International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-1 Number-1 Issue-1 September 2011

A Survey on Close-degree of Concept Lattice and Attribute Reduction in Data Mining Services

Sapna Sahu[1], Surendra Mishra[2]
M.Tech Scholar, Department of Computer Science, SSSIST Sehore, India[1]
Head, PG Department of Computer Science, SSSIST Sehore, India[2]

Abstract

With recent technical advances in processing power, storage capacity, and inter-connectivity of computer technology, data mining is seen as an increasingly important tool by modern business to transform unprecedented quantities of digital data into business intelligence giving an informational advantage. The manual extraction of patterns from data has occurred for centuries. Concept lattice is a new mathematical tool for data analysis and knowledge processing. Attribute reduction is very important in the theory of concept lattice because it can make the discovery of implicit knowledge in data easier and the representation simpler. In this paper we discuss and analyze some close-degree of concept lattice and attribute reduction technique for applying data mining services which is fruitful for many applications and business processing.

Keywords

Data Mining, Concept Lattice, Attribute Reduction, Close-degree

1. Introduction

Data mining in customer relationship management applications can contribute significantly to the bottom line. Rather than randomly contacting a prospect or customer through a call center or sending mail, a company can concentrate its efforts on prospects that are predicted to have a high likelihood of responding to an offer. More sophisticated methods may be used to optimize resources across campaigns so that one may predict to which channel and to which offer an individual is most likely to respond across all potential offers. Additionally, sophisticated applications could be used to automate the mailing. Once the results from data mining are determined, this "sophisticated application" can either automatically send an e-mail or regular mail. Finally, in cases where many people will take an action without an offer, uplift modeling can be used to determine which people will have the greatest increase in responding if given an offer. Data clustering can also be used to automatically discover the segments or groups within a customer data set. Data Mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. The theory of concept lattice [1], proposed by Wille in 1982, is a new mathematical tool for data analysis and knowledge processing, which has been successfully applied in knowledge engineering, data mining, information retrieval and other fields [2-5].

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.

With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments.

For example, Blockbuster Entertainment mines its video rental history database to recommend rentals to individual customers. American Express can suggest products to its cardholders based on analysis of their monthly expenditures.

WalMart is pioneering massive data mining to transform its supplier relationships. WalMart captures point-of-sale transactions from over 2,900 stores in 6

countries and continuously transmits this data to its massive 7.5 terabyte Teradata data warehouse. WalMart allows more than 3,500 suppliers, to access data on their products and perform data analyses. These suppliers use this data to identify customer buying patterns at the store display level. They use this information to manage local store inventory and identify new merchandising opportunities. In 1995, WalMart computers processed over 1 million complex data queries.

Knowledge reduction is an important aspect of knowledge discovery. In [6], the attribute reduction of concept lattice was put forward, which maintained the epitaxial of concept unchanged. Now, the attribute reduction algorithms of concept lattice are mostly based on the discernibility matrix.

We provide here an overview of executing data mining services. The rest of this paper is arranged as follows: Section 2 introduces Concept Lattice; Section 3 describes about Attribute Reduction; Section 4 shows the evolution and recent scenario; Section 5 describes the challenges. Section 6 describes Conclusion and outlook.

2. Concept Lattice

Conceptual Knowledge Discovery in Databases (CKDD) has been developed in the field of Conceptual Knowledge Processing. Based on the mathematical theory of Formal Concept Analysis, CKDD aims to support a human-centered process of discovering knowledge from data by visualizing and analyzing the formal conceptual structure of the data.

The overall theme and contribution of the volume "Advances in Knowledge Discovery and Data Mining" [7] is a process-centered view of KDD considering KDD as an interactive and iterative process between a human and a database that may strongly involve background knowledge of the analyzing domain expert. In particular, R. S. Brachman and T. Anand [8] argue in favor of a more human-centered approach to knowledge discovery support referring to the constitutive character of human interpretation for the discovery of knowledge and stressing the complex, interactive process of KDD as being led by human thought. Following Brachman and Anand, CKDD pursues a human-centered approach to KDD based on a comprehensive notion of knowledge as a part of human thought and argumentation. In addition, certain objects have certain attributes; in other words, objects are related to attributes. Taken together, the set of objects, the set of attributes and the relation defined among the objects and attributes is known as a formal context (Wormuth & Becker, 2004). Let's consider an example based on one given by Lienhard, Ducasse, and Arevalo (2005). Consider a group of people {Sigmund, George, Anna, Fritz, Carl, Fay}, and a set of beverages {beer, orange juice, tea, wine, coffee}. We can refer to the set of people as the objects and the set of beverages as the attributes or properties. We can ask the question which of these beverages people prefer. The set of concepts in forms a structure known as a complete partial order. A partial order is one way of formally representing or modelling hierarchy in a dataset. Lienhard, Ducasse, and Arevalo (2005, p.75) define a context C as the triple (O, A, R), where O and A are the sets of objects and attributes, and R is a binary relation between O and A. Let X be a subset of O and Y a subset of A, where $\sigma(X)$ represents all the attributes common to X, and $\tau(Y)$ represents all the objects common to Y. A concept is defined as the pair (X,Y) such that $Y = \sigma(X)$ and $X = \tau(Y)$. Further, a concept (X_1, Y_1) is a subconcept of concept (X_2, Y_2) if X_1 is contained in X_2 or equivalently if Y_2 is contained in Y_1 (Lienhard, Ducasse, & Arevalo, 2005, p.75). Likewise we can define a concept (X_1, Y_1) as a superconcept of a concept (X_2, Y_2) if the inverse properties hold (Lienhard, Ducasse, & Arevalo, 2005, p.75). Given these relations, the set of concepts for a context can be represented as a concept lattice. A concept lattice for the concepts is presented in Figure 1.

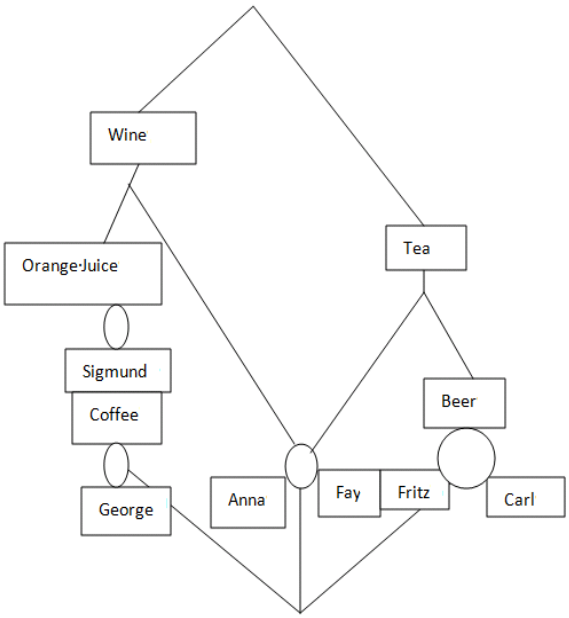

Figure1: Concept Lattice

3. Attribute Reduction

Knowledge discovery processes unleash a powerful medium to further strengthen the business intelligence program of any enterprise. However, the key to successful knowledge discovery processes lies in strong commitment from the BI leadership. Here

are some key requirements for a successful knowledge discovery program:

There should be a strong commitment from the senior management and a strong sponsorship for the knowledge discovery program.

- The business case for the application is clearly understood and measurable, and the objectives are likely to be achievable given the resources being applied.
- The application has a high likelihood of having a significant impact on the business.
- Business domain expertise is available.
- Good quality relevant data in sufficient quantities is available.
- The right people - domain, data management and data mining experts - are available.

If some of the condition attribute values in an information system are unknown, missing or knowing partially, then such an information system is called an incomplete information system and is still denoted with convenience by the original notation S = (U, A, V, f) . That is, if there exists at least an attribute aÎ B Í A , such that a V includes null values, then the system is called an incomplete information system, the sign * usually denotes null value; otherwise the system is called a complete information system.

In order to make some approximation, we collect similar objects from the set O to form a subset and name it as a granule. In a granule, one object is regarded as the same as the others because the inherent difference between two objects disappears when they are assigned to the same granule. The information system may be divided or covered by the set consisted of these granules, which gives an approximation to the IS and can be named as a granule view of it.

Attributes reduction is useful in such type of granual reduction. For an information system IS=(O,AT), each subset of attributes A is the subset of AT determines an indiscernible relation IND(A) as follows:

$$IND(A) = \{(x, y) \in O \times O \mid a \in A, a(x) = a(y)\}$$

A is the subset of AT is a possible reduce of DT for x, Ox∈, if and only if A is a set such that

$$I_A(x) \subseteq \overline{AT}_{Xd(x)}$$

A is called minimal possible reduce, if it is the minimal set satisfied Eq.(1). With the decrease of the cardinality of A, $I_A(x)$ enlarged. However, Eq.(1) sets an upper bound for $I_A(x)$ to confine the granule which contains object x.

Attribute reduct simplifies an information system by discarding some redundant attributes. In the view of approximation, the information system is reduced to a granule view of it. However, it is necessary to analyze the approximation by these reduces.

By an information system we mean a pair S=(U, A), where U and A are finite, non-empty sets called the universe and a set of attributes respectively. With every attribute a∈A, we associate a set V a of its values, called the domain of a. Any subset B of A determines a binary relation ind (B) on U, which will be called an indiscernibility relation and is defined as follows:

(x,y)∈ind(B) if and only if f (x, a)=f(y, a) for every a∈B, where f(x,a) denotes the value of attribute a for element x. It can be seen that ind(B) is an equivalence relation. The family of all equivalence classes of ind(B), i.e. the partition determined by B, will be denoted by U/ind(B), and U/ind(B) is defined as follows:

U/ind(B)={[x] B : x∈U}, where [x] B ={y: (x,y)∈ind(B)} is a equivalence class for an example x with respect to concept B.

The indiscernibility relation will be used next to define two basic operations in rough set theory as follows:

$$B_*(X) = \bigcup \{Y \in U/ind(B) \mid Y \subseteq X\},$$
$$B^*(X) = \bigcup \{Y \in U/ind(B) \mid Y \cap X \neq \Phi\}$$

4. Evolution and Recent Scenario

In 2006, Xiaobing Pei et al. [9] proposed about approach to attribute reduction. The attribute reduction is one of key processes for knowledge acquisition. They proposed with approximate approach to attribute reduction. The concept of minimal discernible attributes set is introduced and a calculation method for it is investigated. And then, the judgment theorem with respect to keeping positive region invariability is obtained, from which an approximate approach to attribute reduction.

In 2000, Chih-Yung Chang et al. [10] proposed about algorithm which analyzes locality reference space for each reference pattern, partitions the multi-level cache into several parts with different size, and then maps array data onto the scheduled cache positions such that cache conflicts can be eliminated. To reduce the memory overhead for mapping arrays onto partitioned cache, a greedy method for rearranging array variables in declared statement is also developed. Besides, we combine the loop tiling and the proposed schemes for exploiting both temporal

and spatial reuse opportunities. To demonstrate that our approach is effective at reducing number of cache conflicts and exploiting cache localities, we use Atom as a tool to develop a simulator for simulation of the behavior of direct-mapping cache.

In 2008, Jen-Wei Huang et al. [11] proposed about a progressive algorithm Pisa, which stands for Progressive mining of Sequential patterns, to progressively discover sequential patterns in defined time period of interest (POI). The POI is a sliding window continuously advancing as the time goes by. Pisa utilizes a progressive sequential tree to efficiently maintain the latest data sequences, discover the complete set of up-to-date sequential patterns, and delete obsolete data and patterns accordingly. The height of the sequential pattern tree proposed is bounded by the length of POI, thereby effectively limiting the memory space required by Pisa that is significantly smaller than the memory needed by the alternative method, Direct Appending (DirApp).

In 2010, Shiow-yang Wu et al. [12] proposed about a complex activity is modeled as a sequence of location movement, service requests, the co-occurrence of location and service, or the interleaving of all above. An activity may be composed of subactivities. Different activities may exhibit dependencies that affect user behaviors. They argue that the complex activity concept provides a more precise, rich, and detail description of user behavioral patterns which are invaluable for data management in mobile environments. Proper exploration of user activities has the potential of providing much higher quality and personalized services to individual user at the right place on the right time.

In 2010, Nitin S. Sharma et al. [13] proposed about clustersbased freight data mining strategy using socio-economic variables to aggregate the most granular representations of freight flows called Traffic Analysis Zones (TAZs) in the Mobile Metropolitan Area (MMA). Such aggregation results in an intermediate level of freight distribution between the county and traffic zone levels, at a resolution meaningful for metropolitan-level planning.

In 2010, Huili Meng et al. [14] proposed about the reduction of the concept lattice. First, they present a close-degree of concept to measure the close-degree of two concepts with the attributes reduction. Based on the close-degree of two concepts, we propose the close-degree of concept lattice. Then they use the close-degree of concept lattice as heuristic information and design an attribute reduction algorithm. The reduction algorithm attempts to get one reduction of the attribute set of concept lattice. Last, they give an application example for proving the validity of the algorithm.

5. Challenges

Besides utilizing links in data mining, we may wish to predict the existence of links based on attributes of the objects and other observed links in some problems. Examples include predicting links among actors in social networks, such as predicting friendships; predicting the participation of actors in events, such as email, telephone calls and co-authorship; and predicting semantic relationships such as \advisor-of" based on web page links and content.

Another important direction in information network analysis is to treat information networks as graphs and further develop graph mining methods. Recent progress on graph mining and its associated structural pattern-based classification and clustering, graph and graph containment indexing, and similarity search will play an important role in information network analysis.

Many domains of interest today are best described as a network of interrelated heterogeneous objects. As future work, link mining may focus on the integration of link mining algorithms for a spectrum of knowledge discovery tasks. Furthermore, in many applications, the facts to be analysed are dynamic and it is important to develop incremental link mining algorithms.

6. Conclusion and Outlook

With recent technical advances in processing power, storage capacity, and inter-connectivity of computer technology, data mining is seen as an increasingly important tool by modern business to transform unprecedented quantities of digital data into business intelligence giving an informational advantage. The manual extraction of patterns from data has occurred for centuries. Concept lattice is a new mathematical tool for data analysis and knowledge processing. Attribute reduction is very important in the theory of concept lattice because it can make the discovery of implicit knowledge in data easier and the representation simpler. In this paper we discuss and analyze some close-degree of concept lattice and attribute reduction technique for applying data mining services which is fruitful for many applications and business processing.

In future we can analyse the aspects with their practical implementations and also the simulation

result which shows the advantage graph of our method.

References

[1] Wille R. Restructuring lattice theory: an approach based on hierarchies of concepts. In: Rival I, ed. Ordered Sets. Reidel: Dordrecht-Boston, 1982, 445-470.

[2] Oosthuizen G D. The Application of Concept Lattice to Machine Learning. Technical Report, University of Pretoria, South Africa, 1996.

[3] Mineau G W, Godin R. Automatic structuring of knowledge bases by conceptual clustering. IEEE Trans Knowledge Data Eng, 1995, 7(5): 824-829.

[4] Ho T B. Incremental conceptual clustering in the framework of Galois lattice. Lu H, Motoda H, Liu H.KDD: Techniques and Applications. Singapore: World Scientific, 1997, 49-64.

[5] Kent R E, Bowman C W. Digital Libraries, Conceptual Knowledge Systems and the Nebula Interface. Technical Report, University of Arkanaas, 1995.

[6] Zhang Wenxiu, Wei Lin, Qi Jianjun. Attribute reduction theory and approach to concept lattice. Science in China Series E-Information Science, 2005, 35(6): 628-639.

[7] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (eds.): Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, Cambridge 1996.

[8] Brachman, Ronald J., and Tej Anand. "The process of knowledge discovery in databases." In Advances in knowledge discovery and data mining, pp. 37-57. American Association for Artificial Intelligence, 1996.

[9] Xiaobing Pei, YuanZhen Wang," An Approximate Approach to Attribute Reduction", International Journal of Information Technology, Vol. 12 No.4 2006.

[10] Chih-Yung Chang, Jang-Ping Sheu and Hsi-Chiuen Chen," Reducing Cache Conflicts by Multi-Level Cache Partitioning and Array Elements Mapping", 2000 IEEE.

[11] Jen-Wei Huang, Chi-Yao Tseng, Jian-Chih Ou, and Ming-Syan Chen," A General Model for Sequential Pattern Mining with a Progressive Database", IEEE, 2008.

[12] Shiow-yang Wu, and Hsiu-Hao Fan, "Activity-Based Proactive Data Management in Mobile Environments", IEEE march 2010.

[13] Nitin S. Sharma, Gregory A. Harris, Michael D. Anderson, Phillip A. Farrington and James J. Swain," Freight Data Mining Strategy Using Socio-economic Variables For Metropolitan Planning", 2010 IEEE International Conference on Granular Computing.

[14] Huili Meng, Jiucheng Xu, "The Close-degree of Concept Lattice and Attribute Reduction Algorithm Based on It", 2010 IEEE International Conference on Granular Computing.