www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume2 Issue 8 August, 2013 Page No. 2635-2640

Mining Association Rules for Web Crawling using Genetic Algorithm

J. Usharani, Dr. K. Iyakutti

Assistant Professor Madurai Kamaraj Universit College Madurai

Professor, Dept of Physics and Nanotechnology, SRM University, Chennai, India

Abstract
With the recent advancement of Internet and Web Technology, web search has taken an important role in our ordinary life. A Web crawler is a software program that automatically retrieves the Web pages when a query is placed in the search engine. To discover interesting patterns or relationship between data in large database Association rule mining is used. Association rule mining can be an important data analysis method to discover associated web pages. The Apriori algorithm is a proficient algorithm for determining all frequent web pages. The Frequent web pages form the Association rules. We proposed a novel approach - genetic based apriori algorithm for web crawling. The proposed method yields promising results compared to the ordinary apriori algorithm and we present empirical results to substantiate this claim.

1. Introduction

With the steep increase in the information on the World Wide Web, there is a great demand for developing efficient and effective methods to organize and retrieve the information available. The Web mining extracts useful information from the web pages. Web mining techniques seek to extract knowledge from Web data, including web documents, hyperlinks between documents. The purpose of the association rule is to find correlations between the processes of any application. The traditional way to find frequent item sets is to use the apriori algorithm.[1]
A genetic algorithm is a type of searching algorithm. It searches a solution space for optimal solution to the problem. [4]Applying the genetic algorithm for web crawlers is likely to yield good results.

2. Association Rule Mining

In general, the association rule is an expression of the form x-→y where x is antecedent and y is consequent. An antecedent is an item found in the data. A consequent is an item that is found in combination with antecedent. The main aim is extracting important correlation among data items in the database. Basically it extracts the pattern from the data based on the two measures such as minimum confidence and minimum support

Support it indicates of how frequently the items appear in the database. Confidence indicates the number of times the if/then statement have been found to be true.

Support

It is the probability of item or item sets given transactional database

$$\text{Support}(x)= \frac{n(x)}{n}$$

where n is the total number of transaction in the database and n(x) is the number of transaction that contains the item set x.

Support(x→y)=Support(xUy)

Confidence

It is conditional probability for an association rule x→y as defined as

$$\text{Confidence}(xUy)= \frac{\sup port(xUy)}{\sup port(x)}$$

The various association rule mining algorithms were used to different applications to determine interesting frequent patterns. One of the association rule mining algorithm such as apriori algorithm .

3.Apriori Algorithm

Apriori algorithm is the best- known algorithm to mine association rules. It uses breadth first search technique to counting the support of item sets and a candidate generate function with exploits the downloaded closure property of support. This algorithm uses an iterative approach called level-wise search, in which n-item sets are used explore n-1 item set..It finds the association between the various items of database. This algorithm also called 'generate and test type 'iterate over the data base in the multiple passes.[3] Apriori was the first scalable algorithm designed for association-rule mining algorithm. The apriori algorithm searches for large item sets during its initial database pass and uses it as the seed for discovering other datasets during subsequent passes.

The General process of Apriori Algorithm

First minimum support is applied to find all frequent item sets in a database.

Second ,these frequent item sets and minimum confidence constraint are used to form rules.

Let D the task-relevant data, be a set of database transactions where each transaction T is a set of items, called $T_{id.}$.Let $I=\{I_1,I_2,\ldots\ldots I_m\}$ be a set of items. An item set contains k items is a k-item set. If a k-item set satifies minimum support then it is a frequent k-item set denoted by k_{Lk}.Firstly,Apriori algorithm generated a set of candidates, which is candidate k-item sets,donoted by ck.if the candidate item sets satisfies minimum support then it is frequent item sets.

Apriori Algorithm

C_k:Candidate itemset of size k

L_k:Frequent itemset of size k

L_1:{Frequent items};

For(k=1;Lk!= Φ;k++) do begin

C_{k+1}=candidates generated from Lk;

For each transaction t in database do

Increment the count of all candidates in Ck+1 that are contained in t

L_{k+1}=Candidates in C_{k+1} with min_support

End

Return $\cup_k L_k$;

4. Web crawler

A Web crawler is a program that automatically collects Web pages to create a local index and/or a local collection of web pages. Roughly, a crawler starts with an initial set of URLs, called seed URLs. It first retrieves the pages identified by the seed URLs, extracts any URLs in the pages, and adds the new URLs to a queue of URLs to be scanned. Then the crawler gets URLs from the queue (in some order), and repeats the process. Web Crawler are mainly used to create a copy of all the visited pages for later processing by a search engine by a search engine that will index the downloaded pages to provide fast searches.

5.Genetic Algorithm

Genetic Algorithm (GA) is an adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics.[2]This heuristic is used to generate useful solutions to optimization and search problems. GAs are inspired by Darwin's theory about Evolution "Survival Of Fittest". Genetic algorithm is an iterative procedure that represents its candidate solution as string of genes called Chromosomes. A group of individuals called population. Population is modified in the each iteration. Genetic Algorithm's iterations are called generation. Genetic Algorithm apply genetic operators such as selection, crossover and mutation. It generates solutions for successive generations. The genetic algorithm process terminates when an optimum solution is found.

6.**Proposed Methodology**

The Association rule mining is used to determine the association between web pages based on the keywords. The Discovery of association rules is based on the URLs. Exciting associations and relationship between web pages are determined by association rules. Web page contains huge amount of data items hence the Apriori algorithm is chosen for handling these web pages. The main goal of the proposed system is to apply genetic based apriori algorithm for web crawling.

Apriori Algoithm for Web crawling [7]

D:Database,collection of pages each involing a set of keywords
I:Item set,set of web pages
T:Transaction,where each T is an as set of items such that TCI
K:Item set that contains k items
C_k:set of candidate k-item sets.It has two fields support count and item set.
Lk:set of candidate k-item sets which have passed the minsup threshold value.
Apriori algorithm consider each item checks its support and rejects the item with support less than the minimum support and adds thereafter one more item with previous item one by one followed by check for the support and so on until the largest item set with support greater than minimum support is found at each iteration the crawler can keep the copy of the item .

Interestingness Measures

Support

The support Supp(x) is an itemset X is defined as the proportion of transaction in the dataset which contain the item set.

$$\text{Supp}(x) = \frac{number\_of\_transaction}{Total\_number\_transaction}$$

Confidence

The confidence is the conditional probability that, given X present in a transition, Y will also be present. Confidence measure, by definition:

$$\text{Conf}(x \rightarrow y) = \frac{\sup p(xUy)}{\sup p(x)}$$

Lift ratio

The lift ratio is the confidence of the rule divided by the confidence assuming independence of consequent from antecedent.

The Lift Ratio of an Association Rule is defined as follows

$$\text{Lift}(x \rightarrow y) = \frac{\text{support}(xUy)}{\text{support}(y) - \text{support}(x)}$$

Many other interesting measures have been proposed and studied in the literature [5][6].

Methodology

In this paper genetic algorithm is applied over the rules fetched from Apriori association rule mining for web crawling.

The proposed Algorithm is

1.Load the given dataset

2.Apply Apriori algorithm to find the frequent itemsets with the minimum support.

3.Apply genetic algorithm for generation of all rules.

4.Generate N chromosomes randomly depending upon dataset length..

5.Apply GA operators,crossover and mutation on the selected members.

6. Define the fitness function for specific support and confidence values.

Represent each frequent item set as binary string, Apply roulette wheel sampling method for selection operator. Fitness of each individual is computed as

If(rules_sup>=min_sup)&&(rules_conf>=min_conf)

Fitness=rules_sup*rules_conf

Else

Fitness=0

Experiment and Results

Initially ten documents are retrieved from the web. The keywords extracted from the documents are as shown below in table1.Keyowods and corresponding support count is calculated for each as shown in table2.

Table 1 Sample Data

URLs	Keywords
W1	Java,Class,Applet
W2	Java,Platform Independent
W3	Class,Applet
W4	Class,Thread,Java
W5	Java,Class,Thread,Object
W6	Java,Thread,Applet,Inheritance
W7	Java,Object,Language
W8	Thread,Class,Object
W9	Applet,Java.class
W10	Java,Applet,Thread,class

Table 2 Table of Support (Ck where k=1)

Keywords	Support
Java	8/10 = 80%
Class	7/10 =70%
Applet	5/10 = 50%
Platform Independent	1/10 =10%
Thread	5/10 =50%
Object	3/10 =30%
Inheritance	1/10 =10%
Language	1/10 =10%

The experimental results show the performance of our genetic based apriori algorithm is better than Apriori algorithm

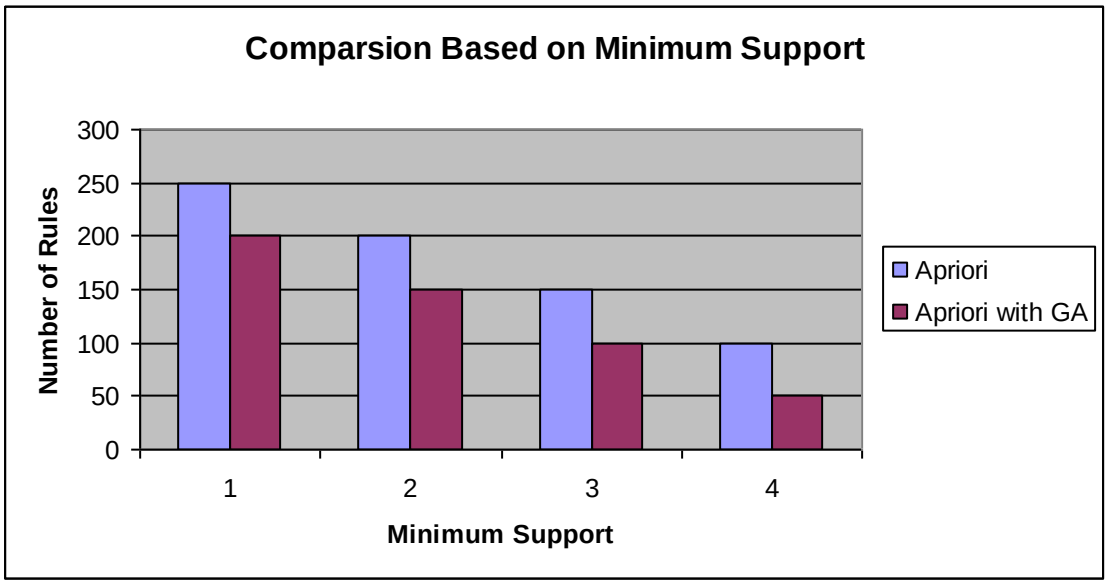

Figure 1 – Comparison based on minimum support

When the Support is increased the number of rules generated decreases. The application of the GA reduces the number of interesting rules generated.

7. Conclusion

A novel genetic based apriori algorithm was proposed for web crawling. Based on the empirical results collected, the proposed approach was found to yield good results when compared with ordinary apriori. As a part of future work, the genetic algorithm can be applied with various other fitness functions and parameters.

8. References

[1]Ramez Elmasri, Shamkant B.Navathe,"Fundamentals of Database Systems", PEARSON, fifth edition, 2009, 978-81-317-1625-0.

[2] Soumadip Ghosh,Sushantha Biswas,Dabasree Sarkar,Partha Pratim Sarkar,
Mining Frequent Itemseets Using Genetic Algorithm

International Journal of artifical Intelligence &Application (IJAIA)Vol,No 4,October 2010
[3] Charanjeer kaur,Association Rule Mining Using Apriori Algorithm:A Survey
Internationl Journal of advanced Research in Computer Enginerring & Technology(IJARCET) Volume2,Issue 6,June 2013
[4]Kalyanmoy Deb,"Introduction to Genetic Algorithms",Kkanpur Genetic Laboratory(KKangal),Depart of Mechanical Engineering IIIT Kanpur 2005
[5] Liqiang geng and Howard J.Hamilton,Interestingness Measures for Data Mining: A Survey, ACM Computing Surveys, Vol. 38, No. 3, Article 9, Publication date: September 2006
[6]S.Kannan and R.Bhaskar,Association Rule Pruning based on Interestingness Measures with Clustering, IJCSI International Journal of Computer Science Issues, Vol. 6, No. 1, 2009
[7]Extraction of Relevant Web Pages Using Data Mining, Shodhganga.infilbnet.ac.in/jspui/bitstream/10603/…/12_chapter%203.pdf
shodhganga.inflibnet.ac.in/jspui/bitstream/10603/.../12_chapter%203.pdf