Representation of Concept Hierarchy using an Efficient Encoding Scheme

Ruchika Yadav
Research Scholar,
Department of Computer Science and Application,
Kurukshetra University,
Kurukshetra, Haryana, India

Kanwal Garg, Ph.D.
Assistant Professor,
Department of Computer Science and Application,
Kurukshetra University,
Kurukshetra, Haryana, India

Mittar Vishav
Assistant Professor,
Department of Computer Science and Engineering,
HCTM Technical Campus,
Kaithal, Haryana, India

ABSTRACT
The premise of this paper is to use an efficient encoding scheme which will be used to encode high level concept hierarchy of a transactional table. This table will work as the base to generate multiple level association rules. These rules discovers the hidden knowledge align at higher level of abstraction. Therefore the numeric encoding of the concept hierarchy improves the time complexity and space complexity of task relevant data.

Keywords
Concept hierarchy, Encoding scheme, Transaction databases.

1. INTRODUCTION
Data mining with its unlimited diversity of techniques and approaches may be applicable to retrieve knowledge at any kind of information repositories[1] like sensor network data mining [2], gene ontology mining [3], cloud computing [4], spatial data mining [5,6], network intrusion detection [7,8] and many more. The information retrieval from the transactional database is becoming very tedious because it may include large number of concept hierarchies. Table 1.1 shows the specimen instances of a computer store specifying its daily sales.

Table 1: A Sales_Transaction Table

Transaction_id	Bar_Code_Set
4900	{23890, 23489, 86853, 75339}
4901	{55321, 21675, 29573, 86213}
4902	{78657, 21675, 21227, 23367}
4903	{72284, 26072}
4904	{26072, 22480, 20193, 23890, 26781}
4905	{86853, 20507, 28267}
4906	{20193, 22981, 23489, 72284}

In this table; First tuple states the detail of sales i.e. against transaction id 4900 the customers bought 4 specific items e.g. 23890 stands for computer desktop of make 'LG', 23489 stands for mouse of make 'Apple' and 55321 stands for laptop of make 'Dell' and so on in other tuples. This is a traditional approach in which item is recorded against a randomly generated Bar code number.

1.2 Problem Identification
After peruse the Table 1.1 the author examined that the instances so generated for the daily sales of computer store are suitable for association rules mining. Since the item code so generated are random in nature. Therefore it is very difficult to identify similar item in the same basket.

This paper has been divided in six sections. In which section 1 introduced the transactional database, its parametric coding and problem identification, section 2 provides the brief overview of concept hierarchy, section 3 explains different encoding schemes, section 4 highlights the important comparative points for encoding schemes, section 5 exemplify an appropriate encoding scheme, section 6 concludes the paper followed by references.

2. CONCEPT HIERARCHY
Concept hierarchy organize data or concepts in hierarchical forms or in certain partial order, which are used for expressing knowledge in brief, high-level terms, and make possible mining knowledge at multiple levels of abstraction [9].

A conceptual hierarchy consists of a set of nodes organized in a tree, where the nodes represent values of an attribute called concepts [10]. A special node, "ANY", is reserved for the root of the tree. A number is assigned to the level of each node in a conceptual hierarchy. The level of the root node is one. The level of a non-root node is one plus the level of its parent level number. Since values are represented by nodes, the levels of nodes can also be used to define the levels of values. Concept hierarchy allows raw data to be handled at a higher and more generalized level of abstraction. Concept hierarchies have been classified into four major categories which are explained below.

2.1 Schema Hierarchy
Schema hierarchy defines the total or partial order among attributes in the database. It may express existing semantic relationships between attributes. An example of schema concept hierarchy is shown in Figure 1. This Figure explores that the partial order along with a sequence of attributes: D is at one level lower than C, which is in turn at one level lower than B and so on.

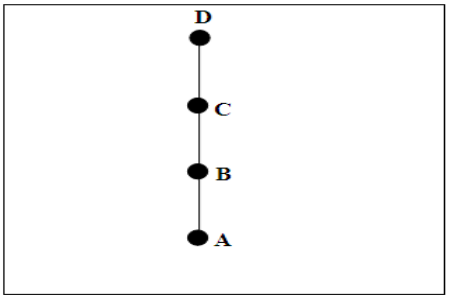

Figure 1: A Schema Concept Hierarchy

In a database, more than one schema hierarchies can be created by using different sequences and grouping of attributes.

2.2 Set-Grouping Hierarchy

A set-grouping hierarchy organizes values for a given attribute or dimension into groups or constant range values [11]. It is also called instance hierarchy because the partial order of the hierarchy is defined on the set of instances or values of an attribute. These hierarchies have more operational sense and so preferred than other hierarchies. The set-concept hierarchy can be expressed as shown in Figure 2.

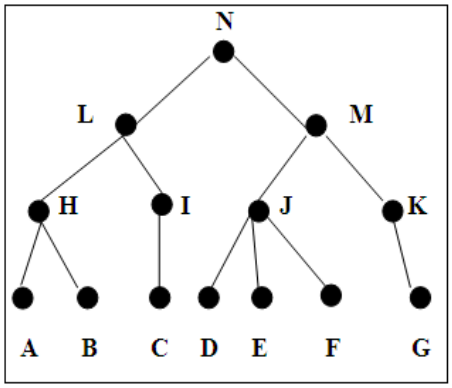

Figure 2: A Set Grouping Hierarchy

This Figure explains that attribute N is divided into two groups L and M. These are divided on the basis of domain of attribute N. Further partial order of the attributes L and M has been defined on the set of instances or values of these attributes. So, partial order of L and M are H, I, J and K respectively.

A set-grouping hierarchy can be used for transforming a schema hierarchy or another set-grouping hierarchy to form a sophisticated hierarchy.

2.3 Operation-Derived Hierarchy

Operation-derived hierarchy is defined by a set of operations on the data. These operations are specified by users, experts, or the data mining system. These hierarchies are generally defined for numerical attributes. Such operations can be as simple as range value comparison, as complex as a data clustering and data distribution analysis algorithm.

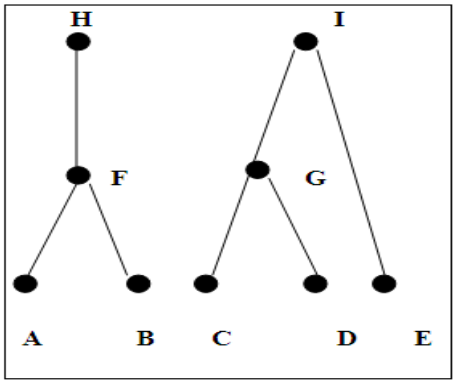

Figure 3: Operation-Derived Hierarchy

Figure 3 depict an operation-derived hierarchy for a simple range value comparison operation on given data. The attributes H and I has two different range values. The attribute with high value is placed as right child and with less value is at left. likewise in this figure with the subpart G has value less than I and E has greater than I and so on.

2.4 Rule-based Hierarchy

In a rule-based hierarchy either a whole concept hierarchy or a portion of it is defined by a set of rules and is evaluated dynamically based on the current data and rule definition. A lattice-like structure is used for graphically describing this type of hierarchies, in which every child-parent path is linked with a generalization rule. The following Figure 4 highlights a rule based concept hierarchy.

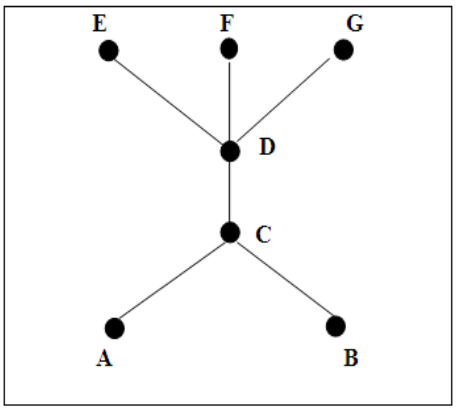

Figure 4: A Rule Based Concept Hierarchy

Finally, for a particular data mining task some specified hierarchies may not be desirable. So, there should be mechanisms for automatic generation of concept hierarchies based on the data distributions in a data set. The static and dynamic generation of concept hierarchy is totally depends on data sets. In this context, the generation of a concept hierarchy based on a static or dynamic data set is called static or dynamic generation of concept hierarchy.

3. ENCODING OF CONCEPT HIERARCHY

This section covers the implementation of concept hierarchies. To incorporate the concept hierarchies into a data mining system, encoding plays a key role. The encoding of a concept hierarchy is attempted in such a way that the partial order of the hierarchy is exactly represented by the codes. In the following subsections, three type of encoding schemes are explained.

3.1 Post-Order Traversal Encoding

Wang and Iyer[12] had proposed a post-order traversal encoding method of the hierarchy. Figure 5 explores the post-order traversal encoding of a small hierarchy with 16 nodes.

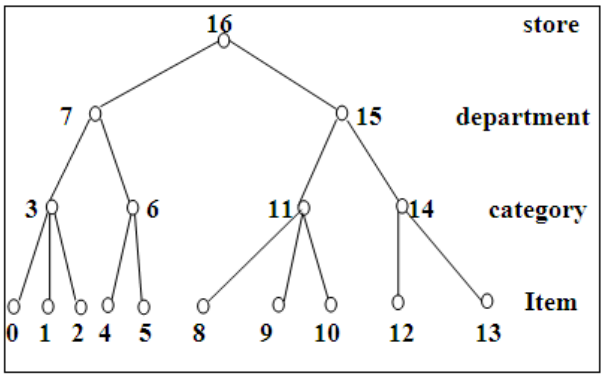

Figure 5: Post-Order Traversal Encoding

This encoding method is based on the post-order traversal of tree. According to this, for any node with label n, if the smallest label of its descendents is m, then m < n and it has exactly (n-m) descendents with labels from m to (n-1). Thus all the integers in the range [m, n-1] give the labels of all its descendents.

This encoding scheme is suitable for the roll-up and drill-down operation in OLAP, especially when cooperated with the DB2 features [13]. But there are limited applications of this encoding scheme.

3.2 Binary Encoding Method

A new hierarchy encoding method was proposed by Yijun Lu [14], which can be treated as a generic purpose encoding strategy that is suitable for any data mining functionalities. The main idea is to assign a unique binary code to each node of a hierarchy which consists of n fields, where n is the level number of the node in the given concept hierarchy. A binary encoded hierarchy is demonstrated in Figure 6.

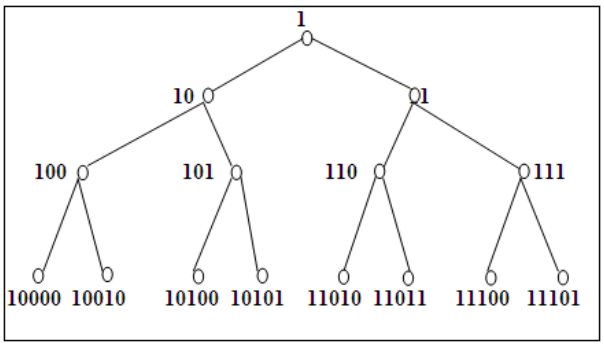

Figure 6: A Binary Encoded Concept Hierarchy.

Once a hierarchy is encoded then the only task is to retrieve the codes of the hierarchy from the memory and recognize generalization and specialization by manipulating the codes. The idea of assigning binary numbers to the nodes of concept hierarchies has been useful in various areas such as logic programming, digital source coding and data compression.

3.3 Positional Encoding Method:

A new concept hierarchy encoding scheme is presented by Han and Fu in 1995 [15]. In this positional encoding method nodes are encoded with respect to their position in the hierarchy using sequences of numbers. In this encoding scheme nodes are encoded with respect to their position in the hierarchy using sequences of numbers. This type of encoding can be done during the collection of task relevant data. This takes fewer bits than the corresponding object identifier. In this encoding scheme, actual items are represented by terminal nodes on the tree. The internal nodes represent classes or concepts formed from lower level. This encoding scheme is suitable for multiple level association rules mining, especially when describes the association among items at different concept levels.

4. PERFORMANCE ANALYSIS AND COMPARISION

In this section, the performance of using concept hierarchy is analyzed on the basis of previous studies. Analytical estimates for both storage requirement and disk access time are given for the following two approaches:

4.1 Without Encoding

In this approach a concept hierarchy is stored using a collection of several tables in which real concepts are used as join key. A concept hierarchy consists of a number of tables, each of which is a map table from a lower level to its next higher level. Another approach is adopted by usual OLAP systems [12], is similar to above approach but instead of using real concept name as join key between tables, here a unique integer identifier is assigned to each node for the purpose of table join.

4.2 With Encoding

This approach uses one relational table for each concept hierarchy. The encoding of a concept hierarchy is attempted in such a way that the partial order of the hierarchy is exactly represented by the codes. Post-order encoding is adopted for the drill-down operation in OLAP. However binary encoding and other positional encoding are suitable for data cube creation as well as for all the functional modules such as summarizer, comparator, associator, classifier and predictor.

Notice that, the storage space required for encoded tables is less than the real concept names used in the related operations and also the less CPU time is consumed for executing functional modules with encoded tables.

Yijun Lu concluded that the encoding approach performs better than without-encoding approach in terms of storage and disk access time. The encoding method gives us a way to use less storage and obtain efficient processing of data mining tasks. The proposed encoding algorithm is useful and efficient especially for concept generalization [14].

There are several encoding schemes and each scheme is used for different applications of concept hierarchies. As shown in Table 2.

Table 2: Application Areas of Encoding Schemes

S. No.	Encoding Scheme	Applications Areas
1.	Post-order Traversal Encoding	Roll-up and Drill-down operations in OLAP.
2.	Binary Encoding	Logic programming, Digital Source Coding and Data Compression.
3.	Positional Encoding	Data Cube Creation, Multiple Level Association Rules Mining, Classification and Prediction.

Since the different encoding schemes are based on different application area. Therefore because of this dependability comparison among different encoding schemes is very difficult. Hence comparative study for these encoding schemes is ineffective.

5. CASE STUDY: A COMPUTER STORE

In this section, the researcher has designed a case study of a computer store to exemplify an appropriate encoding scheme. With reference to the problem identified in section 1.2 the researcher proposed an efficient encoding scheme as shown in upcoming paragraph.

Han & Fu [15] discussed positional encoding scheme is suitable for mining association rules at different levels of abstraction, especially when describes the association among items at different concept levels. During multilevel association rule mining, the taxonomy information for each (grouped) item in Figure 7 is encoded with respect to their position in the hierarchy using sequences of numbers.

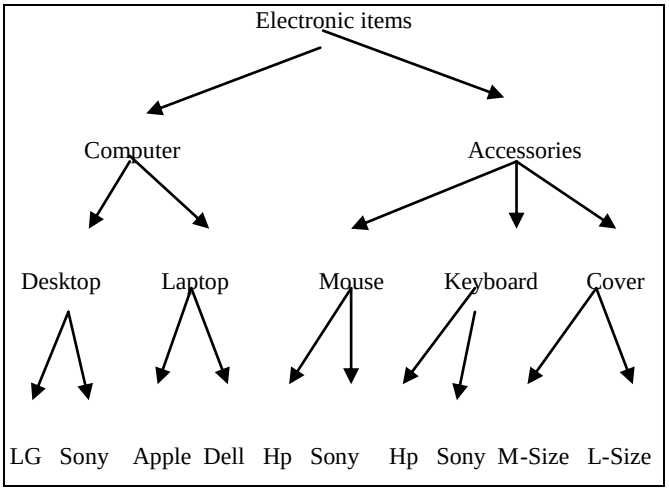

Figure 7: Taxonomy for Relevant Data Items

For example, the item 'Computer Desktop LG' is encoded as `111' in which the first digit, `1', represents `computer' at level-1, the second, `1', for the item type 'desktop' at level 2, and the third, `1', for the `company' at level-3. The repeated items (i.e., items with the same encoding) at any level will be treated as one item in one transaction. The codes of item name of concept hierarchy given in Figure 7 are represented below in Table 3.

Table 3: Codes of Item Name

Item Name (terminal node)	**code**	**Item Name (Internal node)**	**code**
Computer Desktop LG	111	Accessories Cover L-Size	232
Computer Desktop Sony	112	Computer Desktop	11*
Computer Laptop Apple	121	Computer Laptop	12*
Computer Laptop Dell	122	Accessories Mouse	21*
Accessories Mouse Hp	211	Accessories Keyboard	22*
Accessories Mouse Sony	212	Accessories Cover	23*
Accessories Keyboard Hp	221	Computer	1**
Accessories Keyboard Sony	222	Accessories	2**
Accessories Cover M-Size	231		

The taxonomy information for each (grouped) item in Table 1 A sales_transaction is encoded as a sequence of digits in the transaction Table 4.

Table 4: Encoded Transaction Table

TID	**Items**
T_1	{111,121,211,221}
T_2	{111,211,222,232}
T_3	{112,122,221,411}
T_4	{111,121}
T_5	{111,122,211,221,413}
T_6	{211,323,524}
T_7	{323,411,524,713}

6. CONCLUSIONS

This research paper concludes that it is difficult to analyze a traditional transactional database. Therefore an encoding scheme is introduced for better analysis and appropriate production of frequent item sales. This encoding scheme also improves the space complexity and time complexity. The resultant encoding transaction table helps for the conversion of transactional data into encoded transactional table.

7. REFERENCES

[1] Han, J., Kamber, M.: Data mining: concepts and techniques. Morgan Kufmann Publisher (2000).

[2] D. Han, Y. Shi, W. Wang et al., Research on multi-level association rules based on geosciences data," Journal of Software, vol. 8, no. 12, 3269–3276, 2013.

[3] Pietro Hiram Guzzi, Marianna Milano and Mario Cannataro, Mining association rules from gene ontology and protein networks: promises and challenges. In Proceeding 14th International Conference on Computational Science, Published by Elsevier Vol.29, 1970-1980, 2014.

[4] KW Lin and DJ Deng, A novel parallel algorithm for frequent pattern mining with privacy preserved in cloud computing environments. International journal of Ad Hoc and Ubiquitous Computing, Inderscience publication, 205-215, 2010.

[5] Annalisa Appice, Margherita Berardi, Michelangelo Ceci, and Donato Malerba, Mining and filtering multi-level spatial association rules with ARES. Proceedings in 15th International Symposium, ISMIS 2005, Saratoga Springs, NY, USA, 342-353, 2005.

[6] B. Petelin, I. Kononenko, V.Malaˇciˇc, and M. Kukar, Multi-level association rules and directed graphs for spatial data analysis. Expert Systems with Applications, vol. 40, no. 12, 4957–4970, 2013.

[7] H. Han, X. L. Lu, and L. Y. Ren, Using data mining to discover signatures in network-based intrusion detection. Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing, vol. 1, 2002.

[8] H. Zhengbing, L. Zhitang, and W. Jumgi, A novel intrusion detection system (NIDS) based on signature search of data mining. WKDD First International Workshop on Knowledge discovery and Data Ming, 10-16, 2008.

[9] Han, J.: Mining knowledge at multiple concept levels. Proc. 4th Int'l Conf. on Information and Knowledge Management (CIKM'95), Baltimore, Maryland, Nov. (1995) 19–24.

[10] R. Wille. Concept lattices and conceptual knowledge systems. Computer & Math-ematics with Applications, 23, 493-515, 1992.

[11] R. S. Michalski. Inductive learning as rule-guided generalization and conceptual simplification of symbolic description: unifying principles and a methodology. Workshop on Current Developments in Machine Learning, Carnegie Mellon University, Pittsburgh, PA, 1980.

[12] M. Wang and B. Iyer. Ecient roll-up and drill-down analysis in relational database. In 1997 SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 39-43, 1997.

[13] D. Chamberlin. Using the new DB2: IBM's object-relational database system. Morgan Kaufmann, 1996.

[14] Yijun Lu. Specification, generation and implementation concept hierarchy in data mining. December 1997.

[15] J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. In Proc. 1995 Int. Conf. Very Large Data Bases (VLDB'95), Zurich , Switzerland, 420-431, 1995.