International Journal of Soft Computing 2 (3): 464-470, 2007

Association Analysis on Generalized Object-Relational Data Model of Cricket Video Annotations

[1]P. UmaMaheswari and [2]M. Rajaram
[1]Department of Computer Science and Engineering, Sona College of Technology, Salem, India
[2]Department of Electronic and Communication Engineering,
Thanthai Periyar Institute of Engineering and Technology, Vellore, India

Abstract: The ORDB model extends the relational model by providing a rich data type for handling complex objects and object orientation. The major limitation of many commercial data mining algorithms and tools leads to the thought of generalizing the Object -data base. The handling of complex data such as objects effectively is still remaining as a challenging research issue. Data generalization in databases can handle complex data types of the attributes and their aggregations, as necessary and summarizes the information in a relational database by repeatedly replacing specific attribute values with more general concepts according to user-defined concept hierarchies. Mining on video data remains a challenging issue in spite of storing and handling of visual data is very difficult on account of their complex organization structure and Lag of support of existing mining techniques. In this study, video annotation data of a serious of cricket matches are handled effectively in the form of objects that are stored in an Object-relational data base. The association relationships among the action patterns of cricket match are extracted in the form of association rules which are then be represented in the form of textual information. In order to handle the complexity and to make the Object-Relational model of video annotations in the form appropriate for implementing the existing "Apriori algorithm", a generalization technique called "Attribute Focusing and Retrieval" is applied. The associations that are found out of this mining activity can then be converted into useful information, which may help the coaches to train the team to lead for success.

Key words: Data mining, generalization based mining on ORDB, knowledge discovery, attribute focusing, cricket match mining, an application of Apriori algorithm

INTRODUCTION

With the increasing popularity of object based systems in advanced database applications, it is important to study the data mining methods for object-Relational databases because mining knowledge from such databases may improve understanding, organization and utilization of the data stored there. Database size is a most challenging factor which influences more on the various performance characteristics such as scalability, efficiency, compatibility of mining algorithms. It is impractical to indiscriminately mine the entire database, particularly since the number of patterns generated could be exponential with respect to the database size. This leads to the adoption of data generalization. This explosive growth in data and databases has generated an urgent need for new techniques and tools that can intelligently and automatically transform the processed data into useful information and knowledge. Consequently, data mining has become a research area with increasing importance. With the intension to seeks out and discovers interesting patterns in cricket match data, this study highlights the pre-processing of raw data that the program performs, describes the data mining aspects of the software and how the interpretation of patterns supports the process of knowledge discovery. In cricket video annotation, a match is described as a collection of n number of ball shots and every ball shots is a set of actions such as bowling, bating and fielding and also the corresponding result is accumulated with that ball id. An additional provision to represent any special occurrence is also entitled. This initial relation table is too big and more complex with many levels of hierarchy as object structures. All these attributes in Cricket ORDB will not be relevant to the mining process. We need only those which are more specific in functionalities and information. To identity such data, data generalizations will be done as a preprocessing task in order to reduce the complexity as well as size. This generalized relation then be applied for Apriori process to extract interesting patterns. the

Corresponding Author: P.UmaMaheswari, Department of Computer Science and Engineering,
Sona College of Technology, Salem, India

interesting patterns will be evaluated to pick out more confidence rules which are then represented as textual information that may help a coach to assess the effectiveness of certain coaching decisions and formulate game strategies for subsequent matches.

Motivations: We observed that many research works are in progress and also completed on numeric and alphanumeric data (Rakesh *et al*., 1993) whereas very less effort has been taken for multimedia data especially on video sequence data. Video contains visual information which is the most powerful and at the same time the most complex of all the media used for conveying information. Retrieval of video data and mining them to extract useful information is an interesting problem. The challenging feature of mining on complex objects for which still there is no appropriate and effective techniques available influence me a lot to divert my concentration on object based data model (Jawai and Michel, 2002). The extended futures of Object-relational data models to handle objects increased the confidence in my way to overcome the challenging issues (Jawai and Michel, 2002). I have chosen video annotations because of its simplicity and definiteness and it is comparatively easy for creating annotations because every action is definite and limited within a scope.

Related works

Advanced scout (Bhandari *et al*., 1997): It is a PC-based data mining application used by National Basketball Association (NBA) (McMurry, 1995) coaching staffs to discover interesting patterns in basketball game data. Advanced Scout software is described from the perspective of data mining and knowledge discovery. This study highlights the pre-processing of raw data that the program performs, describes the data mining aspects of the software and how the interpretation of patterns supports the process of knowledge discovery. The underlying technique of attribute focusing as the basis of the algorithm is also described. With this information, a coach can assess the effectiveness of certain coaching decisions and formulate game strategies for subsequent games.

Mining frequent patterns in OODBM: (Jan and Petr, 2003; Petr, 2002) This research tackles mining frequent patterns and classification in object-oriented data. A new OR-FP algorithm for mining frequent patterns in object-oriented data is introduced and its implementation described. The OR-FP system loads the data from the Oracle object-relational database system and requires only minimum mandatory settings. Propositionalization with OR-FP for two benchmark domains is described. A modification of OR-FP (Petr, 2002) for processing XML data is discussed. the authors suggested that the features of OOD require more sophisticated data mining techniques. Here a Cinema database it taken for application. it is also analyzed that the OR-FP algorithm is not able to handle continuous data and therefore it is necessary to discretize continuous attributes. In that implementation, Equal Frequency Intervals discretization method is used. Though the system is not able to handle methods in the object-oriented data, they could simply include parameter-free functions which could be seen as virtual attributes due to returning some value for each object. The greatest problem is that the methods can modify objects or that they can be returning different values each time they are called. Thus, the methods employment is said to be still a future challenge. The author lets a modifiation of OR-FP for XML data and evaluation of this modifiation as future work.

A study on Generalization based data mining in object-oriented data bases has already done using object cube models to investigate three aspects:

- Generalize the complex objects
- Class – based generalization
- To extract different kinds of rules. In that research, object cube model is proposed for class based generalization. This study directs future research of developing set of sophisticated generalization operators for complex data generalization.

MATERIALS AND METHODS

Data preparation

Data collection: The raw data from Cricket matches is initially collected using a specialized system designed for logging cricket match data. Data include the type of shot, the entities involved in a ball shot, the role of actors, relationship between entities, description of shot and the outcome. This is a manual task where an skilled assistant sit on and keen the data appropriately on the input form specially designed in a front end, by consequently viewing the cricket video.

Data pre-processing: Cleaning, transformations and enrichment: On completing the annotation creation process, system performs a series of consistency checks to ensure that the data are as accurate as possible before any analysis occurs. In this system, consistency checks are designed to detect errors made during the annotation data collection process. A data error may be a missing action or an invalid event or any irrelevant action

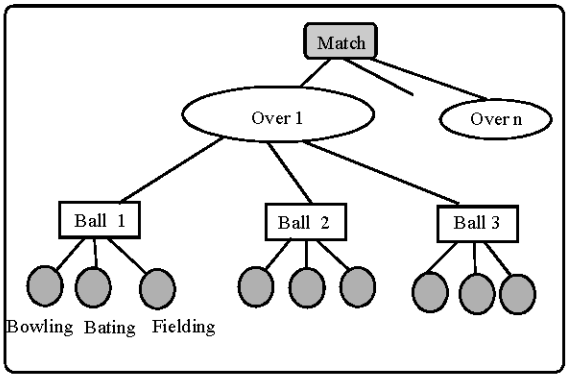

Fig. 1: Granularities in cricket data model

sequence. Corrections are made using a rule base and/or with the input of a domain expert (typically a coach). For example, filling the no entry columns by "Null" value, Indicating maiden over etc. After the consistency checks, the data are transformed and reformatted.

Structuring the data (data modeling)

Object-action model: This model is to describe the entire scenario of each event that occurs in the match with the objects involved with their name, its role, its attributes, the relationship among the entities(objects) those who had involved in a same event and the description of the action carried out by those entities in that particular event. So that we can get required information about each event occurrence simply and appropriately for further search and extraction process. For this, object-relational data model which encapsulates and relates the objects that are involved in a particular action is used. In a Cricket video, a ball sequence is consider to be an event. Each event consists of many actions. An action consists of many objects. The data model comprises different granularity levels as shown in Fig. 1.

The above said granularity levels are incorporated in the form of alphanumeric representation of the ball identifier which is used as the key attribute of the annotation data table and match master tables.

Video data structure

The data structure: {Ballid, action1, action 2, action 3, action 4, result}

Ballid: is 7-digit alphanumeric code which is automatically generated by the system based on the match, over and ball number details.

Action: For simplicity, the number of actions may be limited to 4 for bowling, bating and fielding, respectively. The fourth is for special occurrence if any. The action itself is a set which consists of attributes as follows

Action {{<Entity List1>, <Entity List 2>,
…..},Relation, Descriptionlist}

Entity: An entity in action is an object which plays that role described to yield certain reaction. Entity itself is a set comprising of three attributes as noted below.

Entity { *Entity Name, Role , Attribid* }

Entity name: The name of the entity represented as a string.

Role: Case role of entity (agent, instrument, At-loc, From- loc and To-poss}

Agent is the object that caused the event to happen.

Relationship: The relationship is the one between the entities in the entity list. This relation can be of two types :

Description list - It is the set of tupelos of conceptual descriptions of the relationship of the form :

Description {Desscription1- Description 2-
Description n}

The relationship, in most cases, needs further description. Whereas, the Descry-List denotes a list of much finer descriptions of the concept explained earlier.

An instance of the data structure Index, in the context of visual containing cricket matches, is given below:

Entity list : { <Bowler1, Agent,{id}>,
<Batsman1, To_Poss,{id}>}
Relationship : Bowls- to
Description : { inswinger- good-off stump-
outside }

Data generalization: To achieve transformation of ORDB model of the original cricket annotation data base, a simple generalization mechanism is discussed in this chapter (Jiawei *et al*., 1992; Jawai and Michel, 2002; Generalization-based data).

A simple generalization mechanism (Jiawei *et al*., 1992)

Selective retrieval: "Retrieve attribute A if it is task relevant"

The general idea of attribute focusing is to first collect the task relevant data using an object oriented query and then store them in a separate relational table that is constructed appropriately.

In this study, attribute retrieval technique is used to retrieve task relevant attributes, which leads to transform the ORDB model of source data into simplified Relational model. The source data set is scanned manually and the task relevant attributes are fixed for retrieval. Then those task relevant attributes are retrieved using queries

Algorithm Selective-Retrieval *(Cricket)*
Cricket : ORDB of cricket match
Gneralized-cricket : RDB of cricket match
Var : Tuple T, Attribute A
Repeat
for each attribute A in cricket
if A is relevant

add A to generalized-cricket
for each value of A
append Generalized- Cricket {
value of A,T }
until there are no more attribute is remaining for check.

Algorithm illustration: The structure of ORDB annotation table before generalization (Han *et al*., 1998)

The ORDB structure which is given in the previous section is taken for generalization. The structure of relational table after generalization as such:

Generalized-Cricket { RID, Player 1, Player 2
Player 3, Bowl-Desc, Bat-Desc,
Field- Desc, Result }

Identifier of any object cannot be generalized, so that it remains unchanged in generalized relation as RID. Entity names are generalized as player 1, player 2 and player 3. Description list of every action is shortened as a character sequence where the levels of description is represented by separation with hypen "-" . The result is given as it is in the original data base.

Thus the complex ORDB model is transformed into simple relational model to ease the mining process in which the same Apriori algorithm for relational data bases could be used in further processing to extract useful information.

Association rule mining: This study briefs how the mining operation is performed on the transformed and generalized model of cricket annotations with the aim to find out some interesting associative relationships among the data set by searching and analyzing frequent patterns. In this research the frequent item sets (patterns) will be the sequence of actions occur frequently for certain event in cricket. The association rule mining technique, known as *Apriori algorithm* is used to achieve this task. (Jiawei *et al*., 1992).

An *association rule* is a rule which implies certain association relationships among a set of objects (such as "occur together" or "one implies the other") in a database. Given a set of transactions, where each transaction is a set of literals (called items), an association rule is an expression of the form X => Y, where X and Y are sets of items. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y.

Algorithm: CRICKET-MINE,
Input: generalized annotation DB; min-supp- threshold, min-sup
Output: L, frequent item sets in D.
Method:
i = 0;
Ci = {{A} / A is a variable};
While Ci is not empty do
Database pass:
For each set in Ci, test whether it is frequent;
Let Li be the collection of frequent sets from Ci;
Candidate formations:
Let Ci+1 be those sets of size i + 1
Whose all subset are frequent;
End.

Rule representation: The objective of such a presentation is to ensure that a coach easily understands the results. The text presentation also offers a suggestion as to why the particular pattern is interesting - explicitly pointing out the ways that this particular pattern deviates from an expected norm - in essence presenting an initial argument and easily interpretable justification of Interestingness. Knowledge representation is an important task where high concentration is required. This is indulged to represent the discovered knowledge which is in the form of association rules in the form well understandable by the user who then after apply the for further interpretation to take valid decisions.

Automatically generated text describes the rule patterns as such

Type 1 : " A " occur "Confidence" % with "B"
Type 2 : "Player 1" made "confidence %" of his "A" attempts when "player 2" made "B"

A subsequent requirement of knowledge discovery would be for the domain expert to determine the

underlying cause of this pattern. The process of interpreting patterns represents knowledge discovery and traditionally requires.

RESULTS AND DISCUSSION

Ordb structure of cricket annotations

Select * from cricket
2 /

BALLID

ACTION1(ID, ENTITY1(ID, NAME, ROLE, ATTRIBUTE), ENTITY2(ID, NAME, ROLE, ATTRIBUTE), RELATION, DESCRIPTION (DESC1,...DESC N)

ACTION2(ID, ENTITY1(ID, NAME, ROLE, ATTRIBUTE), ENTITY2(ID, NAME, ROLE, ATTRIBUTE), RELATION, DESCRIPTION (DESC1,...DESC N)

ACTION3(ID, ENTITY1(ID, NAME, ROLE, ATTRIBUTE), ENTITY2(ID, NAME, ROLE, ATTRIBUTE), RELATION, DESCRIPTION (DESC1,...DESC N)

ACTION4(ID, ENTITY1(ID, NAME, ROLE, ATTRIBUTE), ENTITY2(ID, NAME, ROLE, ATTRIBUTE), RELATION, DESCRIPTION (DESC1,...DESC N)

Result A sample annotation:

"Waugh bowls an in swinger good length delivery to shewag Shewag missed that ball, the ball races towards the outsideboundary The keeper picks up the ball and it was declared as no ball."

A sample record

M10101

ACTION1 (1, ENTITY1(1, 'Waugh', 'Agent', '10'), ENTITY2(1, 'Sehwag', 'to poss', '20'), 'Bowls', DDESC('Good Length'))
ACTION2 (2, ENTITY1(1, 'Sehwag', 'Agent', '20'), ENTITY2(1, 'Balls', 'Instrument', '0'), 'Batting', DDESC('Missed'))
ACTION3(3, ENTITY1(1, 'Ball', 'Instrument', '0'), ENTITY2(1, 'Keepar', 'to-loc', '30'), DDESC('Outside Edge'))
ACTION4 (1, ENTITY1(1, 'null', 'null', 'null'), ENTITY2(1, 'null', 'null', 'null'), DDESC('null'))

No Ball

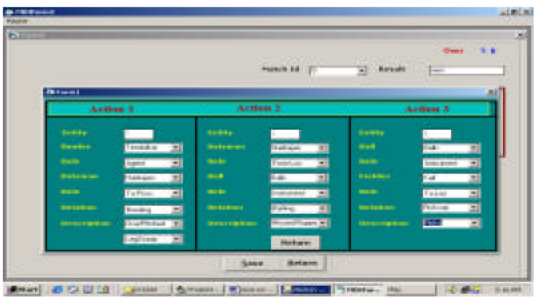

Fig. 2: Annotation creation interface (sample)

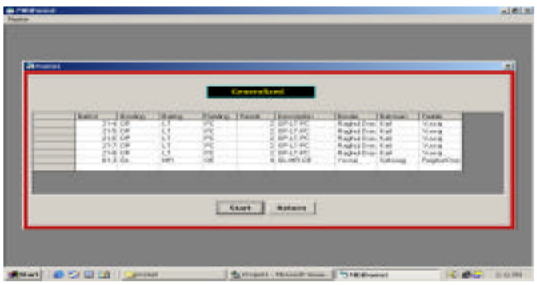

Fig. 3: A sample generalized table window

The annotation creation interface designed in VB is given in Fig. 2. and a sample generalized table window is shown in Fig. 3.

Association Rule Mining on Cricket data set (sample run): In this forum, the generalized set of cricket annotations are taken for mining operation to find all possible frequent patterns those which are sequence of action descriptions related to specific action type that are repeatedly occurred and same event sequence related with the same pair of players or with particular player.

Finding frequent item sets: Mining: From the input table presented above, with the constraints of Min-support = 0.49 and Confidence Threshold = 80 %, for an illustration purpose, Description field only taken for first level mining to discover description based knowledge is shown in Table 1-5.

Table 1: Initial candidate table

GL-HL-IS
OS-HL-YK
GL-OSHL-YK
OS-YK

Table 2: Frequent-1 candidate table

Candidate (C1)	Support	L 1
GL	0.5	Y
OS	0.75	Y
HL	0.75	Y
IS	0.25	N
YK	0.75	Y

Table 3: Frequent-2 candidate table

Candidate -C2	Support	L 2
GL-OS	0.25	N
GL-HL	0.5	Y
GL-YK	0.25	N
OS-HL	0.5	Y
OS-YK	0.75	Y
HL-YK	0.5	Y

Table 4: Frequent-3 candidate table

Candidate (C3)	Support	L 3
GL-HL-OS	0.25	N
GL-OS-HL-YK	0.25	N
GL-HL-YK	0.25	N
OS-HL-YK	0.25	Y

Table 5: Frequent-3 candidate table

Candidate (C4)	Support	L 4
OS-HL-YK	0.5	Y

Explanation: In the above illustration, the data base was scanned 4 times to find the frequent item sets of at all search levels with different combinations of candidate item, as mentioned in the algorithm. The candidate for which the support is not satisfactory i.e.) support < 49 %, are considered to be infrequent and they have been removed from the next iteration candidate list.

Generation of association Rules from frequent item sets(Jiawei *et al.*, 2002): Once the frequent item sets from records in a database have been found, it is straight forward to generate strong association rules from them by evaluating the confidence of each rule. Those rules which satisfies minimum confidence threshold are said to be strong and taken as useful.

$$\text{Confidence}(A => B) = P(B / A)$$

$$= \frac{\text{Support- count}(A \cup B)}{\text{Support} - \text{count}(A)}$$

In this illustration, the frequent item set found is,

$$L = \{ \text{OS-HL-YK} \}$$

Non- empty subsets are:

{OS-HL}, { OS-YK}, {HL-YK}, {OS}, {HL}, {YK}

The association rules are:

OS-HL => YK Confidence = 2/3 = 66%
OS-YK => HLConfidence = 2/3 = 66%
HL-YK => OSConfidence =2/2 = 100%
OS => HL-YK Confidence =2/3 = 66%
HL => OS-YK Confidence =2/3 = 66%
YK => OS-HL Confidence =2/3 = 66%

Though the minimum confidence threshold is 80 %, the rule HL-YK => OS only is the more interesting and strong to generate knowledge.

HL-YK occur 100 % with OS

The useful information behind this statement is,

Huge with Yorker style always
happens with "out swing"
bowling"

CONCLUSION

This study deals with a simple generalization technique called Attribute-Focusing and Association rule mining. The same approach may be extended with effective generalization rules and techniques like Attribute Oriented Induction and Data Cube construction to increase the efficiency of the mining process and henceforth provide a road map for mining on complex structures as video data. This can be extended for the video sequences of other games like basketball, football etc with necessary modifications and advanced data base concepts.

ACKNOWLEDGEMENT

I deliver my whole hearted thanks to my supervisor Dr.Raja Ram for his valuable guidance and encouragements to complete this research.

REFERENCES

Jiawei Han, Y. Cai and N. Cercone, 1992. "Knowledge Discovery in Data base: An attribute-oriented Approach.VLDB, Vancouver, Canada, pp: 547-559.

Rakesh Agrawal, Tomasz Imielinski and Arun Swami, 1993. Data Mining: a perspective, IEEE Journal on knowledge and Data Engg, Special issue on Learning and Discovery in Knowledge - Based Databases, 5: 914-925.

Bhandari, Edward and Group, 1997. Advanced Scout: Data Mining and Knowledge Discovery inNBA Data Data Mining and Knowledge Discovery, Kluwer Acad. Pub., 1: 121-125.

Jawai Han and Michel Hamber, 2002., The Concepts of Data Mining, TMH.Bhandari, I. (1995). Attribute Focusing: Data mining for the layman (Research Report RC 20136). IBM T.J. Watson Research Center.

Petr Kuba, 2002. Mining frequent patterns in object-relational databases. In Proceedings of the International Conference on Knowledge Based Computer Systems (KBCS), Mumbai, pp: 59-68.

Jan Bla´ák and Petr Kuba, 2003. Mining frequent patterns in complex structured data. In Proceedings of DATAKON, Brno, pp: 193-203.

Oracle8i SQLJ Developer's Guide and Reference Release 8.1.5 February, 1999 Part No. A64684-01 Han, J., S. Nishio, H. Kawano, and W. Wang, 1998.

Generalization-based data mining in object-oriented databases using an object-cube model. Data and Knowledge Engineering, 25: 55-97.