

Association Rule in Web Usage mining

Prof. Marathe Dagadu Mitharam

ABSTRACT:-

Association Rule is a fundamental of Data mining task. Its objective to find all co-occurrence relationship called, Association among data item. Let $I = \{i1, i2, \dots, im\}$ be a set of **items**. Let $T = (t1, t2, \dots, tn)$ be a set of **transactions**.

Goal: - Association and Correlation Analysis

In this paper describes

- 1) Discovery and analysis of web usage patterns using Association analysis.

* * * _ _ * * *

1) INTRODUCTION:-

Association rule discovery and statistical correlation analysis can find groups of items or pages that are commonly accessed or purchased together. Association based on Apriori algorithm. This algorithm finds groups of item using support and confidence. Satisfying a user specified minimum support threshold). Such groups of items are referred to as frequent itemsets & frequent itemsets graph.

Log files generated by web servers contain enormous amounts of web usage data that is potentially valuable for understanding the behavior of website visitors.

e.g.

$$\text{Support} = \frac{(X \cup Y).\text{Count}}{n}$$

$$\text{Confidence} = \frac{(X \cup Y).\text{Count}}{X.\text{Count}}$$

Using above examples we can accept the **minsub** and **minconf**. To calculate minsub and minconf as follows.

T1	C++, JAVA, RUBY
T2	C++, ASP
T3	ASP, VB
T4	C++, JAVA, ASP
T5	C++, JAVA, PHP, ASP, RUBY
T6	JAVA, PHP, RUBY
T7	JAVA, RUBY, PHP
JAVA, PHP	RUBY → [sup=3/7, conf=3/3]

In above 7 transactions JAVA, PHP & RUBY show 3/7 times. Every item checks itemset to every using **Joining** and **Pruning** steps. In web usage mining such rule can be use to optimize structure of website.

e.g.

Language, /product/software COLLEGE Website → RCP ACS

3) EXPERIMENT- FINDING WEB USAGE ASSOCIATION RULES:-

Relation: weather					
No.	outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	play Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

Instances: 14

Attributes: 5

outlook

temperature

humidity

windy

play

The purpose of this experiment was to give some insight into the usefulness of association rules when they are applied to the web log data set of an education institution and others. We expected to find rules that correlate to web pages that contain information about sunny, rainy or temperature etc.

If check sunny, false → yes [sub 1/14, conf 1/1]

Suppose this is transaction table and find out Frequent Itemset then,

T1 C++, JAVA, RUBY

T3 ASP, VB

T5 C++, JAVA, PHP, ASP, RUBY

T7 JAVA, RUBY, PHP

T2 C++, ASP

T4 C++, JAVA, ASP

T6 JAVA, PHP, RUBY

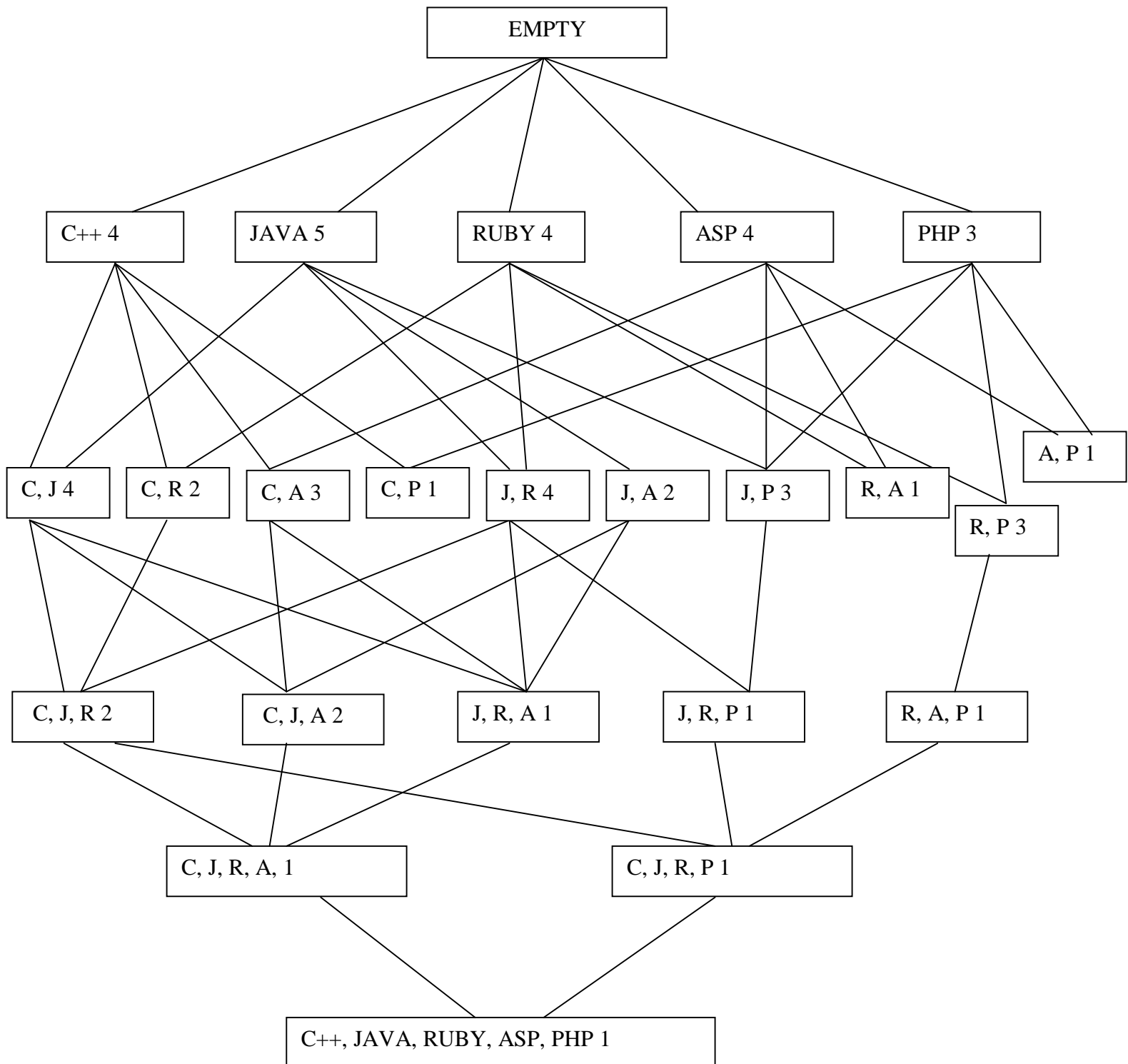
Size1		Size 2		Size 3		Size 4	
Item set	Supp.	Item set	Supp.	Item set	Supp.	Item set	Supp.
C++	4	C++, JAVA	3	C++, JAVA, RUBY	2	C++, JAVA, RUBY, ASP	1
JAVA	5	C++, RUBY	2	C++, JAVA, ASP	2	C++, JAVA, RUBY, PHP	1
RUBY	4	C++, ASP	3	JAVA, RUBY, ASP	1		
ASP	4	C++, PHP	1	JAVA, RUBY, PHP	3		
VB	1	JAVA, RUBY	4	RUBY, ASP, PHP	1		
PHP	3	JAVA, ASP	2				
		JAVA, PHP	3				
		RUBY, ASP	1				
		RUBY, PHP	3				
		ASP, PHP	1				

Fig. 1 Web transactions and resulting frequent itemsets (minsup = 1)
In the Fig 1 find out frequent itemset by using joining and pruning methods of association rule.

4) FREQUENT ITEMSET GRAPH:-

Finds items *C++* and *RUBY* as candidate recommendations. The recommendation scores of item *A* and *C* are 1, corresponding to the confidences of the rules, *JAVA, ASP -> C++* and *JAVA, ASP -> RUBY*, respectively.

A problem with using a single global minimum support threshold in association rule mining is that the discovered patterns will not include “rare” but important items which may not occur frequently in the transaction data.



C=C++
R=RUBY

J=JAVA
P=PHP

A=ASP

Fig 2. Frequent itemset

5) CONCLUSION:-

This paper has attempted to for the purpose of web usage mining. The proposed methods were successfully tested on the log files using association rule.

We conducted a comprehensive analysis of web usage association rules found in a website of an educational institution. Our experiments confirmed that one of the major issues in association rule finding is the existence of too many rules, all of which satisfy defined constraints, but it is difficult to exploit and identify those that are truly interesting to the user.

Using this paper the user easily solved the support, confidence and frequent itemset in server log file.

6) REFERNECES:-

- 1) Web data mining – Bing Liu
- 2) PPT for Association Rule- Bing Liu

3) Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N.(2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. ACM SIGKDD, Jan 2000.

4) Jaideep Srivastava Paper

5) WCA. Web characterization terminology & definitions.

6) <http://www.w3.org/1999/05/WCA-terms/>. Vigente al 19/11/2005

Author Name:- Prof. Marathe Dagadu Mitharam
R.C.Patel A.C.S. College, Shipur
At/Post- Thalner Tal- Shirpur Dist- Dhule
(Maharashtra)