Association Rule Mining for Large Database

Mrs.Kalpana Wani
Department of Computer Science,
PIIT New Panvel, Navi Mumbai, India

Prof. Madhu Nashipudimath
Department of Information Technology
PIIT New Panvel, Navi Mumbai, India

Abstract

As we know World Wide Web plays vital role in serving the needs of the user's on web. The web log files are generated as a result of an interaction between the client and the service provider on web. Web log file contains the massive hidden valuable information pertaining to the visitors, if mined can be used for predicting the navigation behavior of the users. However the task of discovering frequent sequence patterns from the web log is challenging. This system focuses on adopting an intelligent technique that can provide personalized web service for accessing related web pages more efficiently and effectively, so that it can be determined which web pages are more likely to be accessed by the user in future. Proposed system uses two intelligent algorithms for predicting the user behavior's namely FP Growth and Eclat.These algorithms save the time and space problem of existing system. Further from the frequent pages pattern Direct and Indirect Association Rules are generated and based on that Ranking is provided to pages which will help recommendation system to recommend similar search pages.

Keywords— Frequent pattern, Association Rule, Indirect association rule, Complex association rule.

1. Introduction

The World Wide Web serves as a vast, widely distributed, global information service center for advertisement, consumer information, e-commerce, education, financial management, government, news and many other information services. So, it has become much more difficult to access relevant information from the web with the explosive growth of information available on the internet. Therefore, further research work needs to be carried out for extracting the appropriate content as per the user's needs. This can be performed using Web mining as it helps in extracting the common sequences of the user's accessed web pages. In the web environment, association rules are typically applied to HTTP server log data that contain historical user sessions. Web sessions are gathered without any user involvement and additionally, they reliably reflect user behavior while navigating throughout a web site. For that reason, web sessions can be regarded as an important source of information about users. Association rules that reveal similarities between web pages derived from user behavior can be simply utilized in recommender systems. The main goal of such a recommendation is to suggest to the current user some web pages that appear to be useful.

2. Literature Survey

The literature review focuses on the study, comparison and contrast of the available preprocessing techniques. Data Cleaning is done to remove the inappropriate records with unsuccessful status [17]. The ability of using the data mining techniques to extract information from the server logs was first introduced by [14], [15], and [16]. The elements [11] of the web usage access are the users and web pages accessed by the user. The goal of web usage mining is to analyze the user access behavior patterns. Web mining can be practices in three different domains i.e. the content mining, hyper link web structure mining and web usage mining. These approaches effort to extract valuable information from the web which are then applied to some real world problems. The user's surfing behavior analysis follows three phases [6]: data collection and preparation, pattern discovery and content recommendation.

The output generated from the pattern analysis consists of sequences of accesses with corresponding probabilities. Frequent Pattern describes how often the pages are accessed together in a sequence. This has been introduced in 1993 by Argawal et al. [17] .The algorithms used to mine the usage are association rule mining and sequence analysis. Association rule mining discovers relationships between different web pages within a web site this approach helps to find out the order in which the pages are visited, reduces the bandwidth usage and storage needs, which undoubtedly results in improving the system efficiency and effectiveness i.e. an improved system. The objective of the association rules mining [18], [12],[10] is to discover correlations of association between existing records in a dataset. In [12],[10] fundamental association rules and indirect association rule have been described In [8], more practical and efficient methods are being presented, in order to find association rules in the case of less frequent items. Web log [13] can do this only after analyzing data resulted from the users' current and history data.

3. Problem Statement

The objective of this project is to analyzing the navigation behavior of the web access users from the accessed (filtered) data by applying two algorithms to find frequent access pages and finds the direct and indirect association rule between the frequent access pages. Besides many advantages,

association the loss of some vital information. Typical association rules focus on the co-occurrence of items (purchased products, visited web pages, etc.) within the transaction set. Web pages are connected with each other using hyperlinks and they usually determine all possible navigational paths. To reach a page, the user is often forced to navigate through other pages, e.g., a home page, a login page, etc. Additionally, the web site content is usually organized by the designer into thematic blocks which are not always suitable for particular users. For all these reasons, some personalized recommendation mechanisms are very useful in most web portals.

4. Proposed System

The Proposed system will analyse user behavior efficiently and effectively, so that it can be determined which web pages are more likely to be accessed by the user in future. Performance of web page Recommendation System is improved by using algorithm which will save time and memory space for discovering frequently access pages and finds the Direct and Indirect Association Rules from Frequent Access Pages to recommend more number of pages to recommendation system by assigning Ranks to the frequent access pages.

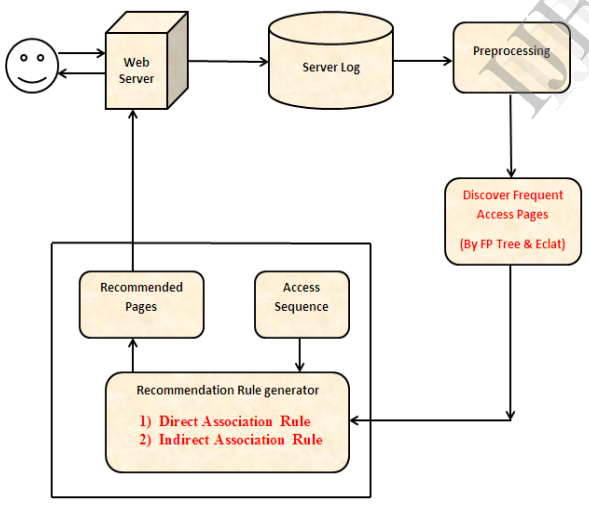

Fig.1. Architecture Diagram for Proposed System

5. Steps for Implementation

Following steps are used for implementation:

Step 1: Data Cleaning & Preprocessing

Preprocessing is necessary, because Log file contain noisy & ambiguous data which may affect result of mining process. Some of web log file data are unnecessary for analysis process and could affect detection of web attack. Data preprocessing is an important steps to filter and organize only appropriate information before applying any web mining algorithm. Preprocessing reduce log file size also increase quality of available data. The purpose of data preprocessing is to improve data quality and increase mining accuracy. Preprocessing consists of field extraction, data cleansing, user identification, session identification.

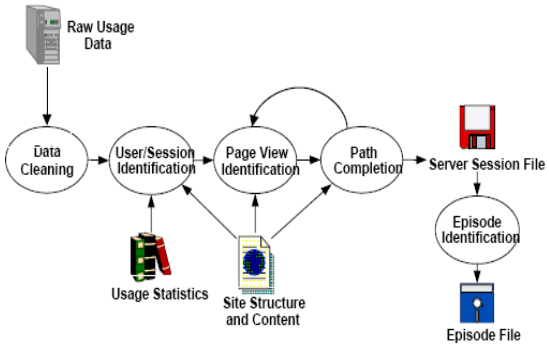

Fig .2. Block diagram for Data Preprocessing

Step 2: Finding Frequent Pattern

First count the occurrences of pages in database and then reorder the pages with higher count page first and so on. Draw the FP Growth tree and start mining FP tree by considering the suffix as page with lowest count first. Find the conditional pattern base for each page and if length of conditional pattern base is less than some threshold value then find the frequent pattern using Éclat algorithm and if length is greater than or equal to the threshold value then use FP Growth algorithm for finding the frequent pattern of visited pages. Sample session dataset and remaining calculation is shown below.

Table 1. Session Database

Session Id	Pages
1	P1,P2,P4
2	P1,P4
3	P1,P2,P4
4	P1,P3
5	P2,P4,P5,P6
6	P2,P4
7	P4,P5,P6
8	P2,P4,P5,P6
9	P1,P6
10	P1,P3

Table 2.Page Count

P4:7
P1:6
P2:5
P6:4
P5:3
P3:2

Table 3. Session Database with ordered Pages

Session Id	Pages	Ordered Pages
1	P1,P2,P4	P4,P1,P2
2	P1,P4	P4,P1
3	P1,P2,P4	P4,P1,P2
4	P1,P3	P1,P3
5	P2,P4,P5,P6	P4,P2,P6,P5
6	P2,P4	P4,P2
7	P4,P5,P6	P4,P6,P5
8	P2,P4,P5,P6	P4,P2,P6,P5
9	P1,P6	P1,P6
10	P1,P3	P1,P3

After rearranging the pages, form the FP Growth Tree starting from the lowest to highest page count as shown below. Count the no.of occurrences of each page.

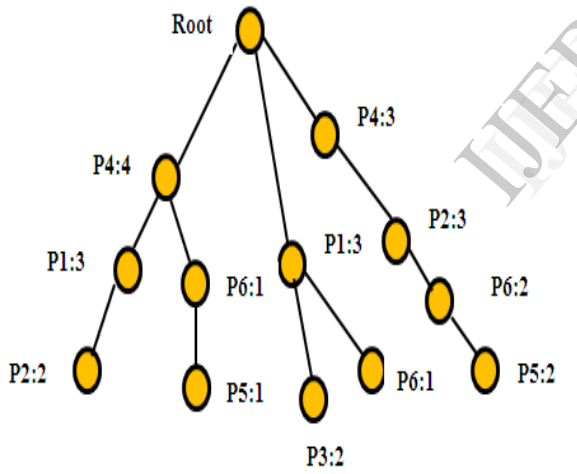

Fig.3. FP-Growth Tree

Table 4. FP-Tree Mining

Item	Condition Pattern Base	Conditional Pattern	Frequent Pattern
P3	{P1:2}	{P1:2}	{P1,P3:2}
P5	{P4:1,P6:1} {P4:2,P2:2,P6 :2}	{P4,P6:1} {P4 P2 P6:2}	{P4 P6 P5:3} {P4 P2 P6 P5:2} {P4 P2 P5:2} {P4 P5:3} {P2 P5:2} {P5 P6:3}
P6	{P4:1} {P1:1} {P4,P2:2}	{P4:1} {P4 P2:2}	{P4 P6:3} {P2 P4:2} {P4 P2 P6:2}
P2	{P4 P1:2} {P4:3}	{P4:5,P1:2}	{P4 P1 P2:2} {P4 P2:5} {P1 P2:2}
P1	{P4:3}	{P4:3}	{P4 P1:3}

After finding the frequent pattern find the confidence and support value for each frequent pattern as shown below.

Table 6. Frequent Patterns with Support and Confidence

Frequent Pattern	Confidence	Support
{P1,P3}	30%	20%
{P4 P6 P5}	P4 P6->P5 = 100% P4 P5->P6 =100% P5 P6->P4 =100%	30%
{P4 P2 P6 P5}	P4 ->P2 P6 P5 = 28% P2 ->P4 P6 P5 = 40% P6 ->P4 P2 P5 = 5% P5 -> P4 P2 P6 = 60% P2 P6 P5->P4 =100% P4 P6 P5->P2 = 100% P4 P2 P5->P6 = 100% P4 P2 P6->P5 = 100% P4 P2->P6 P5 =40% P4 P6->P2 P5 = 60% P4 P5->P2 P6 =60% P6 P5->P4 P2 =60% P2 P5->P4 P6 =100% P2 P6->P4 P5 =100%	20%
{P4 P2 P5}	P4 P2->P5 =40% P4 P5->P2 = 60% P2 P5->P4 = 100%	20%
{P4 P5}	42%	30%
{P2 P5}	40%	20%
{P5 P6}	100%	30%
Frequent Pattern	**Confidence**	**Support**
{P4 P6}	42%	30%
{P2 P6}	40%	20%
{P4 P2 P6}	P4 P2->P6 = 40% P4 P6->P2 = 60% P6 P2->P4 = 100%	20%
{P4 P1 P2}	P4 P1->P2 = 60% P4 P2->P1 = 40% P1 P2->P4 = 100%	20%
{P4 P2 }	71%	50%
{P1 P2 }	33%	20%

Similarly find the confidence and support value for remaining frequent pattern as shown above.

Step 3: Direct Association Rule Generation between the Frequent Access Pages

After getting the frequently visited page sequence, filter the pattern by applying some minimum confidence and support value and form the one page association rule between pages of frequent pattern and calculate its support and confidence values and filter it with respect to minimum confidence and support value as shown below.

Table 7. Formation of Direct Association Rule

Sr.No	Rule	Confidence	Support
1	P1->P4	0.5	0.2
2	P2->P4	1	0.5
3	P2->P5	0.4	0.2
4	P2->P6	0.4	0.2
5	P4 ->P2	0.6	0.5
6	P4->P6	0.5	0.4
7	P5->P2	0.6	0.2
8	P6->P2	0.4	0.2
9	P6->P4	0.8	0.4
10	P6->P5	0.4	0.2

Step 4: Indirect Association Rule Generation between the Frequent Access Pages

In direct association rule of pages there are many pages which are connected to each other indirectly via some transitive pages. Find such pages which are indirectly connected and filter it with respect to minimum confidence and support value.

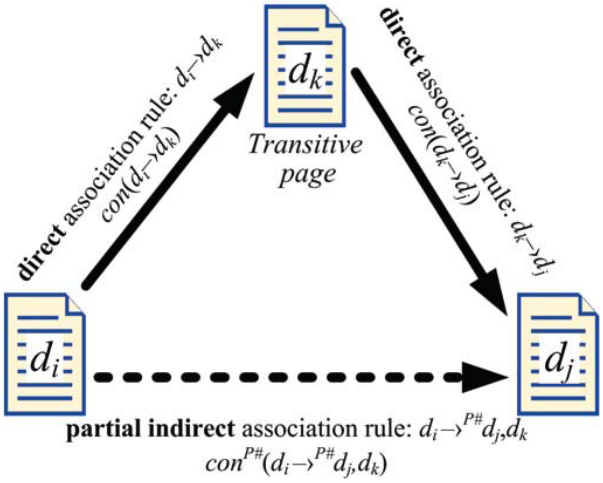

Fig. 4. Indirect Association between Two Web Pages.

Step 5: Complete Indirect Association Rule Generation between the Frequent Accesses Pages

The complete indirect association rule di→#dj aggregates all partial indirect association rules from di to dj with respect to all existing transitive pages dk ∈ Tij and is characterized by complete indirect confidence con#(di→#dj)

$$con^{\#}(d_i \rightarrow^{\#} d_j) = \frac{1}{max_T} \sum_{d_k \in T_{ij}} con^{P\#}\left(d_i \rightarrow^{P\#} d_j, d_k\right)$$

Where

$$max_T = \max_{d_i, d_j \in D} \left(card\left(T_{ij}\right)\right)$$

Table 8. Formation of Indirect Association Rule

L1 (i--k)	L2 (k--j)	Transitive pages	Lk (k---j)	Complete Indirect Rule
P4 ->P2 P5->P2 P6->P2	P2->P4 P2->P5 P2->P6	P2	P2->P4 P2->P5 P2->P6	P4->#P5,P2 P4->#P6,P2 P5->#P4,P2 P5->#P6,P2 P6->#P5,P2
P1->P4 P2->P4 P6->P4	P4->P2 P4->P6	P4	P4->P2 P4->P6	P1->#P2,P4 P1->#P6,P4 P2->#P6,P4 P6->#P2,P4
P2->P5 P6->P5	P5->P2	P5	P5->P2	P6->#P2,P5
P2->P6 P4->P6	P6->P2 P6->P4 P6->P5	P6	P6->P2 P6->P4 P6->P5	P2->#P4,P6 P2->#P5,P6 P4->#P2,P6 P4->#P5,P6

Table 9. Indirect Association Rule with Confidence

Sr.No	Rule	Confidence
1	P1-->#P2,P4	0.3
2	P1-->#P6,P4	0.25
3	P2-->#P4,P6	0.32
4	P2-->#P5,P6	0.16
5	P2- ->#P6,P4	0.5
6	P4-->#P2,P6	0.2
7	P4-->#P5,P2	0.24
8	P4-->#P5,P6	0.2 / 0.22
9	P4-->#P6,P2	0.24
10	P5-->#P4,P2	0.6
11	P5-->#P6,P2	0.24
12	P6-->#P4,P2	0.4
13	P6-->#P5,P2	0.16
Sr.No	**Rule**	**Confidence**
14	P6->#P2,P4	0.48
15	P6->#P2,P5	0.24 / 0.36

Step 5: Complex Association Rule Generation

A complex association rule exists if at least one of two component rules exists, i.e., either direct or complete indirect or both of them. The main quality features of both direct and indirect rules—confidences—are combined within complex association rules.
A complex association rule is characterized by the complex confidence, con*(di→∗ dj), as follows

$$Con^*(di \rightarrow^* dj) = \alpha.Con(di \rightarrow dj) + (1 - \alpha).Con\#(di \rightarrow \#dj)$$

Where α is the direct confidence reinforcing factor
$\alpha \in [0, 1]$
Calculate the complex confidence value for different values of α. Setting α we can emphasize or damp the direct confidence at the expense of the Complete indirect one. The greater the value of α, the closer the complex confidence to the direct one.

Depending on highest value of α rank is assigned to the pages and sends it to recommendation system.

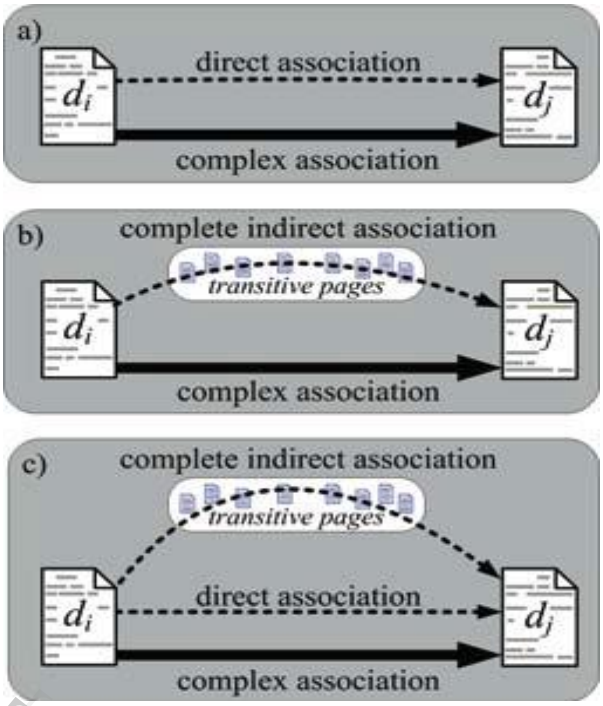

Fig. 5. Complex association results from either a direct association (a) or a complete indirect one (b) or both (c).

By considering different values of α i.e. between [0.2-0.9] value of complex association rule is calculated, its calculation is shown in table on next page. Based on this value rank is assigned to the pages which is shown below in table.

Table 10. Page Ranking List

Page	Direct Pi->Pj	Indirect Pi->#Pj	Complex:con*(Pi->Pj)							
			α= 0.2	α= 0.3	α= 0.4	α= 0.5	α= 0.6	α= 0.7	α= 0.8	α= 0.9
P1	P4	P2 P6	P2P6P4		P2P4P6	P4P2P6		P2P4P6		P4P2P6
P2	P4P5P6	P4P5P6	P2 P6 P4							
P4	P2P5	P2P5P6	P4P2P6				P2P4P6			P4P2P6
P5	P2	P4P6	P4P6P2	P4P2P6			P2P4P5			
P6	P2P4P5	P4P5	P6P2P4						P4P6P2	P2P4P6

Table11. Values of complex confidence

Sr.No	Rule	Direct Pi->Pj	Indirect Pi->#Pj	Complex:con*(Pi->Pj)							
				α= 0.2	α= 0.3	α= 0.4	α= 0.5	α= 0.6	α= 0.7	α= 0.8	α= 0.9
1	P1-->*P2	✗	✓	0.24	0.21	0.18	0.15	0.12	0.9	0.6	0.03
2	P1-->*P3	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
3	P1-->*P4	✓	✗	0.1	0.15	0.2	0.25	0.30	0.35	0.40	0.45
4	P1-->*P5	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
5	P1-->*P6	✗	✓	0.2	0.175	0.026	0.125	0.1	0.075	0.05	0.025
6	P2-->*P1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
7	P2-->*P4	✓	✓	0.456	0.524	0.592	0.66	0.728	0.796	0.864	0.932
8	P2-->*P5	✓	✓	0.208	0.013	0.015	0.016	0.015	0.013	0.010	0.005
9	P2-->*P6	✓	✓	0.32	0.042	0.048	0.05	0.048	0.42	0.032	0.018
10	P3-->*P1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
11	P4-->*P1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
12	P4-->*P2	✓	✓	0.019	0.025	0.028	0.030	0.288	0.252	0.192	0.010
13	P4-->*P5	✗	✓	0.176	0.154	0.132	0.11	0.088	0.066	0.044	0.022
14	P4-->*P6	✓	✓	0.019	0.025	0.028	0.03	0.028	0.025	0.019	0.010
15	P5-->*P1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
16	P5-->*P2	✓	✗	0.12	0.18	0.24	0.30	0.36	0.42	0.48	0.54
17	P5-->*P4	✗	✓	0.48	0.42	0.36	0.30	0.24	0.18	0.12	0.06

18	P5-->*P6	✗	✓	0.192	0.168	0.144	0.12	0.096	0.072	0.048	0.024
19	P6-->*P1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
20	P6-->*P2	✓	✗	0.08	0.12	0.16	0.20	0.24	0.28	0.32	0.36
21	P6-->*P4	✓	✓	0.051	0.067	0.076	0.08	0.076	0.067	0.512	0.028

6. Conclusion

Web usage mining has valuable uses to the marketing of businesses and a direct impact to the success of their promotional strategies and internet traffic. Analysis of web log data will help companies to develop promotions that are more effective. A quick web page recommendation system help user to get required information with multiple options without going through entire web site pages. Proposed system will provide efficient method for extraction of the frequent access pages and method for finding Direct and Indirect Association Rule which will provide multiple pages for Recommendation system.

7. Reference

[1] Bina Kotiyalt, Ankit Kumar2, Bhaskar Pant3, R.H. Goudar4, Shiv ali Chauhan5 and Sonam Junee6." User Behavior Analysis in Web Log through Comparative Study of Éclat and Apriori", Proceedings of the International Conference on Intelligent Systems and Control (ISCO 2013)

[2] Dr. S. Vijayarani, Ms. P. Sathya "An Efficient Algorithm for Mining Frequent Items in Data Streams",in International Journal of Innovative Research in Computer and Communication Engineering Vol. 1, Issue 3, May 2013

[3] Ms. Shweta, Dr. Kanwal Garg ,"Mining Efficient Association Rules Through Apriori Algorithm Using Attributes and Comparative Analysis of Various Association Rule Algorithms "in Volume 3, Issue 6, June 2013 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.

[4] N. VENKATESAN, RAMARAJ," FREQUENT ITEMSET MINING WITH BIT SEARCH", Journal of Theoretical and Applied Information Technology, 15 July 2012. Vol. 41 No.1

[5] Ravi Bhushan and Rajender Nath," Recommendation of Optimized Web Pages to Users Using Web Log Mining Techniques", 978-1-4673-4529-3/12/$31.00c 2012 IEEE

[6] Rachna Somkunwar" Hash-Set Technique of Association Rules"in Volume 2, Issue 11, November 2012 ISSN: 2277 128 X, International Journal of Advanced Research in Computer Science and Software Engineering.

[7] Chowdary Farha ahmed, Byeong-Soo Jeong, "Efficient mining of high utility patterns over data streams with a sliding window model".Springerlink.com, 2011

[8]Yo unghee Kim Won Young Kim and Ungmo Kim "Mining frequent item sets with normalized weight in continuous data streams". Journal of information processing systems. 2010.

[9] L. Zhou, Z. Zhong, J. Chang, J. Li, J.Z. Huang, S. Feng, "Balanced parallel FP-Growth with Map Reduce", in Conference on Information Computing and Telecommunications, IEEE, pp. 243 – 246, 2010.

[10] Ding, Yau, S. S.: TCOM, "An innovative data structure for mining association rules among infrequent items", pp. 290-301. Comput. Math. Appl. 57, 2 (2009).

[11] Ding, 1. Yau, S. S.: TCOM, an innovative data structure for mining association rules among infrequent items, pp. 290-301. Comput. Math. Appl. 57, 2 (2009).

[12] Sang Lin ,Hu-yan Cui ,Ren Ying ,Zhou-lin Lin ," Algorithm Research for Mining Maximal Frequent Itemsets Based on Item Constraints", 2009 Second International Symposium on Information Science and Engineering, ISBN: 978-0-7695-3991.

[13] PRZEMYSŁAW KAZIENKO ." MINING INDIRECT ASSOCIATION RULES FOR WEB RECOMMENDATION", Int. J. Appl. Math. Comput. Sci., 2009, Vol. 19, No. 1, 165–186