rule-based algorithm for the derivation of schema from the source operational model. The proposed ODWC method overcomes the major limitation of manual work requirement to construct the schema and lack of automated assistance for the identification of facts and dimensions from the conceptual models. However, ODWC has been applied on only one case study to date. There is strong need to apply this method on various case studies to strengthen its effectiveness and applicability. Similarly, [10] presented a technique for obtaining, in a fairly automatic way, a data warehouse designed over a set of source operational databases. The proposed technique takes in a list of database schemas in the form of ER model, a dictionary of lexical synonymy properties, and generates a data warehouse as an output. Again the limitation of the technique is that it is only evaluated merely on a single case study.

To achieve the goal of automation, [27], proposed the generation of tool specific schemata for OLAP from conceptual graphical models. A new approach, *Bablefish*, has been introduced and the issues of automatic schema generation have been discussed by the authors. It is suggested that the data warehouse schema can be generated using the conceptual graphical models. The limitation of the approach is that it only takes in to account the initial design stage when no data has been loaded to the system. In real world development, the data warehouse architect needs to design schema along with the data, which is a complex task. Therefore, the proposed approach is not suitable for designing warehouse schemas from existing databases.

The existing work in the area of automation was highlighted by [28] and the authors claimed that automation is focused towards data models, data structures specifically designed for Data warehouse (DW), and criteria for defining table partitions and indexes. The major research contribution of the authors is a step forward towards the automation of DW relational design achieved through a rule-based mechanism, which automatically generates the DW schema with the help of existing DW design knowledge. The proposed rules embed design strategies which are triggered by conditions on requirements and source databases, and perform the schema generation through the application of predefined DW design oriented transformations. For the sake of accomplishing the ease in the automation work, authors in [29] built a conceptual model (*StarER*) for data warehouse design on the basis of user modeling requirements. The *StarER* model combines the star structure, which is dominant in data warehouses, with the semantically rich constructs of the ER model. Comparison of the proposed model with other existing models has been performed, pointing out differences and similarities. Examples from a mortgage data warehouse environment, in which *starER* is tested revealed the ease of understanding of the model, as well as its efficiency in representing complex information at the semantic level.

Likewise, an automatic tool for generating star schema from an Entity-Relationship Diagram (ERD) was introduced by [30]. A prototype named *SAMSTAR* was presented, which was used for the automatic generation of star schema from an ERD. The system takes an ERD drawn by *ERwin* Data Modeller as an input and generates star schemas. SAMSTAR displays the resulting star schemas on a computer screen graphically. With this automatic generation of star schema, this system helps designers reduce their effort and time in building data warehouse schemas. More recently, authors in [31] proposed architecture for the automatic OLAP schema generation from the clustered data. The authors presented a model for the use of data mining results for the generation of automatic schema. The proposed work is similar to our work as we also aim to use clustered data for the automatic schema generation but differs from the fact that our objective is to find the natural grouping in each dimension for a particular cluster. These natural groups become the hierarchies of a given dimension. Such hierarchies facilitate the end user to perform analytical operations on the data for a meaningful analysis.

It is evident from the literature on automatic generation of schema that none of the previous works used hierarchical clustering and visualization techniques to aid automatic detection of hierarchical dimensions and schema generation. In this paper, we use the visualization results along with the hierarchically clustered data to identify possible dimensions, dimensional hierarchies and facts from the underlying dataset. In the next section, we highlight the work done on the integration of data mining and data warehousing which is the primary focus of this research.

2.4. Exploiting data mining techniques in DWs

An infrastructure for parallel multidimensional analysis and data mining was suggested by the authors in [32]. The work done argued the use of OLAP queries which are ad hoc in nature and require fast computational time by pre-aggregating calculations. Data mining uses some of these pre-computed aggregated calculations to compute the probabilities needed for calculating support and confidence measures for association rules. In order to perform OLAP and data mining operations, authors in [33] presented an algorithm and techniques for constructing data cubes and showed that these data cubes can be used for data mining using *Attribute Focusing* technique. Since data cubes have aggregation values on combinations of attributes already calculated, the computations of attribute focusing are greatly facilitated by data cubes. Authors claimed that dimensional hierarchies can be utilized to provide multiple level data mining. This is useful for mining at multiple concept levels and interesting information can potentially be obtained at different levels.

The authors in [8] integrated data warehousing with data mining technique before passing the data to an OLAP engine. With this integration, framework was devised to enhance the visualization capabilities of OLAP and hence making it more useful and intelligent. Authors claimed that with the presence of these clusters the OLAP user have better insight into the data set before performing OLAP operations. The major limitation of the work was that the clusters generated by neural networking technique are to be manually translated in to relational tables. Secondly interactive exploration on the clustered data was missing. To overcome these limitations, authors in [34] proposed and enhanced architecture for the combination of data warehousing and data mining. The proposed architecture

provided a way for integrated enhancement of OLAP's performance and its data visualization using self-organizing neural networks. The proposed architecture was validated using real-life data sets and it was claimed that with the integration of enhanced OLAP with data mining system a higher degree of advancement can be achieved in modern analytical systems.

An enhanced OLAP operator based on the Agglomerative Hierarchical Clustering (AHC) was suggested [35]. The main problem identified is OLAP's capability in aggregating and summarizing complex objects like text, images, sounds and videos. The operator called Operator for Aggregation by Clustering (OpAC) is able to provide significant aggregates of facts referred to complex objects. The strength of the work is the combination of data warehousing and data mining techniques as data mining can discover knowledge from both simple and complex data. However, to classify *N* individuals, AHC generates *N* hierarchical partitions but it does not give guidance as to a best partition to choose. Data miner has to decide the number of clusters that corresponds to the context and to the goal of his/her analysis. A data mining system, *DBMiner*, has been developed for interactive mining on large data warehouses [36]. A very simple architecture has been proposed but the system has not been designed for satisfactory exploratory data mining. The issues related to efficient and effective data mining were discussed by the authors in [37] and it was suggested that data warehousing and mining techniques should work in parallel for the sophisticated analysis.

For the integration of data mining methods with warehousing, authors in [38] proposed and developed an interesting association rule mining approach called Online Analytical Mining of association rules. It integrated data warehousing with association rule mining methods and leads to flexible multidimensional and multi-level association rule mining. In the same way, a methodology was introduced by [39], which derives the rules of web pages tick sequences that are according to the support and confidence level of speculated by the user. This methodology identified a set of frequently accessed web pages on a website by a user. The result is the list of potential customers for a certain product or service on a target web page.

For the same integration purpose a new architecture has been proposed by [40]. It has been suggested that with this integration knowledge discovery in the data warehouse can be enhanced. Mining should be performed on the data cube. Furthermore, the mining engine should support different types of mining methods i.e classification, clustering, association rules and many more. The proposed architecture has not been compared and implemented with any other existing architecture. The proposed architecture allows mining the data cube but not the operational databases.

It is apparent from the review that data warehousing and mining, when used in conjunction with each other, can generate benefits in a number of diverse domains. In this paper, we present a methodology to integrate data mining with warehousing for the efficient analysis and visualization of data. In addition to this, our proposed methodology is not domain specific. Our focus is to provide an easily implementable methodology for the seamless integration of the two mature disciplines.

3. Proposed Methodology

In this section, we propose a methodology for the seamless integration of data mining and data warehousing. Figure 1 gives the overview of the proposed methodology.

For its introduction, we present the motivation of the proposed methodology. The question arises, *what role does data mining play in enhancing the design of data warehouses?* To answer this important question, we begin the explanation starting with the overview of the previous section of literature review. It is evident from the literature review that none of the previous work done in the past has addressed the efficient analysis of numeric and high cardinality nominal variables, effective visualization, automatic schema generation and integration of ware housing and mining in a single framework. Several works [8, 32, 33, 35, 36, 40, 41] suggested that with the integration of data mining with the warehousing system a number of benefits can be achieved. In this paper, we have used the hierarchical clustering technique for the efficient analysis of data as a pre-processing step of data mining.

In addition to this, we use the *Distance-Quantization-Classing (DQC)* approach for the effective visualization of the high cardinality nominal variables. Moreover, we utilize the natural grouping in

each nominal variable along with the numerical values in each cluster to aid the automatic schema generator by providing a natural dimensional hierarchy that has been mined from the data to build a warehouse schema. Most of the previous work for automatic schema generation used ER diagram or conceptual graphical models to automate the process. To our knowledge, none of the work reported in the literature used effective visualization results based on the underlying dataset to find the natural grouping within the high cardinality nominal variables, and to aid the schema generation process. To fulfill the growing requirements and to overcome the existing limitations, a novel methodology for the integration of data mining and data warehousing is required.

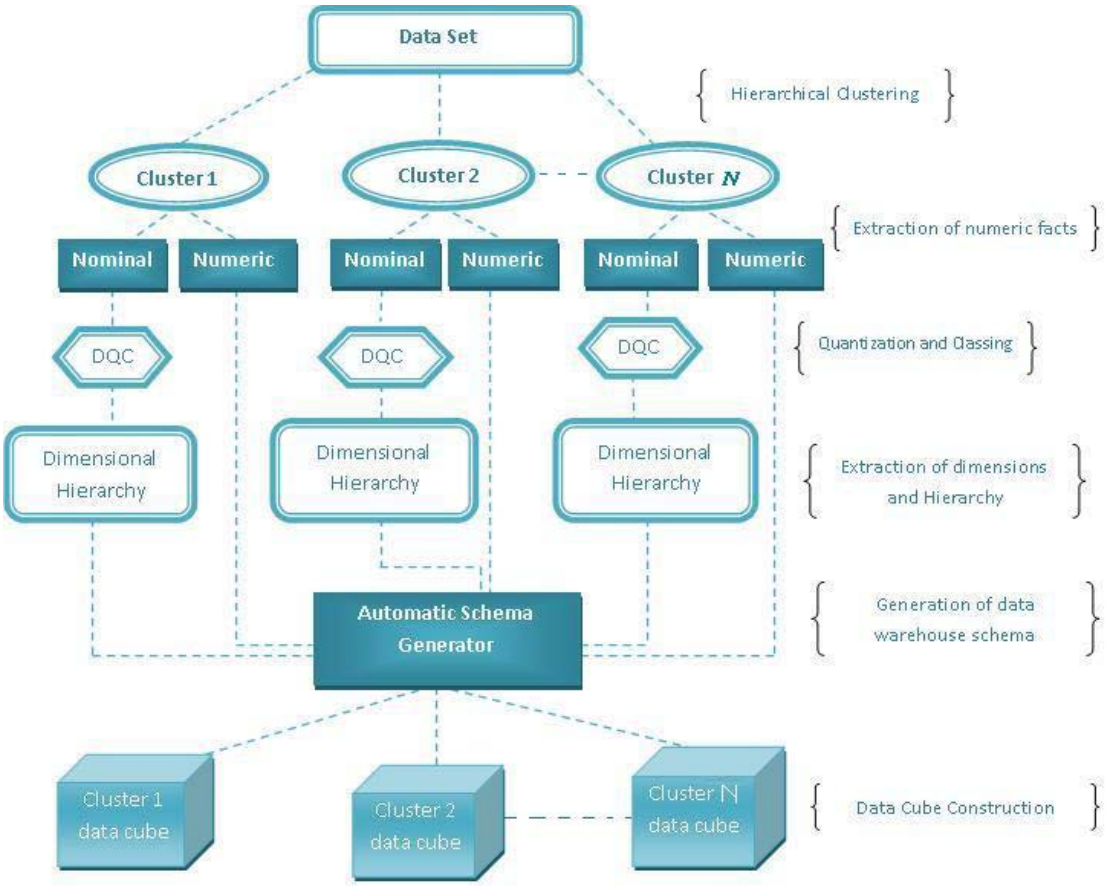

Figure 1. Overview of the proposed methodology

The proposed methodology consists of a series of steps to meet its objective of integrating data mining with warehousing. The description of each step involved in the methodology is explained as follows:

STEP 1: Hierarchical clustering of numeric data

In this step, the hierarchical clustering technique is applied to the data set to generate clusters based on a similarity measure. The most common similarity measures include complete-linkage, average-linkage, and single-linkage. The reason for choosing hierarchical clustering is that it tends to produce natural clusters instead of performing unnecessary merges and splits like other clustering approaches. Furthermore, it allows users to set parameters to determine the proper number of clusters. As most of the clustering algorithms works well on numeric variables, in this step, we target the numeric variables in order to get the optimal results.

STEP 2: Extraction of numerical facts

In step 2, the numerical data present within each cluster is being extracted. The reason for this extraction is that these numerical values are the facts or measures within the data which forms the input

to the automatic schema generation process. In data warehouses the numerical attributes represent core potential measures which the analysts want to visualize from different dimensional perspectives.

STEP 3: Application of Distance-Quantification-Classing (DQC) on nominal data

After getting the hierarchical clusters and extracting the numerical values, DQC approach is applied on each cluster, for the mapping of nominal data into numeric data for effective visualization. The purpose of this application is that in real-world data sets there typically exist a number of nominal variables which have high cardinality. For example country names and produce codes are typical examples of the high cardinality nominal variables. The DQC approach maps the nominal values in to numeric values for effective visualization. Additionally, the approach assigns order and spacing among the variable values in a manner that conveys relationships and associations in the data items. For instance, the DQC approach can group the product codes or country names that are closer to each other based on the other variables in the underlying data set. This assignment of order and spacing is done in such a way that the distance between the two values in the nominal space is preserved in the numeric space.

STEP 4: Extraction of dimensions and dimensional hierarchy

Following step 3, the mapped nominal to numeric values are extracted. These values are responsible for defining the groups in each dimension and the dimensional hierarchy. These extracted values are to be fed into the automatic schema generator to model the dimensional hierarchy in the data warehouse schema.

STEP 5: Automatic generation of warehouse schema

In step 4, the extracted values from the previous step 2 and step 4, become the input to the automatic schema generator. Automatic schema generator first reads the input dimensions and measures. Secondly, handles the dimensional hierarchies. Schema generator module identify the natural groupings of the values within each dimension and name the group *i.e. Group 1, Group 2 to Group N.* Each of the groups created by the schema generator is then assigned the values which are closer to each other in the mapped numeric space (details are discussed in the case study).Thirdly, creates a fact table and manages the relationships among the fact and dimension tables. As an output, this step gives a star schema and also populates the data in the corresponding dimension and fact tables using automatically generated queries.

STEP 6: Construction of data cube

Finally, when the schema has been generated, the data cube is being constructed to allow various data warehouse operations such as drill-down, roll-up, slicing and dicing and pivoting. The construction of data cube allows the flexibility to add/remove the dimensions and to control the granularity of the warehouse analysis.

In the next section, we discuss the implementation of the proposed methodology using two case studies based on real-world data sets.

4. Implementation Details

In this section, based on our implementation, we discuss details of the implementation steps of the proposed methodology. We have performed case studies on two real world data sets from the UCI machine learning repository, namely *Automobile* and *Adult* datasets to validate the results of our proposed methodology.

Case Study: Automobile data set

For the first case study, we have used a small *Automobile* data set. The data set consists of mixed nominal and numeric variables. There are in total 26 (10 nominal and 16 numeric) attributes present in the data set. The first step of the experiment is to perform hierarchical clustering on the data. For this we have used a hierarchical clustering tool, called Hierarchical Clustering Explorer (HCE). HCE is a visualization tool for interactive exploration of multidimensional datasets. One of the goals of HCE is to help users explore and understand multidimensional datasets by maximizing the human perceptual skills that have been underutilized. This tool takes a text file as input and clusters the data using different clustering parameters. The hierarchical clusters produced from the data are shown in the form of a

dendogram. We fed *Automobile* data and set the row wise clustering option and measure was the Euclidean distance. The reason for this choice is that Euclidean distance is the most commonly used type of distance measure in cluster analysis. Using the complete linkage method, three clusters were generated by the tool. Figure 2 shows the hierarchical clusters produced using HCE tool.

The second step of the proposed methodology is to extract the numerical facts from each of the clusters generated. HCE tool provides exporting to data present in each cluster. We export the data present in each cluster and then extract only the numeric attributes and save them in an Excel file. The next step is the application of Distance-Quantification-Classing (DQC) approach. For the mapping of nominal data to numeric data we have used the java based nominal to numeric converter based on DQC approach, developed by [3]. The converter produce two output files as a result, one containing the mapped nominal values, and the other a meta file which has the dimension value ranges in it. Some high cardinality nominal (categorical) attributes for *Automobile* are shown in Table 1.

Table 1. High cardinality variables in the data set

Attribute name	categorical values
Make	22
Fuel System	8
Engine Type	7
No of Cylinders	7
Body Style	5

Figure 2. Hierarchical clusters generated by HCE

After the application of DQC, we have used the *XmdvTool* which uses the parallel coordinates display technique to display the high cardinality dimensions. Figures 3 and 4 shows the parallel coordinate display of cluster 1 and cluster 3 both containing nominal variable data.

It is clear that each display for a cluster gives different grouping based on the underlying data within each cluster.

The next step is to extract the nominal data grouping within each cluster. We use the meta file produced by the converter tool to extract the dimensional value names and ranges XML format. Based on the similarity measure we define groups within each dimension. For instance, it is clear from Figure 3 that the *Make* variable has an obvious group, say Group 1 in which the makes such as *Volvo, Mazda, Nissan* and *Peugot* are combined together. Similarly, Figure 4 gives different groups and each group has its own set of makes, which in general differ from those in the previous cluster. These groups can be used as a hierarchy within each dimension or variable. We claim that each high cardinality variable can become a dimension in the warehouse schema and the individual groups within each dimension will then define the dimensional hierarchy for that particular variable.

The next step of the methodology, after the extraction of the nominal value names and ranges from the XML file, is the generation of warehouse schema. We have built a prototype that reads the two extracted files, the Excel file from step 2 and the XML file from step 4, to generate the warehouse schema. The prototype creates a single fact table having the measures from the Excel file and dimension tables from the high cardinality variables. Finally, automatically generated queries construct a star schema suitable for a given data warehousing platform which in our case happened to be Microsoft SQL Server.

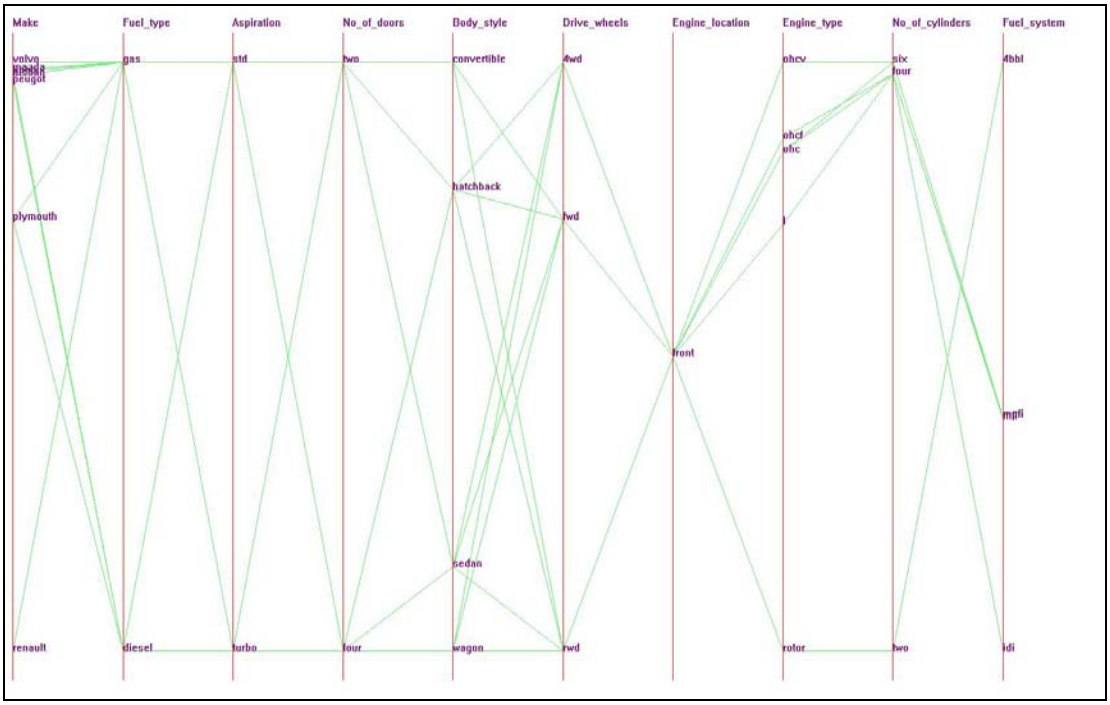

Figure 3. Grouping nominal variables (Cluster 1)

Figure 5 shows the design of the generated schema. It can be seen that each dimension in the generated schema has its individual grouping which are based on the underlying dataset. For instance, *Make* dimension groups all the values present in Cluster 3 into 4 individual groups and each group consists of individual values which are close to each other. For example, Group 1 has *Prosche and Nissan* and Group 2 has *Jaguar, Mercedez-Benz and Toyota.* Likewise, every group formed has its own set of values which happen to be close to each other in cluster 3. Similarly, every cluster obtained has the same dimensions but every dimension has different number of groups and diverse values are present in each group.

Lastly, Microsoft Analysis Services has been used to construct the data cube from the automatically generated schema. Once the data cube, has been constructed the user can perform OLAP operations like drill-down, roll-up, slice-and-dice and pivoting.

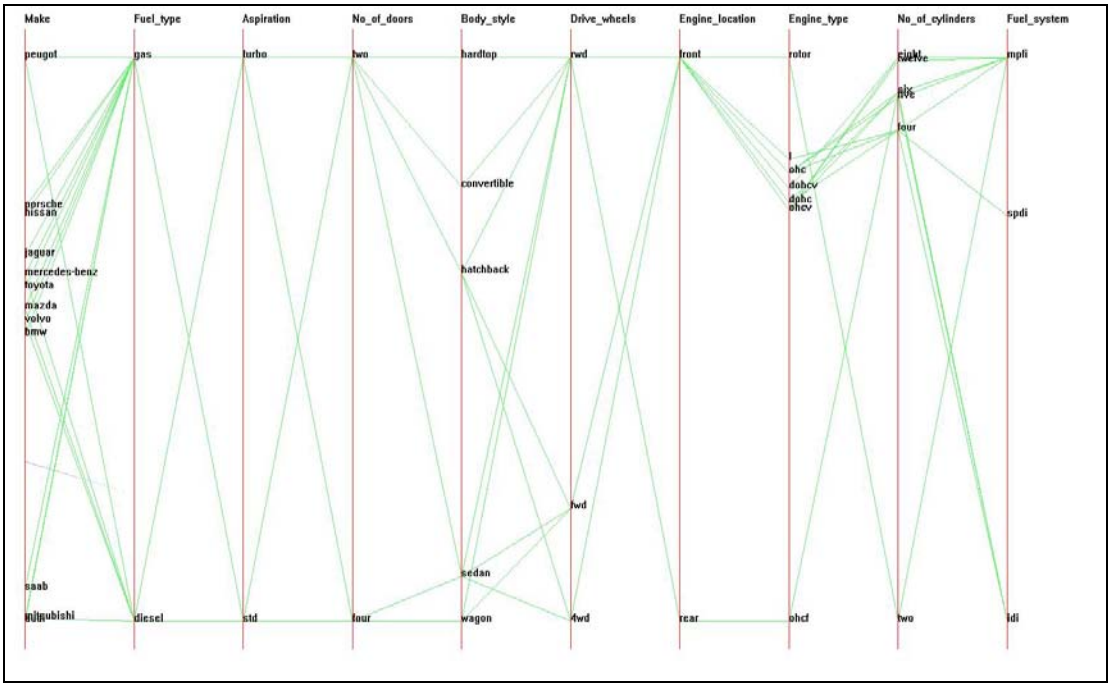

Figure 4. Grouped nominal variables (Cluster 3)

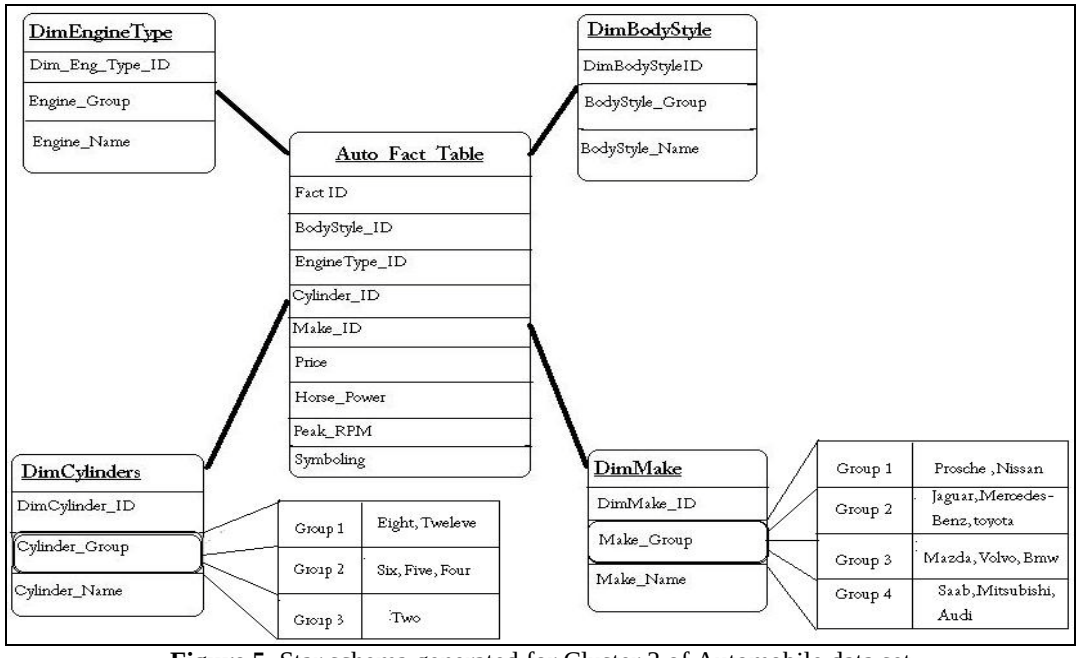

Figure 5. Star schema generated for Cluster 3 of Automobile data set

5. Results and Discussion

The results of the studies are discussed pertaining to the *Automobile* data set. This section is divided into four subdivisions. We discuss the results with respect to each of our objectives that we aimed to achieve in our proposed methodology.

One of the important aspects of the proposed methodology is the efficient analysis of mixed numeric and nominal data. Most of the clustering algorithms work well on the numeric data only but fail to produce meaningful clusters when mixed data is provided. In order to achieve efficient analysis of mixed data we applied a series of different methods. With the application of DQC approach effective visualization has been achieved. Parallel coordinates of the *XmdvTool* highlighted the natural groupings of the underlying data set within each cluster. For the sake of discussion, we compare the grouping among some high cardinality nominal variables of the two clusters, *Cluster1* and *Cluster3* produced from the *Automobile* data set. We compared four variables namely *Make, Body-Style, Engine-Type* and *No-of-Cylinders.* It has been observed that the number of groups for a given variable in *Cluster1* and the values within each group are different when compared with *Cluster3*. For instance, in *Cluster1* result, *No-of-Cylinders* has 1 group that contains (six, four). In *Cluster3* the same variable has 3 groups each containing different set of values. It can be seen from the groups of *Clsuter3* that group2 of *Cluster3* has (six, five and four) as values. Similarly, *Engine-type* variable in Cluster1 has a group which consist of only two type of values (ohcf and ohc) whereas the same variable in *Cluster3* has five distinct values in group2 (l, ohc, dohc, ohcv, dohcv). Here we can see that the (ohcf) engine type is totally absent in the group of *Cluster3*.

This shows that each cluster has its own groupings or relationships based on the underlying data set. In our proposed methodology, we use these variables as dimensions and the groupings within each variable as possible dimensional hierarchy levels. The Parallel coordinates technique helps the user to visualize these grouping among the nominal values which improves the effectiveness of the visual display.

Furthermore, the numeric facts and the dimensions and hierarchy levels are fed into the automatic schema generator to give a star schema as an output. The schema generator produced a schema on using the *Cluster3* data. All the numeric values extracted from each cluster become potential measures and the high cardinality nominal variables become the dimensions. With these dimensions and measures a multi-dimensional data cube was generated. We show a few results; with the help of some OLAP operations how the data cube in Cluster3 can facilitate warehouse analysis for the end-user. Suppose the analyst requires the following. *Which particular automobile make, with which Engine-type and Body-style has the highest aggregated price in Cluster 3.* For such multidimensional queries, using OLAP we select *Price* as measure from the fact table and use the 3 dimension tables. First, we see the desired measure with *Engine-type* dimension and find out which group has the maximum *price* value. In this case, Group2 of *Engine- type* has the highest value. We select group 2 and then select the *Body-style* dimension. Within this dimension we identify that group 1 has the highest value for price. We move forward and select *Make* dimension and drill down into the individual automobile names after identifying the group which has the maximum values and can easily find out that *Mercedes-Benz* has the highest aggregated price.In addition to this final result, we also identified the two groups, Group 2 of engine-type having and Group1 of Body-style having (Convertible, Hatchback) are the two groups from desired dimensions which has the highest values for the Price measure.

Finally, the proposed methodology integrates the data mining technique of hierarchical clustering with data warehousing. Efficient analysis of data, effective visualization and the automation in schema generation are the added advantages of the proposed methodology.

6. Conclusion and future work

In this paper, a novel methodology has been proposed for integrating methods such as clustering and pattern visualization into data warehouse design. We used hierarchical clustering and automated the warehouse schema generation process. Additionally, we demonstrated that efficient analysis and effective visualization of mixed nominal and numeric data is a very important but often neglected task. We have validated the methodology by performing case study on real world datasets from the UCI machine learning repository. Results show that data clustering and visualization methods, working in

conjunction with each other can be used to gain new insights and build more meaningful dimensions. With the seamless integration of data mining and warehousing, the analytical capabilities of the modern analytical system can be enhanced remarkably. The proposed methodology is significant as it incorporates efficient data analysis, effective data visualization and automation of schema generation. Future work is mainly intended towards the incorporation of more sophisticated visualization techniques for the exploration of clustered data. In addition to this, we are working on the overall improvement of the proposed methodology to make it more effective, efficient and suitable for large and complex data sets.

7. References

[1] C. Li, and G. Biswas, "Unsupervised learning with mixed numeric and nominal data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 4, 2002, pp. 673-690.

[2] A. Ahmad, and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data & Knowledge Engineering*, vol. 63, no. 2, 2007, pp. 503-527.

[3] G.E. Rosario, E.A. Rundensteiner, D.C. Brown, M.O. Ward, and S. Huang, "Mapping nominal values to numbers for effective visualization," *Information Visualization*, vol. 3, no. 2, 2004, pp. 80-95.

[4] M. Ankerst, S. Berchtold, and D.A. Keim, "Similarity clustering of dimensions for an enhanced visualization of multidimensional data," *Book Similarity clustering of dimensions for an enhanced visualization of multidimensional data*, Series Similarity clustering of dimensions for an enhanced visualization of multidimensional data, ed., Editor ed.^eds., 1998, pp. 52.

[5] Y.H. Fua, M.O. Ward, and E.A. Rundensteiner, "Hierarchical parallel coordinates for exploration of large datasets," *Proceedings of the conference on Visualization'99*, IEEE Computer Society Press, pp. 43-50.

[6] J.X. Chen, and S. Wang, "Data visualization: parallel coordinates and dimension reduction," *Computing in Science & Engineering*, vol. 3, no. 5, 2001, pp. 110-112.

[7] A.O. Artero, M.C.F. de Oliveira, and H. Levkowitz, "Enhanced high dimensional data visualization through dimension reduction and attribute arrangement," *Proceedings of the confernece on Information Visualization(IV)* pp. 707-712.

[8] S. Asghar, D. Alahakoon, and A. Hsu, "Enhancing OLAP functionality using self-organizing neural networks," *Neural, Parallel & Scientific Computations*, vol. 12, no. 1, 2004, pp. 1-20.

[9] S. Chaudhuri, and U. Dayal, "An overview of data warehousing and OLAP technology," *ACM Sigmod record*, vol. 26, no. 1, 1997, pp. 65-74.

[10] L. Palopoli, L. Pontieri, G. Terracina, and D. Ursino, "A novel three-level architecture for large data warehouses* 1," *Journal of Systems Architecture*, vol. 47, no. 11, 2002, pp. 937-958.

[11] D. Dori, R. Feldman, and A. Sturm, "From conceptual models to schemata: An object-process-based data warehouse construction method," *Information Systems*, vol. 33, no. 6, 2008, pp. 567-593.

[12] J.C. Schlimmer, "UCI repository of machine learning databases, "http://archive.ics.uci.edu/ml/datasets/Automobile" 1987.

[13] J. Seo, M. Bakay, P. Zhao, Y.W. Chen, P. Clarkson, B. Shneiderman, and E.P. Hoffman, "Interactive color mosaic and dendrogram displays for signal/noise optimization in microarray data analysis," *Proceedings of the International Conference on Multimedia and Expo*, pp. 461-464.

[14] "XMDVTOOL HOME PAGE, 2003 http://davis.wpi.edu/~xmdv."

[15] S. Soni, and W. Kurtz, "Analysis Services: optimizing cube performance using Microsoft SQL server 2000 Analysis Services," *Microsoft SQL Server 2000 Technical Articles*, 2001.

[16] Z. Huang, "Clustering large data sets with mixed numeric and categorical values," *Book Clustering large data sets with mixed numeric and categorical values*, Series Clustering large data sets with mixed numeric and categorical values, ed., 1997, pp. 21–34.

[17] 17. W. Tang, and K.Z. Mao, "Feature selection algorithm for mixed data with both nominal and continuous features," *Pattern Recognition Letters*, vol. 28, no. 5, 2007, pp. 563-571.

[18] B. McCane, and M. Albert, "Distance functions for categorical and mixed variables," *Pattern Recognition Letters*, vol. 29, no. 7, 2008, pp. 986-993.

[19] C.C. Hsu, C.L. Chen, and Y.W. Su, "Hierarchical clustering of mixed data based on distance hierarchy," *Information Sciences*, vol. 177, no. 20, 2007, pp. 4474-4492.
[20] B.L. Milenova, and M.M. Campos, "Clustering large databases with numeric and nominal values using orthogonal projections," *In Proceedings of the 29th Conference on Very Large Databases (VLDB), Berlin, Germany*.
[21] B.L. Milenova, and M.M. Campos, "O-cluster: scalable clustering of large high dimensional data sets," *Proceedings of 2002 IEEE International Conference on Data Mining (ICDM'02)*, pp. 290–297.
[22] C. Doring, C. Borgelt, and R. Kruse, "Fuzzy clustering of quantitative and qualitative data," *IEEE Annual Meeting of the Fuzzy Information, 2004. Processing NAFIPS'04*, pp. 84-89.
[23] H. Luo, F. Kong, and Y. Li, "Clustering mixed data based on evidence accumulation," *Advanced Data Mining and Applications*, vol. 4093, 2006, pp. 348-355.
[24] D.W. Goodall, "A new similarity index based on probability," *Biometrics*, vol. 22, no. 4, 1966, pp. 882-907.
[25] A.O. Artero, M.C.F. de Oliveira, and H. Levkowitz, "Uncovering clusters in crowded parallel coordinates visualizations," *Book Uncovering clusters in crowded parallel coordinates visualizations*, Series Uncovering clusters in crowded parallel coordinates visualizations, ed., Editor ed.^eds., 2004, pp. 81-88.
[26] J. Pardillo, and J.N. Mazón, "Designing OLAP schemata for data warehouses from conceptual models with MDA," *Decision Support Systems*.
[27] K. Hahn, C. Sapia, and M. Blaschka, "Automatically generating OLAP schemata from conceptual graphical models," *Proceedings of the 3rd ACM international workshop on Data warehousing and OLAP*, ACM, pp. 9-16.
[28] V. Peralta, A. Marotta, and R. Ruggia, "Towards the Automation of Data Warehouse Design," *15th Conference on Advanced Information Systems Engineering, short paper proceedings (CAISE FORUM)*
[29] N. Tryfona, F. Busborg, and J.G. Borch Christiansen, "starER: A conceptual model for data warehouse design," *Proceedings of the 2nd ACM International Workshop on Data Warehousing and OLAP (DOLAP)*, ACM, pp. 3-8.
[30] I.Y. Song, R. Khare, Y. An, S. Lee, S.P. Kim, J. Kim, and Y.S. Moon, "Samstar: An automatic tool for generating star schemas from an entity-relationship diagram," *Conceptual Modeling-ER 2008*, pp. 522-523.
[31] M. Usman, S. Asghar, and S. Fong, "Data Mining and Automatic OLAP Schema Generation," *Book Data Mining and Automatic OLAP Schema Generation*, Series Data Mining and Automatic OLAP Schema Generation, ed., Editor ed.^eds., 2010, pp.
[32] S. Goil, and A. Choudhary, "PARSIMONY: An infrastructure for parallel multidimensional analysis and data mining," *Journal of parallel and distributed computing*, vol. 61, no. 3, 2001, pp. 285-321.
[33] S. Goil, and A. Choudhary, "High performance OLAP and data mining on parallel computers," *Data Mining and Knowledge Discovery*, vol. 1, no. 4, 1997, pp. 391-417.
[34] M. Usman, S. Asghar, and S. Fong, "Integrated Performance and Visualization Enhancement of OLAP Using Growing Self Organizing Neural Networks," *Journal of Advances in Information Technology*, vol. 1, no. 1, 2010, pp. 26-37.
[35] R.B. Messaoud, O. Boussaid, and S. Rabaséda, "A new OLAP aggregation based on the AHC technique," *Proceedings of the 7th ACM international workshop on Data warehousing and OLAP*, ACM, pp. 65-72.
[36] J. Han, "Towards on-line analytical mining in large databases," *ACM Sigmod Record*, vol. 27, no. 1, 1998, pp. 97-107.
[37] J. Han, S. Chee, and J.Y. Chiang, "Issues for on-line analytical mining of data warehouses," *Proceedings of the Workshop on Research Issues on Data Mining and Knowledge Discovery, Seattle, Washington*.
[38] H. Zhu, "On-line analytical mining of association rules," Department of Computer Science, Simon Fraser, 1998.
[39] J. Fong, H.K. Wong, and A. Fong, "Online analytical mining web-pages tick sequences," *Journal of Data Warehousing*, vol. 5, no. 4, 2000, pp. 59-67.

[40] Marketos G, Theodoridis Y, and Kalogeras S.I, "seismological Data Warehousing and Mining," *International Journal of Data Warehousing & Mining*, vol. 4, no. 1, 2008, pp. 1-16.
[41] M. Usman, S. Asghar, and S. Fong, "A Conceptual Model for Combining Enhanced OLAP and Data Mining Systems," *2009 Fifth International Joint Conference on INC, IMS and IDC*, IEEE, pp. 1958-1963.