http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 11 (8): 1127-1130, 2012
ISSN 1812-5638 / DOI: 10.3923/itj.2012.1127.1130

A Rough Sets Based Data Preprocessing Algorithm for Web Structure Mining

Li Xiang-Wei, Zheng Gang and Kang Yu-Xue
Lanzhou Polytechnic College, Lanzhou 730050, China

Abstract: Aimed to enhance the efficiency of web structure mining, based on the Rough Sets (RS), an effective web structure mining preprocessing algorithm is proposed in this paper. Firstly, to linear the huge web link graph, the Vast Forward Path (VFP) is introduced and extracted from the user access record in web server logs. Secondly, to build the data analysis model, the Information System is constructed using the VFP. Thirdly, the paper make using of the attribute reduction theory of RS; the Information System is reduced by eliminate a lot of abundant attributes. The experiments show that the proposed algorithm can get high efficiency and avoid the abundant web redundant data.

Key words: Rough sets, web structure mining, data preprocessing

INTRODUCTION

With the development of internet and computer storage technology, amount of web access record are stored in web servers in the form of web logs. It is very rich resource information for understanding web user surfing behaviors. Consequence, web structure mining becomes a very important and hot area to find user trends and regularities in web user's navigation patterns. The results of web structure mining can used extensively to improve and optimize web site structure design, enhance the web server quality.

There are many effective researches and efforts towards mining web structure from web logs. Interesting access patterns mined from web logs are useful knowledge to administer web server in practice (Guoyan, 2011). Examples of application of such knowledge include improving structure designs of web sites, understanding user reaction and motivation and building adaptive web sites or recommended different web site to different user according to user interesting (Tanasa and Trousse, 2004).

Essentially, a web structure mining is a sequential pattern in a large pieces of web logs, which is pursued intensively by web access users. Some research efforts try to employ techniques of sequential pattern mining, which is mostly based on association rule mining and discovering web navigation patterns from web logs (Al-Hamami *et al*., 2006). This statistical approach, called principal clusters analysis, for analyzing millions of user navigations on the web (Meng *et al*., 2009). This technique identifies prominent navigation clusters on different topics. Furthermore it cans deter-mine information items that are useful starting points to explore a topic, as well as key documents to explore the topic in greater detail. The general idea of an intelligent web algorithm that employ predictive models of web requests is to extend the least recently used policy of web and proxy servers by making it sensitive to web access models patterns and decision trees (Romero *et al*., 2009; Mustapasa *et al*., 2010). A new data source called intentional browsing data for potentially improving the effectiveness of WUM applications, which is a category of online browsing actions, such as "copy", "scroll", or "save as" and is not recorded in web logs server (Bedi and Chawla, 2007; Facca and Lanzi, 2005).

THE FUNDAMENTAL THEORY OF ROUGH SETS AND ATTRIBUTES REDUCTION

Indiscernibility relation: Let $U \neq \phi$ be a universe of discourse and X be a subset of U. An equivalence relation, R, classifies U into a set of subsets $U/R = \{X_1, X_2, ..., X_n\}$ in which the following conditions are satisfied:

- $X_i \subseteq U$, $X_i \neq \phi$ For any i.
- $X_i \cap X_j \neq \phi$ For any i, j.
- $\cup_{i=1,2,...,n} X_i = U$

Any subset X_i, which called a category, class or granule, represents an equivalence class of R. A category in R containing an object $x \in U$ is denoted by $[x]_R$. For a family of equivalence relations $P \subseteq R$, an indiscernibility relation over P is denoted by IND(P) and is defined by Eq. 1:

$$IND(P) = \bigcap_{R \in P} IND(R) \quad (1)$$

Corresponding Author: LI Xiang-Wei, Key Laboratory of Gansu Advanced Control for Industrial Processes, Lanzhou 730050, China

Lower and upper approximations: The set X can be divided according to the basic sets of R, namely a lower approximation set and upper approximation set. Approximation is used to represent the roughness of the knowledge. Suppose a set X⊆U represents a vague concept, then the R-lower and R-upper approximations of X are defined by Eq. 2 and 3:

$$\underline{R}X = \{x \in U : [x]_R \subseteq X\} \quad (2)$$

Equation 4 is the subset of X, such that X belongs to X in R, is the lower approximation of X:

$$\overline{R}X = \{x \in U : [x]_R \cap X \neq \phi\} \quad (3)$$

Equation 5 is the subsets of all *X* that possibly belong to X in R, thereby meaning that X may or may not belong to X in R and the upper approximation R contains sets that are possibly included in X. R-positive, R-negative and R-boundary regions of X are defined respectively by Eq. 4-6.

$$POS_R(X) = \underline{R}X \quad (4)$$

$$NEG_R(X) = U\text{-}RX \quad (5)$$

$$BNR(X) = RX\text{-}\underline{R}X \quad (6)$$

Attributes reduction and core: In RS theory, an Information System Table is used for describing the object of universe, it consists of two dimensions, each row is an object and each column is an attribute. RS classifies the attributes into two types according to their roles to Information System Table, i.e. Core attributes and redundant attributes. According to this theory, the minimum condition attribute set can be received, which is called reduction. One Information Table might have several different reductions results simultaneously. The intersection of the reductions results is the Core of the Information System Table and the Core attribute are the important information that include in original Information System.

A subset B of a set of attributes C is a reduction of C with respect to R if and only if:

- $POS_B(R) = POS_C(R)$ and
- $POS_{B-\{a\}}(R) \neq POS_C(R)$ for any $a \in B$

And, the Core can be defined by Eq. 7:

$$CORE_C(R) = \{c \in C | \forall c \in C, POS_{C-\{c\}}\} \quad (7)$$

THE ROUGH SET BASED DATA PREPROCESSING ALGORITHM FOR WEB STRUCTURE MINING

The proposed algorithm consists of the following five steps. Figure 1 shows the flow chart of the algorithm.

Extraction user access sequence: In the internet surfing action, user access is a random process, so it can generate a lot of random user access record in web server. Along with time passing, the user access record stored in web server becomes more and more. By effective preprocess, the user access record can represent by user access sequences such as {A, B, A, C, D, E}. User access sequence represents the user access action during user surfing, it contains much effective implicit and important information that can be used extensively to web design or personal recommendation. By analyzing and structure mining, we can achieve this intelligent information.

Generating user access directed graph: Internet is a huge and global unstructured unit. It includes not only different Web pages but also some stochastic and frequent access link structure. If we regard many Web pages as crunodes and link structure as directed edge and using the weights represents the number of link, the whole huge internet can be considered as weight directed

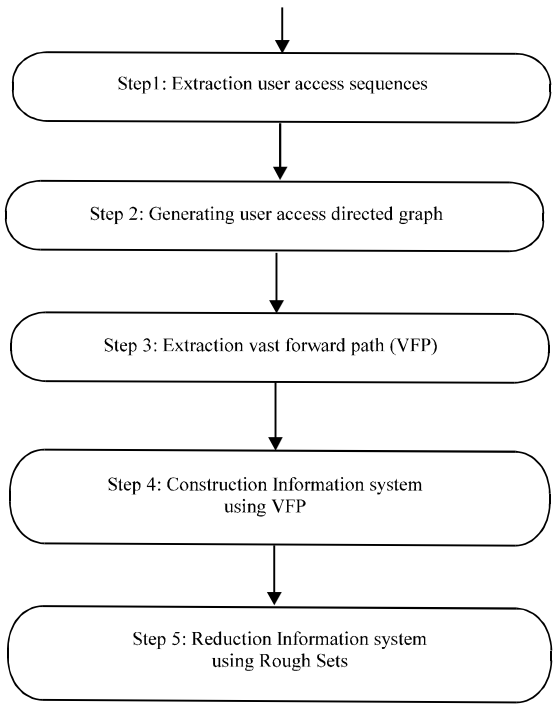

Fig. 1: The flow chart of proposed algorithm

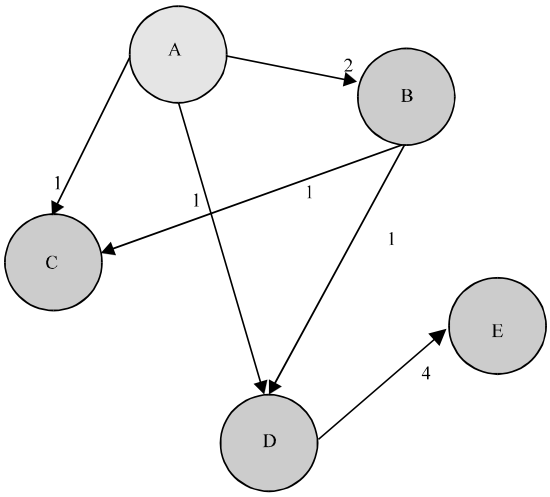

Fig. 2: The user access directed graph

digraph. We can represent weight directed digraph as Eq:

$$G = (V,E) \quad (8)$$

where, V is the sets of page, sometimes also called vertex and E is the sets of link in different page, sometimes also called edge. The input edge of vertex V rep-resent the citation of vertex v and output edge of vertex V represent the citation to the other edge. The weight represents the number of the citation. A classical user access directed graph is showed as Fig. 2.

Extraction of vast forward path: In order to construct the model, the paper introduces directed graph represent the user access action model. While directed graph is very sophisticated, unstructured and even have loop, we must take any other measure to solve this problem. In this paper, the Vast Forward path (VFP) is proposed to overcome it.

Definition 1: The vast forward path (VFP): User access path are the page sequences that accessed during the browsing process by internet user and the vast forward path are the access path that consist of page which turn back to the previous page. For example, suppose that a user access sequences are A-B-A-C-D-C; the vast forward path can be represent A-B and A-C-D.

The vast forward path implies that all the relational crunodes have some relations or some same property. We can regard all these pages as the same categories.

Table 1: The information system constructed by VFP

	A	B	C	D	E
A	0	2	1	1	0
B	0	0	1	1	0
C	0	0	0	0	0
D	0	0	0	0	4
E	0	0	0	0	0

Table 2: The reduced information system using rough set

	A	B	C	E
A	0	2	1	0
B	0	0	1	0
C	0	0	0	0
D	0	0	0	4
E	0	0	0	0

Construction of information system: According to the vast forward path implies that all the relational crunodes have some relations or some same property. We can regard all these pages as the same categories. previous description, the web structure access logs is a graphic structure which is nonlinear structure, so many linear theory which is mature can not used to analyze. Based on the analysis of web access logs, the Information System can be construct using user access records in web server logs, since user access records represent the preference of browser and the analysis of user access records can find interesting information. As the Fig. 2 show, we construct the Information System with web access nodes A, B, C, D and E and the row represent the attributes, the column represent the access object, the Information System can be constructed as Table 1.

Reduction information system using rough sets: According to the RS attributes reduction theory, the Table 1 can be reduce as Table 2, since the attribute C and the attributes D have the same role to user A, B, C, D, E. so we can reserve any of the attributes of C or D and can reduce the other attributes. With the growth of the access data, more and more data can be reduced, corresponding to the whole internet access logs, the process efficiency can be enhanced dramatically while the whole structure of the user access logs reserved. Many effective analysis can be play on the reduced Information System.

RESULTS AND DISCUSSION

At present, it is an urgent problem for the applications of web structure mining how to find the user preferred patterns accurately and how to optimize web site and improve commercial strategy. In this study, we present a novel Rough Sets based web structure mining preprocess algorithm, the algorithm firstly extract most forward path from the user access logs and then construct the Information System use user access logs, finally, by

introducing the RS attributes reduction theory, the paper achieved the reduced Information System, since the reduced information system avoid the affection of redundant data and reserved the original structure of user access logs, the reduced Information System can be widely used to the web structure analysis and achieve high performance.

ACKNOWLEDGMENTS

Author, acknowledge the Doctor Subject Foundation of the Ministry of Education of China under Grant No. 20106201110003. The Gansu Natural Science Foundation of China under Grant No. 1107RJZA170.

REFERENCES

Al-Hamami, A.H., M.A. Al-Hamami and S.H. Hasheem, 2006. Applying data mining techniques in intrusion detection system on web and analysis of web usage. Inform. Technol. J., 5: 57-63.

Bedi, P. and S. Chawla, 2007. Improving information retrieval precision using query log mining and information scent. Inform. Technol. J., 6: 584-588.

Facca, F.M. and P.L. Lanzi, 2005. Mining interesting knowledge from weblogs: A survey. Data Knowledge Eng., 53: 225-241.

Guoyan, H., 2011. Mining web frequent multi-dimensional sequential patterns. Inf. Technol. J., 10: 2434-2439.

Meng, X.J., Q.C. Chen and X.L. Wang, 2009. A tolerance rough set based semantic clustering method for web search results. Inform. Technol. J., 8: 453-464.

Mustapasa, O., D. Karahoca, A. Karahoca, A. Yucel and H. Uzunboylu, 2010. Implementation of semantic web mining on e-learning. Soc. Behav. Sci., 2: 5820-5823.

Romero, C., S. Venturaa, A. Zafraa and P. de Brab, 2009. Applying web usage mining for personalizing hyperlinks in web based adaptive educational systems. Comput. Educ., 53: 828-840.

Tanasa, D. and B. Trousse, 2004. Advanced data preprocessing for intersites web usage mining. Intell. Syst., 19: 1541-1672.