*Advanced Engineering Forum*
*ISSN: 2234-991X, Vols. 6-7, pp 625-630*
*doi:10.4028/www.scientific.net/AEF.6-7.625*
*Online: 2012-09-26*

# Construction Search Engine Based on Formal Concept Analysis and Association Rule Mining

HongSheng Xu[1, a]

[1]College of Information Technology, Luoyang Normal University, Luoyang,471022, China

[a]xhs_ls@sina.com

**Keywords:** formal concept analysis; association rule; data mining; search engine

**Abstract.** In the form of background in the form of concept partial relation to the corresponding concept lattice, concept lattice is the core data structure of formal concept analysis. Association rule mining process includes two phases: first find all the frequent itemsets in data collection, Second it is by these frequent itemsets to generate association rules. This paper analyzes the association rule mining algorithms, such as Apriori and FP-Growth. The paper presents the construction search engine based on formal concept analysis and association rule mining. Experimental results show that the proposed algorithm has high efficiency.

## Introduction

Traditional search engine technology to meet certain needs of the people, but because of its universal nature, still can not satisfy different backgrounds, different professions, and different periods of the information retrieval. Also, the search engine returns the results of the query words are a huge number of user needs is just one little part, in general, users rarely turn many pages are turned the first few pages, so the user clicks on the URL strong locality. So, how to provide users with more accurate and more effective results have been efforts in the direction of search engine development.

This philosophy based on the concept of understanding, the German Wille, R Professor of formal concept analysis for the discovery of the concept, sort and display. In the form of conceptual analysis, the extension of the concept is understood as the collection of all objects belonging to this concept, but the connotation is considered to be characteristics common to all of these objects (or attributes) set to achieve the concept of philosophy in an understandable form of it. The concept lattice structure is in order to reflect a complete contact and the concept of generalization and case relations between the objects and attributes in the concept of hierarchy.

Association rule mining in large amounts of data to find interesting association or contact between the itemsets is an important topic in the research of KDD (Knowledge Discovery in Database). With the large amounts of data constantly collect and store a lot of people in the industry are increasingly interested in mining association rules from their databases. The paper presents the construction search engine based on formal concept analysis and association rule mining.

## Formal concept analysis model

The study of formal concept analysis involves four main aspects: the study of basic theory, the concept lattice generation method, the visualization of the concept lattice and the applied research of the concept lattice. The formal concept analysis as a formal mathematical methods, and artificial intelligence, it is database technology, software engineering, computer science, closely linked, but relatively independent. At present, the theory of formal concept analysis has been successfully applied to software engineering, data mining, information retrieval and other fields.

Concept lattice as the core data structure in the theory of formal concept analysis in software engineering, knowledge discovery, cluster analysis, rules, Web knowledge discovery and information retrieval, a variety of semantic analysis involving data field has been widely used.

The study of formal concept analysis involves four main aspects: the study of basic theory, the concept lattice generation method, the visualization of the concept lattice and the applied research of the concept lattice. At present, the theory of formal concept analysis has been successfully applied to software engineering, data mining, information retrieval and other fields.

Definition 1 a formal context K: = (G, M, I), composed by the set G, M, and the relationships between them, the elements of G are called objects (objects), the elements of M are called attributes (Attributes).

Set <H, £> posets for arbitrary B ⊆ H, if there is a∈B is an arbitrary element x, satisfy x £ a, called the upper bound of a subset of B. Similarly, if B is any element x, meet a £ x, called the lower bound of a subset of B. If the posets <H, £> each element are a formal concept, the partial order focus on any one node of the upper bound is called the super-concept of the node, the lower bound is called the node of the sub-concepts. All these sub-concepts - the super-concept of the relationship between the composition form of the concept of partial order set, as is shown by equation 1.

$$\bigvee_{g\in A}\left(\{g\}'',\{g\}'\right)=\bigwedge_{m\in B}\left(\{m\}',\{m\}''\right), \tag{1}$$

The concept lattice can be a graphical representation of the labeled line graph (lapelled line diagram), also known as the Hasse diagram of concept lattice. The map was generated as follows: if C1 <C2, and the grid element C3 makes the C1 <C3 <C2, then there exists an edge from C1 to C2. Node in the line graph representation of concepts line up the Asian concepts - ultra-concept relationship. For an object, if C is the smallest concept that contains the object, the name of the object to be attached to the C corresponding to the node. For a feature, if C contains the characteristics of concept, the characteristics of the name were attached to the C corresponding to the node. Concept lattice label chart is often used as a mode of communication, which makes the concept of the background of a given data structure becomes clear and easy to understand, in order to achieve the visual display of the concept lattice.

Set (A, £) is a partially ordered set if for any non-empty collection S ⊆ A, there are ∨ S-(A, £) is called a semi-lattice, similar to, if for anynon-empty set A⊆ S there ∧ S (a, £) be called a complete intersection semi-lattice. Both fully and semi-lattice (A, £) is completely cross-semi lattice, then it is a complete lattice.

L (K1)∈ Ox, D1) ∪C1 = (O3 and C2 = (O3 ∪ Oy, D2), ∈ L (K2) and satisfies D1, ∪ D2 = D3 and Ox ∩ Oyo, = ϕ, the K1 D1 makes (O3) = D1 (any less than C1 grid node C ′ ′C 1 is equal to or greater than the C3 to know) and g (D1) = of O3 ∪ of Ox, in K2 D2 makes f (O3) = D2 (any less than the C2 grid node C ′, also 2′C is equal to or greater than the known C3) and g (D2) = of O3 ∪ O y, D1 ∪ D2, K1 + K2 =D3 meet f (O3) = D3 and g (D3) = of O3, C3 = (O3 D3) ∈ L. (K1 + K2).

Conceptual analysis on the background of the multi-valued multi-valued background is in order to a single value of the background. It is easy to some technical background of the multi-valued form to convert the single value in the form background. Background conversion of multi-valued single-valued background in two ways - the concept of scaling conceptual scaling, it is the logical scaling logical scaling which is to complete, as is shown by figure1.

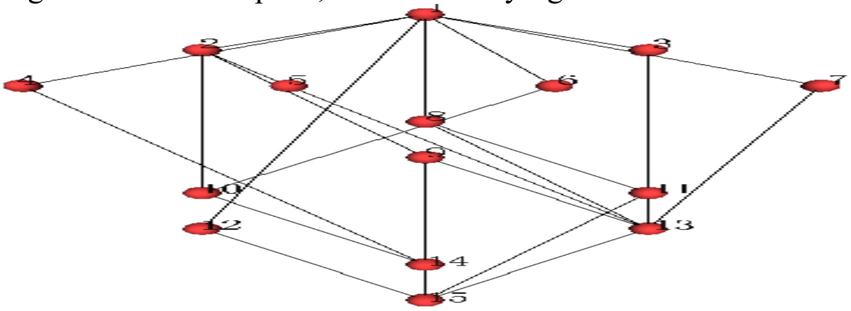

Figure. 1 Concept lattices Hasse diagram example

The concept lattice process is actually the link between a concept and the concept formation process. Therefore, the concept lattice, making the grid algorithm has a very important position. For the same data, generated by the grid is unique, ie, not the data or attributes of the order of one of the advantages of this concept lattice. The concept lattice construction algorithm method can be divided into two categories: traditional serial construction algorithm and the rise in recent years, parallel construction algorithm. Traditional serial algorithm can be divided into a batch algorithm and the incremental construction algorithm.

Let the transaction Tr has the item set T to be inserted into the lattice L in newborn grid node, if a grid node N1 = (C1, D1) to meet: (1) Intersection = D1 ∩ T-and L in any of the node N2Intent to have (N2) the ≠ Intersection; (2) for any L meet the N3> N1 node N3, Intent to (N3) ∩ T ≠ Intersection; then N1 is called a node of a lattice can be generated by the N1a newborn grid node (C = C1 +1, Z = D1 ∩ T).

Formal context by the set of objects, it is attributes and two binary relations, which is the basis of the concept lattice. The amount of data of different forms of background may be different, corresponding to the speed of the concept lattice are also different. How to efficiently construct the concept lattice is one of the most concerns undoubtedly one of the best ways to reduce the amount of data of the formal context, as is shown by figure2.

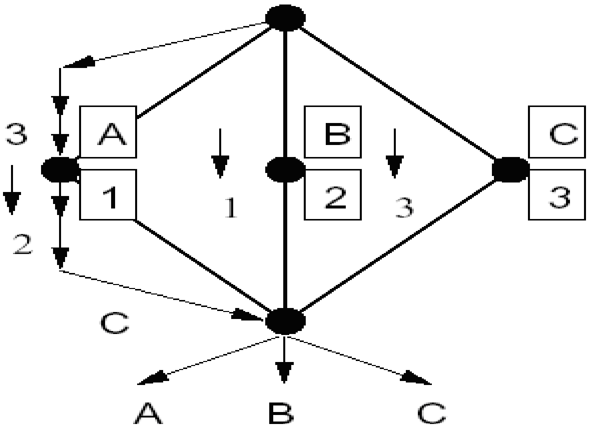

Figure. 2 Concept lattices and Formal context example

Concept lattice in addition to the concept of classification and definition from the data, it can also be used to find dependencies between objects and attributes. This has two meanings: (1) scan some or all of the lattice structure generates a rule set that can be used in the future; (2) browse the lattice structure in order to test a given rule is established.

Batch algorithm, the idea of the incremental construction algorithm is similar. The basic idea is you want to insert the object and the concept of the grid intersection, take different actions based on the results submitted. One of the most typical algorithms is the Godin algorithm.

Incremental construction of concept lattice is given the original formal context K=(X, D, R), corresponding to the original concept lattice L and the new object x * solving the formal context K*=(X∪{x*}, D, R), corresponding to the lattice L *. In the incremental generated concept lattice the process of solving the problem to be solved are mainly three: (1) the generation of the entire new node; (2) to avoid duplication of generation of the existing grid node ;( 3) side of the update. In order to effectively address these three problems, each node in the original concept lattice, according to the relationship between the connotation description and the new object, you can define the different types of it.

## Association rule mining algorithm

Association rules in data can be divided one-dimensional and multidimensional. One-dimensional association rules, we only involve one-dimensional data, such as the user to purchase items; multidimensional association rules, data to be processed will involve more than one dimension.

Replaced by another sentence, one-dimensional association rules to deal with some of the relationship of the individual properties; multidimensional association rules to deal with some of the relationships between the various properties., the support and confidence is defined as equation 2.

$$Support(A \Rightarrow B) = P(A \cup B)$$

$$Confidence(A \Rightarrow B) = P(B \mid A) \tag{2}$$

Single-dimensional association rules show that contact (the same attribute or dimension within the association); within the dimensional and multidimensional association rules show that the dimension between the contact (the relationship between the property and dimension).

These e-shopping site to use the rules of the association rules mining, and then set the user intends to purchase with the bundle. There are some shopping sites use them to set the corresponding cross-selling is to buy a commodity customers will see the addition of a commodity advertising.

The association rule mining algorithm considers the following two: (1) to reduce I / O operations. Association rule mining of data sets are sometimes up to GB and even TB magnitude, frequent I / O operation is bound to affect the efficiency of mining association rules, reduce I / O operation is to reduce the number of scan data set D; (2) reduce the need to calculate the number of support set (usually called the candidate set), with the frequent item sets the number of close. Reduction in the number of candidate itemsets can save computing time and storage space required to address some of the candidate itemsets, as is shown by equation3.

$$confidence(A \Rightarrow B) = P(B \mid A) = \frac{Support\_count(A \cup B)}{Support\_count(A)} \tag{3}$$

We can observe that a good modeling of the relationship between the item sets can be used concept lattice. To set a concept lattice grid node (X, Y) if X is as the transaction sets, Y as the item sets, a concept lattice grid node itemset Y transaction sets. In fact, the rule mining, the contents of X is not important; important is the number of X. On this basis, we will design the grid nodes in the concept lattice into the data structure as follows: Y is used to store node set of connotations, that is, itemsets, C is used to the extension of the number of storage nodes.

The researcher proposed a fundamentally different from the Apriori algorithm does not produce the candidate set of frequent itemsets generation algorithm, called the frequent pattern growth (frequent pattern growth), or referred to as the FP-growth, using a special data structure FP-tree storage transaction database information, the maximum degree of compression of data and ensure data completeness, to generate frequent itemsets by the FP-tree traversal algorithm than in the past has greatly improved efficiency. FP-growth to take the following divide and conquer strategy: provide frequent itemsets database is compressed into a frequent pattern tree (or FP-tree), but still retains the itemset associated information; then this compressed database into a set of conditions database (a special type of projection database), each associated with a frequent item, and were digging for each database.

## Construction search engine based on formal concept analysis and association rule mining

Pages of information or text information stored in the database of search engine system, we can think that it is constituted by one or more background. Then we can put the organization of knowledge in the database into one or more forms of background. Through the application of the concept lattice construction algorithm, we will be able to get the corresponding knowledge of the concept lattice [3].

This step of the completion of the concept lattice structure, the semantic correlation calculated preparations. The concept lattice construction algorithm used in this paper is an improved algorithm Godin algorithm [4]. The algorithm by a number of experiments and the proof of the use of multiple projects, in the concept lattice structure is correct and efficient. Algorithm in the literature is mainly reflected in two aspects of the algorithms and data structures: first, use a depth-first graph traversal algorithm to ensure that will not produce the side of the condition is not met, and to reduce traverse times; by increasing the description of the grid nodes connotation of a collection of the number of elements, and parent node information, auxiliary find newborn grid nodes child nodes. The data mining algorithms based on improved-Godin and Apriori are following.

```
IF inf(L) = $(\phi, \phi)$ THEN;
for ($k=2; L_{k-1} \neq \varnothing; k{+}{+}$) {
  if $l_1[1]=l_2[1] \wedge l_1[2]=l_2[2] \wedge \ldots \wedge l_1[k-2]=l_2[k-2] \wedge l_1[k-1]<l_2[k-1]$;
  for(i=0;i< ColNode.Length;i++) do
   {Intent(inf(L)) $\leftarrow$ Intent(inf(L)) $\cup$ $f^*(\{x^*\})$
  $C_k = \{X \bowtie Y \mid X \in L_{k-1}, Y \in L_{k-1};\ X(i)=Y(i),\ i=1,2,\ldots,k-2;\ X(k-1)<Y(k-1)\}$
  Rules[N] := GenerateRulesForNode(N);
  R := R$\cup$Rules[N];
  $L_k = \{c \in C_k \mid WSup(c) \geq wminsup\}$
  Intersection= $f(x^*) \cap$ Intent(Ci)
  return R1 := (R1$-$Gk1)$\cup$Gk$\cup$Gn;
```

Windows XP operating system using Visual C#.net2010 to achieve the above rule set and computing algorithm. For randomly generated data sets, the relationship between the probability of 25%, the number of attributes is 50, we do the extended test, the number of objects in increments of 352, recorded by the sub-formal context, The paper presents the construction search engine based on formal concept analysis and association rule mining, FCA-Apriori compared to FP-Growth the results shown in Figure 3.

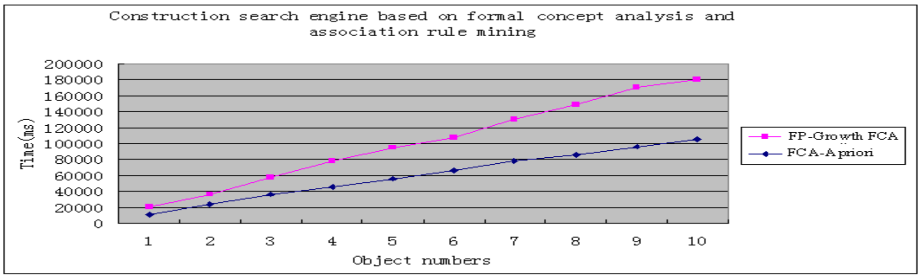

Figure. 3 The search engine based on FCA-Apriori compared to FP-Growth map

The purpose of the experiment the overall efficiency of the contrast to the concept lattice of the two access methods - is isomorphic to generate and direct construction - until the final concept lattice, including all the time overhead in the process from beginning to enter the formal context. Isomorphic to generate third-order form of the background of nuclear as a support, directly constructed using the Godin algorithm. The concept lattice model used as a data organization and formalization of data analysis tools; both theoretical research and practical applications are of great significance, it has a broad and successful application in many fields.

## Summary

The paper presents the construction search engine based on formal concept analysis and association rule mining. Formal Concept Analysis The starting point for research is a form of background, in the form of background is the formal description of the background knowledge. By the formal context can be a certain method of calculating the formal concept. The properties in the form of background can be a single value can also be multi-valued, so that the form of the background is also single-valued and multi-value. Generally during the multi-value form of background research, through a certain transformation rules to the formal context of the multi-value converted to single-valued formal context.

## References

[1] Ganter B, Wille R. Formal Concept Analysis:Mathematical Foundations [M]. Springer Verlag , Berlin, 1999

[2] C.H.Cai, W.C.Fu Ada, C.H.Cheng, W.W.Kwong. Mining association rules with weighted items . In Proc. of the Int’l Database Engineering and Applications Symposium. 1998. 68～77

[3] Ho T B. Incremental conceptual clustering in the framework of Galois lattice, in KDD: Techniques and Applications, H. Lu, H. Motoda and H. Luu (Eds.), World Scientific,1997. 49-64.

[4] Wang L-D., Liu X-D. Concept analysis via rough set and AFS algebra Information Sciences 2008, 178(21), 4125-4137.

[5] Anderson, B. and Moore, A.. Adtrees for fast counting and for fast learning of association rules. KDD-98. USA. 1998

[6] Faïd M, Missaoui R, Godin R. Knowledge discovery in complex objects. Computational Intelligence, 1999, 15(1): 28-49