Intelligent Query Answering By Knowledge Discovery Technique

Tin Tin Yee, Khin Sandar
Computer University (Pathein), Myanmar
tin903@gmail.com , drkhinsandar@gmail.com

Abstract

In this paper, we propose the framework of knowledge discovery technique for intelligent query answering. In a database system, there may exist two kinds of queries: data queries and knowledge queries. Data query finds concrete data stored in a database and corresponds to a basic retrieval statement in a database system. Knowledge query finds rules, patterns and other kinds of knowledge in a database and corresponds to querying database knowledge including deduction rules, integrity constraints, generalized rules, frequent patterns and so on. Our framework is based on attribute-oriented induction approach for the discovery of multiple, statistical rules in large database.

1. Introduction

Nowadays, huge amount of data are already being and will continue to be collected in a large of databases by various kinds of data gathering tools which creates both a need and an opportunity for extracting knowledge from databases. Knowledge discovery in database (KDD) is nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in database [5]. Many knowledge discovery methods have been developed for mining knowledge from data. In the previous stidues [2, 7, 8], an attribute-oriented induction method has been developed for knowledge discovery in database.This method integrate learning-from-examples techniques with database operations and extract generalized data from actual data in database.

Query answering mechanisms can be classified into two categories based on their method of response: direct query answering and intelligent (or cooperative) query answering. Direct query answering is a direct, simple retrieval of data or knowledge from database; whereas intelligent query answering consists of analyzing the intent of query and providing generalized, neighborhood or associated information relevant to the query [8]. Intelligent query answering can provide interesting services for e-commerce applications.

This paper is organized as follow. In section 2, we present related work. In section 3, we describe the proposed system framework. In section 4, we present four categories of query answering mechanisms in database. In section 5, we presented result. In section 6, we describe future work and in section 7, we present conclusion.

2. Related Work

J. Han and Y. Fu [4] presented a knowledge discovery system prototype, DBLearn, has been constructed based on this methodology and has been experimented on several large relational databases with satisfactory performance.

T. Imielinski [7] described a new concept of an answer for a query which includes both atomic facts and general rules. He provided a method of transforming rules by relational algebra expressions built from projection, join and selection and demonstrated how the answers consisting of both facts and general rules can be generated.

T. Gaasterland [12] proposed a method to relax a query in order to find neighboring information and to control the relaxation process with user constraints.

3. Proposed System Framework

In this section, our proposed system framework for intelligent query answering by knowledge discovery is present. In figure [1], an overview of our proposed system framework is described .

Many knowledge discovery methods have been developed in studies for mining knowledge from data [5], generalization [7, 12], knowledge representation [1], etc. In this paper is based on one generalization method: attribute-oriented induction (AOI). An attribute-oriented induction method has been developed for knowledge discovery in databases. This method integrates a machine learning paradigm, especially learning-from-examples techniques, with databases operations and extracts generalized data from actual data in databases.

The general idea of attribute-oriented induction is to first collect the task-relevant data using a relational database query and then perform generalization based on the examination of the number of distinct values of each attribute in the relevant set of data. The generalization is performed by either attribute removal or attribute generalization. Aggregation is performed by merging identical generalized tuples and accumulating their respective counts. This reduces the size of generalized data set. The resulting generalized relation can be mapped into different forms for presentation to the user, such as charts or rules.

3.1 Generalization and Extraction of Prime Generalized Relations

Data generalization, statistics summarization and generalized rule extractions are essential techniques for intelligent query answering. The generalization can be performed efficiently by an attribute-oriented induction method [7, 8].Here we present a similar process which extracts a special intermediate generalized relation, prime relation, to facilitate the extraction of different feature tables and the generation of various generalized rules for different purposes of intelligent query answering. Prime relation is a generalized relation in which each nongeneralizable attribute is removed and each generalizable attribute is generalized to the desirable level. The extraction of prime relation can be performed by attribute-oriented induction in the following three steps.

1. Relevant data collection

A set of task-relevant data is collected be using relational database query.

2. Prime relation generation

By removal of nongeneralizable attributes and generalization of the values in the generalizable attributes to the desirable level, some generalized tuples in the relation may become identical. The identical generalized tuples are merged into one tuples, ″count″ which registers the number of original tuples generalized to the current one.

3. Generalized rule extraction

Two methods have been developed for the extraction of generalized rules from prime relation:

(1) To derive a final generalized relation by further application of attribute-oriented induction.

(2) To derive a generalized feature table for intelligent query answering.

A generalized feature table is two-dimensional table generated from prime relation. It represents the occurrence frequency of a set of generalized features in relevance to one or a set of reference attributes in the prime relation. The algorithm for the extraction of a prime relation is described as follow.

Algorithm 3.1: Extraction of prime relation from relational data set

Input:
(1) task-relevant data set R, which is a relation of n A_i $(1 \leq i \leq n)$
(2) a set of concept hierarchies, H_i on attribute A_i
(3) a set of desirability thresholds T_i for attribute A_i

Output: prime relation R

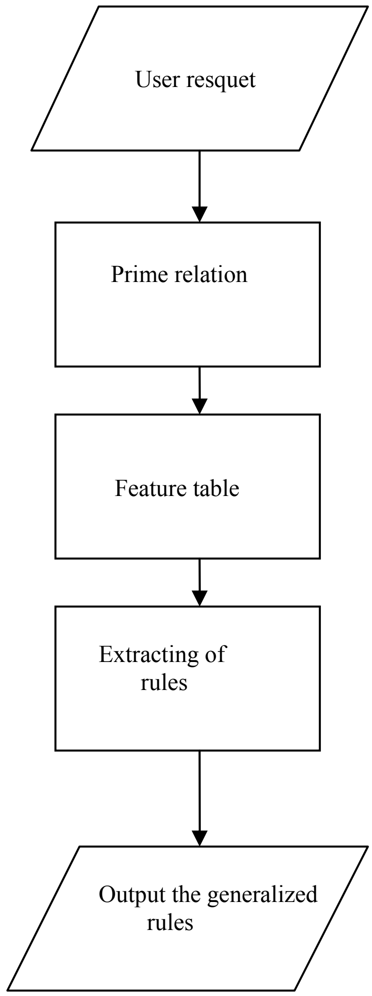

Figure 1. Overview of Proposed System Design

Method.

R_t := R;

/* R_t is temporary relation*/

for each attribute A_i in R_t **do**
{
if A_i is not at the desirable and non-generalizable **then** remove A_i;

if A_i is not at the desirable level but generalizable **then** generalize A_i to the desirable level;
}
/* Identical tuples in R_t are merged with the number of identical tuples registered in count */

R′ :=R_t

Name	Type	Brand	Class	Model	Color	Quantity	Count
TV	LCD	Samsung	Japan	A1	Black	Poor	20
TV	Color	LG	Korea	A2	White	Poor	25
TV	Color	Star	China	A3	Silver	Good	75
TV	Color	Samsung	Japan	A4	Black	Excellent	80
TV	LCD	Daewoo	Korea	A3	Black	Good	30
VCD	1 disc	Star	China	B1	Black	Good	33

Table 1. Simple dataset

Type	Brand	Class	Model	Color	Quantity	Count
LCD	Samsung	Japan	A1	Black	Poor	20
Color	LG	Korea	A2	White	Poor	25
Color	Star	China	A3	Silver	Good	75
Color	Samsung	Japan	A4	Black	Excellent	80
LCD	Daewoo	Korea	A3	Black	Good	30

Table 2. Prime relation

Type	Brand				Class			Model				Color			Quantity			Count
	S	L	ST	D	J	K	C	A1	A2	A3	A4	B	W	Si	P	G	E	
LCD	20	0	0	30	20	30	0	20	0	30	0	50	0	0	20	30	0	50
Color	80	25	75	0	80	25	75	0	25	75	80	80	25	75	25	75	80	180
Total	100	25	75	30	100	55	75	20	25	105	80	130	25	75	45	105	80	230

S=Samsang L=LG J=Japan Si=Silver E=Excellent
K=Korea ST=Star B=Black P=Poor
C=China D=Deawoo W=White G=Good

Table 3. Feature Table for Attribute 'Type'

3.2 Extraction of Generalized Feature Tables and Generalized Rules

To facilitate intelligent query answering, a prime generalized relation can be mapped into several generalize feature tables from which a variety of interesting generalized rules can be extracted. The following algorithm extracts a feature table from a prime relation.

Algorithm 3.2: Extraction of feature table T_A for an attribute A from the prime relation R′

Input: a prime relation R′ consists of an attribute A with distinct values $\{a_1,..,a_n\}$
(1) k other attributes $B_1,\ldots, B_k$ (suppose different attributes have unique distinct values
(2) a special attribute, count

Output: The feature table T_A for attribute A

Method.

1. The feature table T_A consists of m+1 rows and l+1 columns, where m is the number of distinct values in the attribute and l is the total number of distinct values in all of the other k attributes.
2. Each slot in TA is filled by the following procedure,

   for each row r in R′ **do**
   {
   for each attribute B_i in R′ **do**
   T_A [r. A, r, B_i]:=T_A [r. A, r, B_i] + r.count
   T_A [r.A, count]:=T_A [r.A, count] +r.count
   }

3. The last row P in TA is filled by the following procedure,

   for each column S in T_A do
   {
   for each row t (except the last row P) in T_A **do**
   T_A [p , s] := T_A [p , s] + T_A [t , s];
   }

The following algorithm extracts generalized rules from the feature table.

Algorithm 3.3: Extraction of genalized rules from the feature table T_A

Input: - A feature table T_A for the attribute A, where A has a set of distinct generalized value $\{a_1, \ldots, a_m\}$

- Another attribute B in the table has a set of distinct generalized values $\{ b_1, \ldots, b_n \}$
- The slot of the table corresponding to the row with the value ai and the column withthe value bj is referenced by $T_A \{a_i, b_j\}$

Output: A set of generalized rules relevant to A and B extracted from the feature table

Method.

1. For each row a_i, the following rule is generated in relevance to attribute B, which present the distribution of different generalized values of B in class a_i

$$a_i(x) \rightarrow b_1 [p_{i1}] \vee \ldots \vee b_n [p_{in}]$$

where p_{ij} is the probability that the value bj of B is in class a_i, which is computed by ,

$$P_{ij} = T_A [a_i , b_j] / T_A [a_i , count]$$

2. For each column b_j, the following rule is generated in relevance to The last row P in T_A is filled by the following procedure, all the classes, which presents the distribution of the generalized value b_j of B among all the classes

$$b_j(x) \rightarrow a1 [q_{1i}] \vee \ldots \vee am [q_{mj}]$$

Where q_{il} is the probability that value bj of B is distributed in class ai among all the classes,which is computed by,

$$q_{ij}=T_A [a_i , bj] / T_A [total , bj]$$

-LCD(x) →Samsung [40%] V Daewoo [60%]
-LCD(x) → Japan[40%] V Korea[60%]
-LCD(x) → A1[40%] V A3[60%]
-LCD(x) →Black[100%]
-LCD(x) →Poor[40%] V Good[60%]
-Color(x) → Samsung[44.4%] V LG[13.9%] V Star[41.7%]
-Color(x) → Japan[44.4%] V Korea[13.9%] V China[41.7%]
-Color(x) →A2[13.9%] V A3[41.7%] V A4[44.4%]
-Color(x) →Black[44.4%] V White[13.9%] V Silver [41.7%]
-Color(x) →Poor[13.9%] V Good [41.7%] V Excellent[44.4%]

Figure 2. Generalized Rule

4. Four Categories of Query Answering Mechnisms in Database

In database system , there may exist two kinds of queries: data queries and knowledge queries. Data query is find concrete data stored in a database, which corresponds to a basic retrieval statement in a database system. Knowledge query is to find rules and other kinds of knowledge in database, which corresponds to querying database knowledge [14] including deduction rules, integrity constraints, and generalized rules.

However, it is often desirable to provide intelligent and assisted answers to queries instead of direct retrieval of data and knowledge. Therefore, query answering mechanisms in database can be classified into two categories based on their method of response : direct query answering and intelligent query answering. Direct query answering means that a query is answered by returning what is being asked, whereas intelligent query answering consists of analyzing the intent of the query and providing generalized, neighborhood or associated information relevant to the query.

Query answering mechanisms can be categorized into the following four combinations:

- **direct answering of data queries:** direct data retrival in database
- **intelligent answering of data queries:** answer data queries cooperatively and intelligently
- **direct answering of knowledge queries:** a query processor receives a knowledge query and answers it directly by returning the inquired knowledge
- **intelligent answering of knowledge queries:** a knowledge query is answered in an intelligent way by analyzing the intent of the query and providing generalized,neighborhood or associated information

In this paper is using intelligent answering of data queries mechanisms.

Intelligent answering of data queries is mechanisms which answer data queries cooperatively and intelligently. There are many ways for a data query to be answered intelligently, including generalization and summarization of answers (generalized rules), explanation of answers or returning intensional answers, query rewriting using associated or neighborhood information [8], comparison of answers with those similar queries, etc.

5. Result

In this paper, we use 211data set from electronic shop. An analyst or a manager can analyze the reports (generalized rules) of the entire shop over all items. Figure 3,4 and 5 are shown the result.

three disc (X) --> Sony [100%]
three disc (X) --> Japan[100%]
three disc (X) --> DVP-K56P [100%]
three disc (X) --> Silver [100%]
Black&White TV (X) --> Samsung [42.3%] V Daewoo [57.7%]
Black&White TV (X) --> Japan [42.3%] V Korea [57.7%]

Black&White TV (X) --> BW-21T20EL [42.3%] V BD2007 [57.7%]
Black&White TV (X) --> Black [100%]
Color (X) --> Star [41.5%] V LG [13.7%] V Samsung [44.8%]
Color (X) --> Japan [43.7%] V Korea [14.8%] V China [41.5%]
Color (X) --> CS-29K30[0.5%] V CS-29Z40 [0.5%] V CS-29M21FA [43.7%] V T07 [41.5%] V LG-21V [13.7%]
Color (X) --> Black [43.7%] V White [13.7%] V Metal [0.5%] V Sliver [41.5%] V Silver [0.5%]

Figure 3 . Result for Attribute 'Type'

Sony (X) --> three disc [42.9%] V LCD [14.3%] V movie [42.9%]
Sony (X) --> Japan [100%]
Sony (X) --> KLV-32V [14.3%] V DCR-608[42.9%] V DVP-K56P [42.9%]
Sony (X) --> Black [57.1%] V Silver [42.9%]
Panasonic (X) --> Ceiling [96.2%] V Full-size [3.8%]
Panasonic (X) --> Japan [100%]
Panasonic (X) --> NN-C988W [3.8%] V F-600A1 [96.2%]
Panasonic (X) --> Black [3.8%] V White [96.2%]
Mitsubishi (X) --> Ceiling [100%]
Mitsubishi (X) --> Japan [100%]
Mitsubishi (X) --> D12Z [100%]
Mitsubishi (X) --> POOR [100%]

Figure 4. Result for Attribute 'Brand'

Black (X) --> Black&White TV[37.7%] V color[58%] V two disc[0.7%] V LCD[0.7%] V movie[2.2%] V Full-size[0.7%]
Black (X) --> Sony [2.9%] V Crown [0.7%] V Panasonic [0.7%] V Samsung [73.9%] V Daewoo [21.7%]
Black (X) --> Japan [77.5%] V Korea [21.7%] V China [0.7%]
Black (X) --> BW-21T20EL [15.9%] V NN-C988W [0.7%] V KLV-32V [0.7%] V CS-29M21FA [58%] V VCD958 [0.7%] V DCR-608[2.2%] V BD2007 [21.7%]
White (X) --> color [47.2%] V Ceiling [50.9%] V Split [1.9%]
White (X) --> Panasonic [47.2%] V LG [49.1%] V Mitsubishi [3.8%]
White (X) --> Japan [50.9%] V Korea [49.1%]
White (X) --> HSC126RPCO [1.9%] V F-600A1 [47.2%] V D12Z [3.8%] V LG-21V [47.2%]
Sliver (X) --> T07 [100%]
Sliver (X) --> POOR [1.3%] V GOOD [98.7%]
Silver (X) --> three disc [60%] V color [20%] V Compact-size [20%]
Silver (X) --> Sony [60%] V Samsung [40%]

Figure 5 . Result for Attribute 'Color'

6. Future work

This paper has presented sample framework for intelligent query answering by knowledge discovery framework. However, the extraction of the rules provides a flexible means for intelligent query answering, it has two drawbacks: (1) the discovered knowledge is often too task-relevant to be readily applied to other situations, and (2) it is often too costly to extract such knowledge dynamically. This paper can be extended for reports (generalized rule) of time periods such as monthly and yearly .

7. Conclusion

In this paper, a framework has been presented for intelligent query answering by knowledge discovery techniques. Many knowledge discovery methods have been developed for mining knowledge from data. This paper is based on attribute-oriented induction (AOI) method.

8. References

[1] A. Borgida and R . J .Brachman. Loading data into description reasoners. *In Proc.1993 ACM-SIGMOD Int. Conf. Management of Data*, Washington, D.C, May 1993, pp. 217-226.

[2] Y. CAI, N. Cercone, J. Han. Attribute-oriented induction in relational databases. In G. Piatetsky-Shapiro and W. J. Frawley, editors, *Knowledge Discovery in Databases*, AAAI / MIT Press, 1991, pp. 213-228.

[3] W. W. Chu and Q. Chen . *Neighborhood and associative query answering.Journalof intelligent Information Systems* , 1992, pp. 1:355-382.

[4] F. Cuppens and R. Demolombe. Cooperative answering: A methodology to provide intelligent access to databases. *In Proc. 2nd Int. Conf. Expert Database Systems,* Fairfax, VA, April, 1988, pp. 621-641.

[5] W. J. Frawley, G. Piatetsky–Shapiro, and C. J. Matheus. Knowledge discovery in databases: An overview. In G. Piatetsky–Shapiro and C. J Matheus, editors, *knowledge discovery in Databases*, AAAI / MIT Press, 1991, pp. 1-27.

[6] T. Gaasterland ,"Restriction query relaxation through *user constraints". In Proc. International Conference on Intelligent and Cooperative Information Systems*, Rotterdam, Netherlands, May, 1993, pp. 359-336.

[7] J. Han, Y. CAI, and N. Cercone. Data-driven discovery of quantitative rules in relational databases. *IEEE Trans. Knowledge and Data Engineering*, 1991, pp. 5:29-40.

[8] J. Han and Y. Fu . Exploration of the power of attribute-oriented induction in data mining . In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R.

Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, AAAI /MIT Press, 1996, pp. 399-421.

[9] J. Han, Y. Hung, and N. Cercone. Intelligent query answering by knowledge Discovery techniques*. In IEEE Trans. Knowledge and Data Engineer,* 1993.

[10] Jiawei Han, Micheline kamber, *Data Mining Concept and techniques.*

[11] T. Imielinski .Intelligent query answering in ruled based systems. J*. logic Programming*, 1987, pp. 4:229-257.

[12] R . S . Michalski , K. A. Kaufman , and J. S. *Ribeiro. Mining for knowledge in databases: The INLEN architecture, initial implementation and first result*. J. Int. Info. System, 1992, pp. 1:85-114.

[13] A. Motro and Q. Yuan . Querying database knowledge. *In Proc. 1990 ACM–SIGMOD Int. Conf. Management of Data*, Atlantic City, NJ, June 1990, pp.. 173-183.

[14] G. Piatetsky-Shapiro and C. J. Matheus **.** *Knowledge discovery in business database.* Workshop Notes from the Ninth National Conference on Artificial Intelligence: Knowledge Discovery in Databases, 1991, pp. 11-24.

.