International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

A Collective Survey on Methodology of Frequent Item Pattern Mining on Web Information

Sonali Abhane[1], P. D. Lambhate[2]

[1]M.E (Computer) Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Pune, India.
Savitribai Phule Pune University, Pune, Maharashtra, India -411007

[2]Professor (Computer) Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Pune, India.
Savitribai Phule Pune University, Pune, Maharashtra, India -411007

Abstract: ***In today's world of digital era finding semantic relationship between the information is really a big challenge for data analyzers. So most of the time this process needs the external help of the dictionaries to provide the semantic meanings but many times correlations between the words also plays more important role to get the best semantic. So many ideas are been proposed to get the semantic patterns in between the words of the document. But the proposed method put forwards as idea of getting the frequent item patterns from web pages using crawler which is been powered with the best frequent item algorithm like Éclat. This paper deals with a cumulative survey on the method of frequent item mining on web page information.***

Keywords: Web Crawler, Information Gain, Recursive learning, Eclat algorithm, Fuzzy Logic.

1. Introduction

Web crawlers are the program that goes recursively on given URL to find the relevant information. Numbers of crawling algorithms are.

- Breadth first search
- Depth first search
- Page Rank algorithm
- Genetic algorithm
- Naïve Bayes classifier

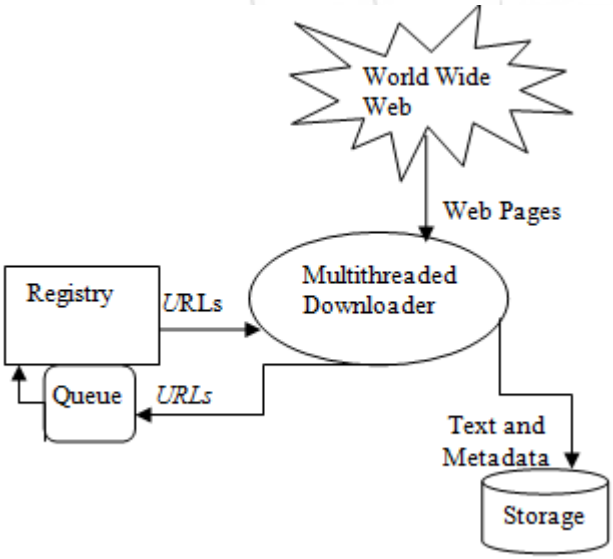

Figure 1: Basic crawling mechanism

Breadth First Search
These algorithms perform search operation on nearby node by starting from root. Once it got the thing for which searching is need to be done, success condition is returned and terminates the search operation. If the software unable to find the relevant information on the same level then it moves to the next level.

Depth First Search Algorithm
In this efficient technique of traversing system starts the traversing from the node and go through their children's. If multiple children's are there then the child on left side is selected first and then the child on the right side is selected.

Page rank algorithm
A page rank algorithm finds the rank of the all web pages of the web site. This ranking is done on the basis number of back links and the citations attached with the web page. Page rank algorithms plays important role as it gives the rank on the basis of usability, so it gets easy to decide the important web pages from the bulk of web pages.

Genetic Algorithms
Working of genetic algorithm is done on the basis of biological evolution.

Naïve Bayes classifier

Preprocessing
Preprocessing is a technique of reducing the amount of data by eliminating the things which are less required. Preprocessing plays important role in the data cleaning as it speed up the process to the great extent.

These classifiers are best on the learning and classification which is probabilistic. The main assumption of this algorithm is that one feature of the data is always independent of the other data. As discussed above the algorithm is best suit for the probabilistic models, it has a less applications in realistic models.

- Data Cleaning
- Data Integration
- Data transformation
- Data reduction

Data Cleaning: It is a technique of elimination outliers, smooth out noised data and adds the missed data. The main

task of the data cleaning is to find the most interesting data from the huge set of data.

Data Integration: As the name indicates, it is a technique of bringing the data from the multiple sources and stores it to the single location. As the data is at single position, system will have less cost of data retrieval.

Data Transformation: It is a process of converting the data from bulk to the appropriate format so it get easy to mine such data. Data transformation comprises of following sub techniques.

- Normalization
- Smoothing
- Aggregation
- Generalization

Data Reduction: Here the integrity of the data is managed while converting the data from complex format to the simple format. Data reduction mainly comprises of two techniques.

- Stemming
- Lemmatization

Fuzzy Logic: Fuzzy logic is a theory which is based on the multi valued logic in which truth values are lies between 0 to 1. Fuzzy logic is a good alternative to Boolean logic where only two values i.e. true and false are considered. Since fuzzy logic supports partial truth values it used on many applications where accurate values are not needed. The complete fuzzy logic methodology comes under four sub methods as,

- The Fuzzifier.
- A Rule Base.
- An Inference Engine.
- Defuzzifier.

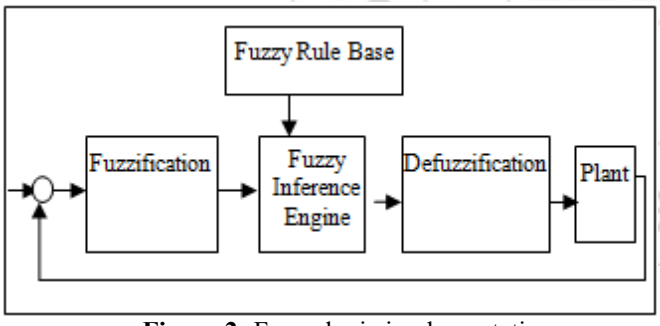

Figure 2: Fuzzy logic implementations

Fuzzy Rule: Fuzzy rules are the rules where output variables are considered. It is a normal IF-THEN rule with the condition and outcome. E.g. consider the fuzzy rule for the temperature management. IF (temperature is cold OR too-cold) AND (target is warm) THEN command is heat

Fuzzy Set Operations: Here the rules set in above step and the result of individual rule is combined. The operations performed on the fuzzy sets are completely different than the non-fuzzy sets. This combination process is also termed as inference engine.

Defuzzifier: The result of the fuzzy inference engine is fuzzy value only. Deffuzzification is a step to finding the crisp values. Numbers of algorithms are proposed for the deffuzzification, so one can use these algorithms as per their context.

Pattern mining: Pattern mining plays important role in data mining task. Association rule mining is an active area under research. It helps in finding the complex relations between the item sets. The format of association rule is as.

$X \in Y$ [Support= s%, Confidence=c%] where

1) Support s, is the probability that rule contains $\{X, Y\}$
   Support $(X \in Y) = P$ (XUY),
2) Confidence c, is the conditional probability that
   Specify the c% of the transactions of database
   Considered must specify $X \in Y$.

The rest of the paper is organized as follows Section 2 discusses some related work and section 3 is dedicated for conclusion.

2. Literature Survey

[1] Proposed a deep survey on various crawling strategies that are being existed. Here complete theory of crawling i.e. how crawling is happen? , how effectively crawlers can work? , how to increase the efficiency of the algorithm is well narrated by the author. To show the comparison between the systems numbers of systems are considered and thus they show that how one system can effectively perform over other. Tabular comparisons of 8 techniques are explained here. A Focused Crawler Based on Naive Bayes Classifier Proposed a new crawler which took Naive Bayes classifier as a base of their operations. Here to explore the features of the content of web pages TF-IDF algorithms are used. And to find the rank of web pages Bays classifiers are being used. Hence the proposed crawler performs better with respect to page ran crawlers and the BFS crawlers. Also the developers efficiently solved the TRAP problem which is normally observed in case of crawlers. The above crawler performs well on small data; its operations on large data are considered as a future work of their study.

Hits algorithm is one of the best methods used for the analysis of links. Techniques make use of this links to assign weight and authority to each web page of the network. But it has been observed that for some tree structure web pages algorithm fails to give precise outcome. [2] Narrates the technique which tries to overcome the disadvantage of above HITS algorithm. To accomplish this task a modification is done in the adjutancy matrix which is given as input to the algorithm. In the next part of the modification they changed the way by using weights is assign to the web pages. In modified algorithm they assign the weight by checking how frequently the web pages are used by the users.

To optimize the process of web crawling two step scheme is proposed by the [4]. In the first stage of process optimal crawls needed for each web page is determined. In the second part optimal achievable schedule is determined so that the crawling time will get reduced. With this reduced crawling time it gets easy to reduce the cost of crawler. To bring down the idea into reality the concept of probability function and resource allocation is considered. The main reason behind the use of this technique is their computational

efficiency which is very high. [5] Described a deep survey on the dimensionality reduction techniques i.e. preprocessing techniques used for the purpose of data reduction. The paper described all those method in lemon manner so it will get very easy for the newbies. Also the deep compassion between the lemmatization and stemming technique is done which gives better conclusion for the selection of one among them. [6]Explained the different preprocessing techniques normally observed in the data mining tasks. Apart from this mining 1 association rule is also performed using association algorithms. **[7]** WEKA is one of the most popular software used for the purpose of data mining tasks. [7] Highlights the different phase of WEKA from the day it released. The main motto behind the paper writing is the update which cam recently for the WEKA software. Here all the new features included in this software is highlighted so one can get easy understanding of the software use. [8] highlights the different techniques used for the purpose of web mining. Number of web mining systems such as semantic search, text and image retrieval, clustering, recommendation and many more are better elaborated here. For the proper understanding of the web mining a detailed graph is given by using one can easily understand the hierarchy. For the purpose of information retrieval fuzzy logic and genetic algorithms are combined and further effectiveness is explained. [9] Proposed a fuzzy logic for the neural network based decision system. The proposed model is completely differing from traditional models as it has advance deciding ability and complex network structure. The learning speed of fuzzy can be extensively increased by combining the supervised and unsupervised training sets. For explaining the feasibility of the system two applications are proposed. One is decision taking system and another is control system.[10] narrates the contribution of fuzzy logic in different domains such as machine learning, image processing, and pattern mining over a last 40 years. Here author states the bold contribution of fuzzy logic in these domains by giving the various case studies, applications and example. [11] Proposed a method of finding genes by considering the frequent patterns in clustering. Before this proposal numbers of algorithms were proposed, out of all that Apriori algorithm is the best one. But Apriori processing is done level by level, because of which processing cost is increases. To overcome above drawback a new system is proposed by [11] which makes use of Fuzzy FP-growth approach which completely eliminates the boundary problem observed in Apriori. Also the increase in number of genes will not going to degrade the scalability of the system. [12] Illustrates the weight of fuzzy logic in the process of interesting data mining by focusing the different important aspects. In spite of different contributions made in the area, authors have full confidence that still there are many areas where the technique can grow efficiently. So to do this they said that main focus should be given on the strengths of the FST and the right issue. Association rule mining is one of the best research areas of data mining. ARM can be effectively used for the purpose of finding the hidden relationship between the items because it is impossible to find such relationships using normal mining tasks. Association rules are normally having two categories as positive and negative. [13] Presents different techniques of finding the positive and negative association's rules. In this study several advantage and disadvantages of said techniques are figured out. [14] Proposed a brand new algorithm for transactional database to find the maximal frequent item sets. The algorithm performs the well when the numbers of datasets of the database are large in size. Here efficient pruning methodology is integrated with the DFS algorithm the fulfill the requirement of the search. By evaluating the system against various real word datasets, it has been found the system shows three to four times more performance than the previous systems. [15] Narrates the GenMax which is a back track searching algorithm efficiently used for the purpose of finding minimal item sets. To reduce the search space optimization of number is used. Two techniques named as progressive focusing and Deepset propagations are used. Deepset propagation is used to speed up the process of frequency computation. For the checking of maximality progressive focusing technique is used. Here GenMax is compared against the MaxMiner and MAFIA which is the state of the art methods for the frequent item set computation. After comparison it had been found that each of these methods has different performance as it is completely depends on the characteristics of the data being used. Also the comparison is done for the condition i.e. for which conditions algorithms behaves well and for which condition they show degraded performance. As the conclusion they said that MAFIA is a best method for finding the frequent item set but the GenMax is the best suit for detection of exact patterns.

To increase the efficiency of finding frequent item set transaction matrix based method is proposed by [16] which make use of transaction reduction. Normally the traditional association rule mining algorithms like Apriori generates the huge set of candidate item set which degrades the performance of the system. So to remove the said deficiencies [16] proposed a method known as MBAT. For the comparison purpose a tabular representation is given which clearly shows the difference between the Apriori and the proposed technique.

Table 1: Algorithmic comparision

Method	*Number of time database scanned*	*Number of candidate item sets generated*	*Computational time*
Apriori Algorithm	Large	Large	Large
Proposed Method	Only once	Very less as compared to apriori algorithm	Very less

3. Conclusion and Feature Scope

Frequent item mining is one of the most relevant used methods for pattern identification process. This paper puts light on frequent item mining on information of the web pages which are efficiently crawled from the web sites for further process using Éclat algorithm. This paper deeply surveyed on the many techniques that can be used on the web mining with pattern identification.

References

[1] Rahul kumar1, Anurag Jain2 and Chetan Agrawal3 "SURVEY OF WEB CRAWLING ALGORITHMS"

Advances in Vision Computing: An International Journal (AVC) Vol.1, No.2/3, September 2014.

[2] Wang, Wenxian, et al. "A focused crawler based on naive bayes classifier."*Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on*.IEEE, 2010.

[3] Miller, Joel C., et al. "Modifications of Kleinberg's HITS algorithm using matrix exponentiation and web log records." *Proceedings of the 24th annual international ACM SIGIR conference*

[4] Wolf, Joel L., et al. "Optimal crawling strategies for web search engines."Proceedings of the 11th international conference on World Wide Web.ACM, 2002

[5] Wu, C. L., K. W. Chau, and Y. S. Li. "Predicting monthly stream flow using data-driven models coupled with data-preprocessing techniques." *Water Resources Research* 45.8 (2009).

[6] "A Comprehensive Approach Towards Data Preprocessing Techniques & Association Rules", Jasdeep Singh Malik,PrachiGoyal, [3]Mr.Akhilesh K Sharma [3]Assistant Professor, IES-IPS Academy, Rajendra Nagar Indore – 452012 , India

[7] Hall, Mark, et al. "The WEKA data mining software: an update." *ACM SIGKDD explorations newsletter* 11.1 (2009): 10-18.

[8] Arotaritei, Dragos, and SushmitaMitra. "Web mining: a survey in the fuzzy framework." *Fuzzy Sets and Systems* 148.1 (2004): 5-19.

[9] Lin, Chin-Teng, and CS George Lee. "Neural-network-based fuzzy logic control and decision system." *Computers, IEEE Transactions on* 40.12 (1991): 1320-1336

[10] Mitra, Sushmita, and Sankar K. Pal. "Fuzzy sets in pattern recognition and machine intelligence." *Fuzzy Sets and systems* 156.3 (2005): 381-386.

[11] Mishra, Shruti, Debahuti Mishra, and Sandeep Kumar Satapathy. "Particle swarm optimization based fuzzy frequent pattern mining from gene expression data." *Computer and Communication Technology (ICCCT), 2011 2nd International Conference on*.IEEE, 2011.

[12] Hüllermeier, Eyke. "Fuzzy methods in machine learning and data mining: Status and prospects." *Fuzzy sets and Systems* 156.3 (2005): 387-406.

[13] Diti Gupta1, Abhishek Singh Chauhan2, "Mining Association Rules from Infrequent Itemsets:A Survey" International Journal of Innovative Research in Science,Engineering and Technology

[14] Burdick, Doug, Manuel Calimlim, and Johannes Gehrke. "MAFIA: A maximal frequent itemset algorithm for transactional databases." *Data Engineering, 2001.Proceedings.17th International Conference on*.IEEE, 2001.

[15] Gouda, Karam, and Mohammed Zaki. "Efficiently mining maximal frequent itemsets." *Data Mining, 2001.ICDM 2001, Proceedings IEEE International Conference on*.IEEE, 2001.

[16] Singh, Harpreet, and RenuDhir. "A New Efficient Matrix Based Frequent Itemset Mining Algorithm with Tags." *International Journal of Future Computer and Communication* 2.4 (2013): 355.

[17] Mbuso Gerald Dlamini, Yo-Ping Huang,Thanduxolo,Shannon Zwane, Siphamandla Dlamini "Extracting Interesting Patterns from E-commerce Databases to Ensure Customer Loyalty" Proceedings of 2015 IEEE 12th International Conference on Networking, Sensing and Control Howard Civil Service International House, Taipei, Taiwan, April 9-11, 2015

Author Profile

Prof. P.D.Lambhate, received her Degree from WIT, solapur, ME(Comp) from BVCOE Pune, Pursing PhD. In computer Engineering. She is currently working as Professor at Department of Computer and IT, Jayawantrao Sawant College of Engineering, Hadapsar, Pune, India 411028, affiliated to Savitribai Phule Pune University, Pune, Maharashtra, India -411007. Her area of interest is Data mining, search engine.

Sonali Abhane, is currently pursuing M.E (Computer) from Department of Computer Engineering, Jayawantrao Sawant College of Engineering,Pune, India. Savitribai Phule, Pune University, Pune, Maharashtra,India -411007. She received her B.E. (Computer) Degree from MKSSS Cummins College of Engg. For Women, Savitribai Phule Pune University, Pune, Maharashtra, India -411007. Her area of interest is programming languages & data mining.