Knowledge View Based on Rough Set and Similarity

Mingjian ZHOU[1, a], Qiang LIAO[1, b],Juncai TAO[2,c]

[1]Department of Computer Science and Technology, Nanchang University, Nanchang Jiangxi, China

[2]Computer Center, Nanchang University, Nanchang Jiangxi, China

[a]email: zhoumingjian@ncu.edu.cn, [b]email:liaoqiang@ncu.edu.cn, [c]email:taojuncai@ncu.edu.cn

Keywords: view; knowledge item; rough set; similarity; knowledge management

Abstract. View maintenance is a hot topic in database. But unlike database view, the knowledge view of knowledge management system has its particularity. The knowledge view is a cache of knowledge item that can be access quickly. Based on rough set, this paper proposes an approach of constructing knowledge view. The approach divides the knowledge set into three region using rough set theory, then reduces the three region by calculating interest similarity of every knowledge item, and finally the knowledge view will be constructed by filtering the knowledge item of merged region that consists of the three reduced region. The experiment results show that the approach can construct knowledge view efficiently.

Introduction

With the development of information technology and the growing popularity of Internet, much more knowledge-intensive corporations have raised concern about Knowledge Management (KM) to increase their competitive ability, since KM is regarded as the formal management of knowledge for facilitating creation, accessing, and reuse of knowledge, typically using advanced technology.

Nowadays, all KM System (KMS) have a large amount of knowledge and are updated continually. This makes user have adequate knowledge to use, but in the same time causes trouble to user. In general, each user has constant interest in a period of time. To solve a problem, the user retrieves many knowledge items from KMS and navigates some of them which he thinks is available to him. When he enters KMS again for the same problem, to avoid the troublesome retrieval, he wishes that KMS will automatically provide the same knowledge to him. On the other hand, if the user viewed a knowledge item that is transferred from remote database, when he want to view the same knowledge item again, he wish to view it directly instead of time-consuming transferring from remote database again. These demands can be satisfied with the concept of view.

Traditionally, a view is defined as a function from a set of base tables to a derived table and the function is recomputed every time the view is referenced. On the other hand, a view is like a cache of data that can be accessed quickly. So views are useful in applications such as data warehousing, replication servers, data recording system, data visualization and mobile systems [1-3]. But, because the object of KMS is different from these applications, the approach of view management in database can not been used in KMS directly. So, developing an approach of view management in KMS is necessary.

[4] proposes an approach of creating user view in KMS based on user navigation map. It used navigation map and no-loop map to generate judge rule which is used in creating the user view. This approach can create user view that is combined with user interest, but it is limited to the knowledge structure defined in [4]. Other researchers [5,6,7] propose many approaches of view maintenance, but these approaches are not suitable for KMS. So, based on rough set and similarity, this paper proposes a knowledge view constructing approach. This approach can construct knowledge view without the limitation of particular knowledge structure.

© 2016. The authors - Published by Atlantis Press

Organization of the Text

In order to comprehensive the approach proposed in this paper, relative conceptions are introduced as follows.

Rough Set Theory. Rough set theory[8,9] combines the classification and knowledge together, and it can portray approximately the uncertain or imprecise knowledge based on the known knowledge in database. Here are a few related basic concepts.

(1) Suppose S= (U, A, V, f) is a information system, where U is a non-empty finite set called universe including all elements, A is a non-empty finite set with attribute for each element, V is the range of A ($V=\bigcup_{a\in A} V_a$, V_a is named the range for a∈A), f is a mapping function, f:U×A→V, for $\forall u \in U$ and $\forall \alpha \in A$, we have f(u,α)∈V.

(2) In S, for each attribute subset B($B\subseteq A$), non-clear relation(equivalent relation) IND can be defined in field U×U: IND(B)={(u,u') ∈U×U: $\forall a \in B$,f(u,a)=f(u',a) }.

(3) Equivalent class in partition U/IND(B) according to non-clear relation IND(B), denoted as $[u]_B$, is defined as follow:

[u]B={u' ∈U:uIND(B)u'}

(4) Let $X \subseteq U$, $B \subseteq A$, IND(B) is a non-clear relation in field U×U, the lower approximation set $\underline{B}(X)$ and upper approximation set $\overline{B}(X)$ can be defined respectively as follow:

$$\underline{B}(X) = \{u \in U : [u]_B \subseteq X\}$$

$$\overline{B}(X) = \{u \in U : [u]_B \cap X \neq \phi\}$$

(5) Let $X \subseteq U$, $B \subseteq A$, the boundary region $BN_B(X)$ and the negative region $NEG_B(X)$ can be defined respectively as follow:

$BN_B(X)= \overline{B}(X) - \underline{B}(X)$

$NEG_B(X)=U - \underline{B}(X)$

The lower approximation is also called the positive region of X, $POS_B(X)$, which is the set of objects that can be correctly divided into X according to the information of U/B. The negative region $NEG_B(X)$ is the set of objects that can be divided out of X according to the information of U/B, and the boundary region $BN_B(X)$ is the set of objects that maybe divided into X according to the information of U/B. So according to the information of U/B the set U can be divided into three region: positive region, negative region and boundary region.

Basic Conceptions.

Definition 1: The knowledge set is defined as KS= (U, A), in which U is the knowledge content set and the cardinality is m, A is the attributes set and the cardinality is n.

Obviously, for any $u_i \in$U, its attributes can be described as $A_i=(a_{i1},a_{i2},\ldots,a_{in})$, so the knowledge set can also be represented as:

$$KS = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}. \quad (1)$$

Definition 2: similarity matrix *sm*. Given a knowledge set KS, the similarity matrix of KS is defined as:

$$sm = \begin{bmatrix} as_{11} & as_{12} & as_{1m} \\ as_{21} & as_{22} & as_{2m} \\ & & \\ as_{m1} & as_{m2} & as_{mm} \end{bmatrix}. \quad (2)$$

$$as_{ij} = 1 - \sqrt{\frac{1}{m}\sum_{k=1}^{m}(\frac{a_{ik} - a_{jk}}{\max_k})^2}\ . \quad (3)$$

In which, $\max_k = \max_{h=1}^{m} a_{hk}$. (4)

Definition 3: Interest similarity *ins*. Given a knowledge set KS and its *sm*, for the *i-th* knowledge, its interest similarity *ins* is:

$$ins_i = \frac{ts_{\max} - ts_i}{ts_{\max}}. \quad (5)$$

$$ts_i = \sum_{k=1}^{m} as_{ik}\ . \quad (6)$$

$$ts_{\max} = \max_{k=1}^{m} ts_k\ . \quad (7)$$

The ts_i is the *i-th* row sum of the similarity matrix, the less sum value is, the more dissimilar the knowledge *i* is to others. So the value ins_i reflect the similarity degree of *i-th* knowledge with others.

The approach of constructing knowledge view

According to the rough set theory, the set U can be divided into positive region, negative region and boundary region. So we can regard the knowledge set consisted of knowledge items that have been navigated by user as the set U, and then divide the knowledge set into three region. By calculating the *ins* value of each region, the interested knowledge item can be distinguished.

The approach has six main steps:

(1) Dividing the viewed knowledge set into positive region, negative region and boundary region;

(2) Calculating the *ins* value of the three region respectively;

(3) Reducing the three regions respectively by removing the knowledge item whose *ins* value is greater than the pre-set threshold;

(4) Merging the three reduced regions, and calculating the *ins* value of the merged region;

(5) Reducing the merged region by removing the knowledge item whose *ins* value is greater than the threshold;

(6) Constructing knowledge view with the knowledge item of the reduced-merged region.

The algorithm corresponding to this approach is described as follow:

Input: the knowledge item set D, the pre-set threshold Lambda

Output: the knowledge view KV

// step1:Dividing D into positive region POS_D, negative region NEG_D and

// boundary region BN_D;

Rough_divide(D);

// step2: reducing the three region

For all pos_i in POS_D{

Compute ins_i;

```
    If (ins_i>Lambda) POS_D=POS_D-pos_i;
  }
   For all neg_i in NEG_D{
    Compute ins_i;
    If (ins_i>Lambda) NEG_D=NEG_D-neg_i;
  }
   For all bn_i in BN_D{
    Compute ins_i;
    If (ins_i>Lambda) BN_D=BN_D-bn_i;
  }
  D2=POS_D ∪ NEG_D ∪ BN_D;
  // step 3: constructing KV
   For all d2_i in D2{
    Compute ins_i;
    If (ins_i>Lambda) D2=D2-d2_i;
  }
  Sort knowledge item of D2 according to the ascending order of their ins_i value;
  KV=Φ;
  For all d2_i in D2
    KV=KV+d2_i;
```

The complexity of the three steps respectively is $O(m*n)$,$O(m_1^2*n)+O(m_2^2*n)+O(m_3^2*n)$, $O(m_4^2*n)$, in which: (1) m is the number of knowledge item of D; (2) n is the number of attributes; (3) m_1,m_2 and m_3 is the number of knowledge item existed in the three region respectively, and $m_1+m_2+m_3=m$; (4)m_4 is the number of knowledge item of D2, and is less than m.

Test results

In order to verify the performance of the approach proposed in this paper, we have implemented the corresponding algorithm. The algorithm is compiled with VC++ 6.0 and its running environment is Microsoft Windows XP Professional with Pentium IV 2.8GHz and 1GB memory capacity.

In the experiment we use the Traditional Chinese Medicine dataset from SIRC-TCM. It consists of more than 9000 records and 15 attributes. The subject of navigating records is cough and rheum. We first select 150 records about this subject to navigate, and then use the algorithm to construct the knowledge view. The 150 selected records are divided into positive region (80 records), negative region (20 records) and boundary region (50 records). By setting the threshold Lambda to 0.5, 0.7 and 0.9 respectively, we complement the experiment and test the recall ratio and precision ratio.

The recall ratio and precision ratio are defined as:

$$recall\_ratio = \frac{number\ of\ knowledge\ item\ satisfied\ with\ interest\ in\ view}{number\ of\ knowledge\ item\ satisfied\ with\ interest}. \quad (8)$$

$$precision\_ratio = \frac{number\ of\ knowledge\ item\ satisfied\ with\ interest\ in\ view}{number\ of\ knowledge\ item\ in\ view}. \quad (9)$$

Fig.1 and Fig.2 show the results as follow.

As is seen from Fig. 1 and Fig.2, with the Lambda value increased gradually, the recall ratio increased. This is because that the reducing criteria are higher with the Lambda value increased. But at all events, the recall ratio is above 0.8, and the precision ratio is above 0.85. So it is suitable for constructing knowledge view.

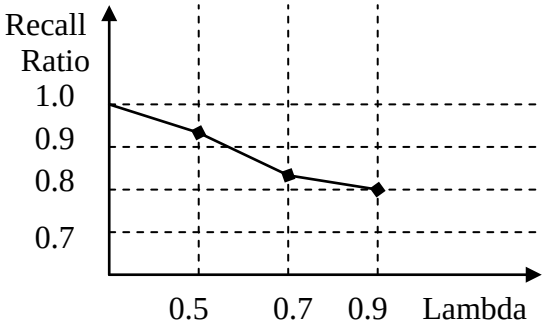

Fig. 1. the recall ratio in different Lambda

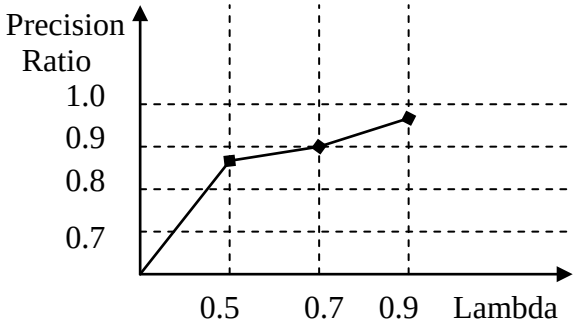

Fig. 2. the precision ratio in different Lambda

Conclusion

The knowledge view is most beneficial for improving knowledge query performance as it stores pre-computed knowledge item. But unlike view maintenance of database, the content of knowledge view is knowledge item. Concerning with this particularity, this paper presents an approach of constructing knowledge view based on rough set and similarity. Experiments show that a knowledge view combined with user interest can be acquired efficiently by this approach.

However, this approach is still in continuous development and improvement. For example, the threshold used in region reducing will directly affect the quality of knowledge view, but its value is set manually. How to adapt the threshold automatically is a topic for future study.

Acknowledgement

In this paper, the research was sponsored by the National Natural Science Foundation of China (Project No.61262049).

References

[1] H.P. Chen, L.F. Chen, J.D. Wang. Reserarch on Issues in Developing Big Data Warehouses. Computer Engineering and Design[J], 2006,27(21) 3956-3958

[2] S. Chen, E.A. Rundenstering. GPIVOT: Efficient Increamental Maitenance of Complx ROLAP Views. 21st International Conf. on Data Engineering(ICDE 05)[C],2005,552-563

[3] A.N.M Rashid, M.S. Islam. Role of Material View Maintenance with PIVOT and UNPIVOT Operators. IEEE International Advance Computing Conferrence(IACC 09)[C], Patiala, India, March 2009, 951-955

[4] M.J. Zhou, J. Gao. User View of Knowledge Management System. Journal of Computer-Aided Design & Computer Graphics[J], 2005 17(5) 1101-1106

[5] A.N.M Rashid, M.S. Islam. An Incremental View Materialization Approach in ORDBMS. International Conf. on Recent Trends in Information, Telecommunication and Computing(ITC 2010)[C],2010, 105-109

[6] X.G. Zhang, L.M.Yang. Incremental View Maintenance Based on Data Source Compensation in Data Warehouses. International Conf. on Computer Application ans System Modeling(ICCASM 2010)[C],2010, 287-291

[7] Wenfei Fan, Xin Wang, Yinghui Wu. Answering Pattern Queries Using Views. IEEE Transactions on Knowledge and Data Engineering[J], 2016 28(2) 326-341

[8] Z Pawlak. Rough Set. International Journal of Computer and Information Science[J], 1982,11(5) 341-356

[9] Z Pawlak. Set Theoretical Aspects Reasoning about Data. Dordrechy the netherlands: Kluwer Acadenmic Publishers,1991