***I.J. Education and Management Engineering,* 2018, 1, 40-47**
Published Online January 2018 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2018.01.05

Available online at http://www.mecs-press.net/ijeme

Quine-McCluskey: A Novel Concept for Mining the Frequency Patterns from Web Data

Bina Bhandari[a], R. H. Goudar[b], Kaushal Kumar[c]

[a]*Graphic Era Hill University, Dehradun, Uttarakhand, India*
[b]*Visvesvaraya Technological University, Belagavi*
[c]*Graphic Era University, Dehradun, Uttarakhand, India*

Received: 01 March 2017; Accepted: 11 September 2017; Published: 08 January 2018

Abstract

With the advancement in the web technology it is considered as one of the vast repository of information. However this information is in the hidden form. Various data mining techniques need to be applied for extracting the meaningful information from the web. In this paper the various techniques are discussed that have been used by many researchers for extracting the information and also shown the disadvantages with the existing approaches. The paper put forward a novel concept of mining the association rule from the web data by using Quine-McCluskey algorithm. This algorithm is an optimization technique over the existing algorithm like Apriori, reverse Apriori, k-map. This paper exhibits the working of the Quine- McCluskey algorithm that can extract the frequently accessed web pages with minimum number of candidate sets generation. However the limitation of Quine-McCluskey algorithm is that it cannot find the infrequent patterns.

Index Terms: Quine-Mccluskey Algorithm, K-Map, Apriori Algorithm, Users Access Pattern.

1. Introduction

The advancement in the web technology has given rise to tremendous amount of data at an exponential rate. The past research studies web shows that web is not only confined for sharing of information rather it has been used for different purposes by different others such as for predicting the future request of the users [13, 14], guiding the user by understanding past history of navigation [15]. One such application of web is used for building the business strategies. Therefore the designers of the web are interested in knowing the interest of

* Corresponding author.
E-mail address:

their customer [11] over the web so that the customers can be retained for the long time over the net. This can be achieved by acquiring the knowledge about the users surfing behaviour over the web that will further help in designing the web page as per the users' choice (Personalization [12], [27]) and recommending the links to the users based on their history of surfing on web (Recommendation). Understanding the behaviour of the web users has led to an increasing number of researchers to focus on it. The web is rich in information however the information it contains is in the hidden form Many researchers have applied the data mining techniques to extract the interesting information from the web. The web mining can be broadly classified into three categories: Web Content Mining, Web Structure Mining and Web Usage Mining [11]. Web Content Mining refers to the extraction of the text, audios, videos etc., structure mining refers to the mining of the links whereas web usage mining refers to extraction of web data which is generated via the client server interaction [16]. This is also known as Log Data or Web Data or click stream data[24]. Along with the users interaction with the server the log also stores the data which is generated by the system software like spider, crawler etc. therefore the log data needs cleaned. Many researchers have focused on data cleaning [24]. According to ([17], [18]) pre-processing is one of the most important process in web usage mining and consumes maximum time.

The data mining techniques are applied to the pre-processed data for extracting the actual navigation behaviour of the users so that the organizer of the web can recommend the links to the users based on their previous surfing history. One such mining used by researchers for mining the web data is association rule mining. Association rule mining ([19], [20]) aims to discover the potentially useful patterns from the web data. It is considered as one of the most vital technique and is well researched by many researchers. This technique was initially introduced by Agrawal et. al. in 1993 for the market basket problem. Association rule mining discovers the frequent patterns, association, correlations among sets of items in the large databases. Let P = $P1, P2, P3\ldots\ldots\ldots, Pm$ be a set of m distinct web pages, a transaction T that consists of set of web pages, D be a set of database which contains the number of transaction records and each transaction contains the different pages accessed by the users. Association rule is an implication in the form of $A \Rightarrow B$, where $A, B \subset P$ are sets of items (Web Pages) called item sets, and $A \cap B = \varnothing$. Here A is called antecedent and B is called consequent and the rule is A implies B.

Association rule mining framed the rule based on two important measures: *support* and *confidence*. The frequently accessed or mostly visited websites by the user can be extracted from the web data through association rule mining by taking minimum support and confidence. Where support for the association rule can be defined as the fraction of records that contain $A \cup B$ to the total number of records in all transactions. The confidence can be defined as the fraction of number of transactions that contain $A \cup B$ to the total number of records that contains A. Apriori is one of the well-known association based algorithm which is widely used for discovering the interesting patterns from the web data. Apriori algorithm is used by many researchers ([19], [20], [21], [22], [23]) due to the ease of understanding and implementation however later on they started working on the limitations of the algorithm. The drawback of Apriori algorithm was that it requires number of passes (scans) over the database; generates (k-1) candidate sets for the k-items and stores the counter for each candidate sets which results in comparatively more storage and processing time and moreover some of the items turn to be infrequent.

The k-map algorithm was used by authors for accessing the frequent and infrequent patterns from the web data. This algorithm converts the data into the tabular format in binary form and thus drastically reduces the scanning of the item sets as compared to the Apriori algorithm. The advantage of this algorithm was that it can find both types of patterns with lesser number of scans but also has certain limitations. This has drawn the attention of the researchers to search for new techniques of mining.

In this paper we used the Quine-McCluskey algorithm for discovering the association rule from web data. The Quine-McCluskey algorithm is based on brute-force method for discovering the main implicants. This method is used for the simplification of the Boolean functions and it is a computer based technique [1].

The contribution of this paper is that it focuses on three different techniques for performing association through the Apriori algorithm, K-Map algorithm and also discusses their limitations . This paper also put a novel concept of mining data through Quine-McCluskeyfor discovering the association among the items that are not connected to each other.

The paper is organized as follows section 2 contains the related work with respect to Web usage mining and association rule mining section 3 contains a discussions with the Apriori algorithm and K-Map, section 4 contains a brief discussion about the Quine-McCluskey algorithm, section 5show the working of the Quine-McCluskey (tabulation) method and finally section 6 concludes the paper work.

2. Related Work

The process of discovering the user navigation behavior from the web data is known as Web usage mining. Web usage mining focuses on understanding the behavior of the users over the web so that this information can be used in various applications of the web such as enhancing the quality of the web, web personalization [3],improving web structure and for improving the web server performance[4] etc.

The unremitting growth of web and its complexity has made the web more complex and hence increases the difficulty of identifying the users' access patterns over the web [28]. The author has proposed a novel method for identifying the users' access patterns. The method is based on Hidden Semi-Markov Model. The author has proposed a state selection algorithm which is based on k-mean clustering for improving applicability for the real websites.

Saglamet. al. in their paper discussed that companies have focused on understanding the requirements of their customers therefore they grouped the customers into different classes and set the potential customer to one class. Based on the potential customers' preference they developed the business strategies which will result in increasing the marketing share [5].

Enrique Lazcorretaet. al. has given a new technique for an automatic personalized recommendation system. The system is based on a single user behavior that is taken into account.The author has analyzed the technique of finding the association rules in the huge databases and also on converting the extracted data for the user-adapted recommendations with the help of two-step modified Apriori technique [6].Yun et. al. stated that the web is dynamic in contents and therefore analyzing the web leads to biases[7].

Paper [24] discusses that the amount of data generated through the social sites such as Facebook, flicker, LinkedIn etc. is in tremendous amount. This is the data that is generated through the client server interaction. Other than sharing of information people are also interested in doing business through the web. Therefore this paper uses Hadoop for mining the log file.

According to [25] preprocessing is one of the main part that gives the actual information about the users surfing behavior. This task covers removing the unwanted data (redundant data, data generated by system software etc.), cleaning of data, identifying the distinct users through the different technique such as time heuristic etc.ones the data is purified then association, classification and clustering techniques can be applied for discovering the users navigation behavior on web. The author has used the classification technique for classifying the users' category by using naïve Bayesian algorithm through the weka tool.

Bamshad Mobasheret. al. presented a scalable framework based on association rule mining for the recommendersystems by usingweb data. They presented data structure forcapturing the discovered frequent patterns that is good for the recommender systems. The algorithm given by them (recommendationalgorithm)produces the recommendations efficiently by utilizing the stored patterns without discovering all the rules from the item sets [27].

3. Discussion on Apriori and K-Map Algorithm

This section contains the discussion about the algorithm used by researchers (Apriori and K-Map) for discovering the association among the itemsets although they are not connected to each other. It also exhibits the limitations of the algorithms which have taken the attention of the researchers for discovering the new algorithms.

3.1. Apriori Algorithm

In the history of Association Rule Mining, Apriori was a great improvement. Apriori follows two steps for generating the frequent patterns from the large databases [23]:

i. Generating the candidate item sets and scanning the database for the support count of the corresponding item sets.
ii. Discovering the frequent item sets.

The algorithm and working of this algorithm is given in ([22], [23]).

Limitations: The drawback of this algorithm is that is quite time consuming and a tedious process to scan the entire database, it is costly to handle the number of candidate sets generated ([23], [29]).

3.2. K-Map Algorithm

The K-map method is used by the author for discovering the positive and the negative patterns. The positive patterns here refer to the frequently accessed web pages and the negative patterns refer to the infrequently accessed web pages. For example A implies B shows the association among the two pages whereas A does not imply B shows a negative relationship between the two pages. According to the [26], the algorithm overcomes the problem of number of scans (Apriori). This approach converts the database into the tabular format which is given in the binary format. The presence of the item is shown by “1” and the absence is shown by “0”. The algorithm and the working is given in our previous work [26]. The advantage of this algorithm is that it can extract both types of patterns.

Limitations: The limitation of the k-map based association rule mining algorithms is that it can perform well upto maximum of six variables. An increment in the variable will result in increasing the complexity of the algorithm[1].

From the above discussions we can conclude that the Apriori and k-map algorithms was not much suitable for the discovering the frequent item sets therefore we introduced a novel concept (Quine-McCluskey) algorithm for mining the frequent patterns from the databases.

4. Quine-McCluskey Algorithm

The Quine-McCluskey algorithm is an alternative method to the k-Map. However both the techniques are used for discovering the association among the item sets but the limitation of the Quine-McCluskey algorithm is that it cannot discover the negative patterns from the databases whereas the k-Map can discover the positive and the negative patterns. The Quine-McCluskey algorithm can be easily implemented in a computer program. The drawback of this algorithm is that this algorithm cannot perform well in terms of memory usage and processing time. Even additions of one or more variable will double the processing time and memory usage [2].The Karnaugh Map [8] and Quine-McCluskey [9] methods both are used for the simplification of the Boolean expression. Although the methods are same however the implementation of the methods are different .This is introduced as a substitute method to K-map. It’s has two steps. In the first step the truth table for the data set is generated and in the second phase the smallest set of prime implicants is taken based on a systematic procedure [10].

5. Working of the Algorithm

In this section the working of the Quine-McCluskey algorithm is shown. The experiment is performed on a small dataset of web data. Table 1 shows number of transactions in a database. In the next step the repeated items are removed from the table and the whole data is placed in the tabular format (in the binary form). Here A, B, C, D are the web pages. Now this tabular data is used for comparing the all pairs of items in the adjacent groups that checks for the change in value i.e. one and only one position. Now the new groups are formed based on the composite terms obtained, if there is a mismatch between the two position then replace it with a dash (-). Every time member of a group is combined with another group, both of them are then exempted from the list of implicants and are considered later [30]. This process continues until we get all the frequent patterns from the web data.

Table 1. Transaction in a Database

S.No	**TID**	**Itemsets**
1	T1	A
2	T2	A, C
3	T3	D
4	T4	B, C, D
5	T5	B, C, D
6	T6	A
7	T7	D
8	T8	C, D
9	T9	C, D
10	T10	A
11	T11	D
12	T12	A, C
13	T13	A
14	T14	B, C, D
15	T15	A, C

The redundant transactions are removed from the Table 1 and a restored in the binary format in the tabular format. The presence of web pages (pattern, item) is represented by “1 and the absence is represented by 0. This is process is done manually.

Table 2. Show the Presence of Item by 1

Web Page Transaction	A	B	C	D
1	1	0	0	0
2	1	0	1	0
3	0	0	0	1
4	0	1	1	1
5	0	0	1	1

In Table 3 the groups are formed. Then thee value for any of the one to one position is checked. The change in the value will add a dash (-) to that position where as similar values at both the position will yield either “1” or “0”. For example the grouping of (1,2) will give 1 at the first position, 0 at the second position, - at the third position and again 0 at fourth position. In the third position the value of C is first transaction is 0 and in second transaction it is 1.

Table 3. Set Generations Based on Item Presence

Web Page Transaction	A	B	C	D
1, 2	1	0	-	0
3, 5	0	0	-	1
4, 5	0	-	1	1

Table 4. Combination of Transactions

Web Page Transaction	A	B	C	D
1, 2, 3, 5	-	0	-	-
3, 4, 5	0	-	-	1

So on Table 4 groups the transactions (1,2,3,5) and (3,4,5), the patterns generated by these groups results in the following Equation (1)

$$\overline{B}+\overline{A}B \tag{1}$$

The Equation (1) shows that the most frequently accessed web page from the web data is D. The association among the item sets can contribute to the designing of the web page. This gives a clear picture about the users' choice by understanding his behaviour of surfing over the net.

6. Conclusion

In this paper we have given a novel method Quine McCluskey for discovering the association rule among the web pages from the web data. This algorithm improves the performance of the traditional methods such as Aprioriand reduces the number of scans drastically as compared to the Apriori algorithm. A comparative study is performed on Apriori, K-map and Quine-McCluskey algorithm which shows our novel concept works better than the rest two algorithms for discovering the association rule.This will give a better understanding about the users surfing behavior and hence this knowledge can be used for designing the web sites as per users' choice i.e. web personalization which can be further used for enhancing the business through the web.

References

[1] R .MohanaRanga Rao, "An Innovative Procedure To Minimize Boolean Function", (IJAEST) International Journal Of Advanced Engineering Sciences And Technologies, Vol. No. 3, Issue No. 1, 012 – 014.

[2] Hatim A. Aboalsamh, "A Novel Boolean Algebraic Framework for Association and Pattern Mining", WSEAS TRANSACTIONS on COMPUTERS, ISSN: 1109-2750, Vol. No. 7, Issue 8, pp. 1352-1361, August 2008.

[3] M. Eirinaki and M. Vazirgiannis, "Web Mining For Web Personalization", ACM Transaction Inter. Tech., Vol. 3, No. 1, pp. 1-27, 2003.

[4] J. Pei, J. Han, B. Mortazavi-Asl, and H. Zhu, "Mining Access Patterns Efficiently From Web Logs", in PADKK '00: Proceedings of the 4th Pacific Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications. London, UK: Springer-Verlag, pp. 396-407, 2000.

[5] B. Saglam, F. S. Salman, S. Sayin, “A Mixed Integer Programming Approach To The Clustering Problem With An Application In Customer Segmentation”, European Journal of Operational Research, 173, pp. 866–879, 2006.

[6] Enrique Lazcorreta, Federico Botella and Antonio Fernandez-Caballero, “Towards Personalized Recommendation By Two-Step Modified Apriori Data Mining Algorithm”,Elsvier Expert Systems with Applications 35, pp. 1422–1429, 2008.

[7] C. Yun and M. Chen, “Mining Web Transaction Patterns In An Electronic Commerce Environment”, In The fourth Pacific-Asia Conference On Knowledge Discovery And Data Mining, pp. 216–219, 2000.

[8] M. Karnaugh, “The Map Method For Synthesis Of Combinational Logic Circuits”, Transaction AIEE, pp. 593-599, 1953.

[9] E. L. McCluskey, “Minimization Of Boolean Functions”, Bell System Technical Journal, 35, pp. 149-175, 1959.

[10] G. De Micheli, “Synthesis and Optimization of Digital Circuits”, McGraw-Hill Science Engineering, 1994.

[11] Murat Ali Bayir, Ismail HakkiToroslu, Murat Demirbas and AhmetCosar, “Discovering Better Navigation Sequences For The Session Construction Problem”, Elsevier, Data & Knowledge Engineering 73, pp. 58–72, 2012.

[12] RanieriBaraglia, FabrizioSilvestri, Dynamic Personalization Of Web Sites Without User Intervention, Communications of the ACM 50, pp. 63–67, Feb 2007.

[13] Haibin Liu, VladoKeselj, “Combined Mining Of Web Server Logs And Web Contents For Classifying User Navigation Patterns And Predicting Users' Future Requests”, Data and Knowledge Engineering Vol. 61, Issue 2, pp. 304–330, 2007.

[14] Nizar R. Mabroukeh, Christie I. Ezeife, “Using Domain Ontology For Semantic Web Usage Mining And Next Page Prediction”, Proceeding of the 18th ACM Conferenceon Information and Knowledge Management, CIKM '09, ACM, New York, pp. 1677–1680, 2009.

[15] Hiroshi Ishikawa, Manabu Ohta, Shohei Yokoyama, Junya Nakayama and Kaoru Katayama, “On The Effectiveness Of Web Usage Mining For Page Recommendation And Restructuring”, Lecture Notes in Computer Science, Springer Berlin, Heidelberg, Volume 2593, pp. 253–267, 2003.

[16] Federico Michele Facca, Pier Luca Lanzi, “Mining Interesting Knowledge From Weblogs: A Survey”, Data & Knowledge Engineering, Vol. 53, Issue 3, pp. 225–241, 2005.

[17] KR Suneetha, DR Krishnamoorthi, Identifying User Behavior By Analyzing Web Server Access Log File”, .IJCSNSInternationalJournalofComputerScienceand Network Security Vol. 9, April 4, 2009.

[18] DoruTanasa, Brigitte Trousse, “Advanced Data Preprocessing For Intersites Web Usage Mining”, IEEEIntelligentSystems Vol. 19, pp. 59–65, (March–April (2)), 2004.

[19] R. C. Agrawal, T. Imielinski, and A. Swami, “Mining Association Rules Between Sets Of Items In Large Databases”, In Proceeding of ACM SIGMOD International Conference on Management of Data, pp. 207–216, 1996.

[20] R. C. Agrawal and R. Srikant, “Fast Algorithms For Mining Association Rules”, InProceeding of 20thInternational Conference In Large Databases, pp. 487–499, 1994.

[21] D. N. Goswami, AnshuChaturvedi, and C. S. Raghuvanshi, “An Algorithm for Frequent Pattern Mining Based On Apriori”, (IJCSE) International Journal on Computer Science and Engineering Vol. 02, Issue 04, pp. 942-947, 2010.

[22] B. Kotiyal, A. Kumar, B. Pant, R.H. Goudar, “User Behavior Analysis in Web Log through Comparative Study of Eclat and Apriori”, Proceedings of7'h International Conference on Intelligent Systems and Control, pp.421-426, 2013.

[23] Qiankun Zhao, Sourav S. Bhowmick, “Association Rule Mining: A Survey”, Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116, 2003.

[24] B. Kotiyal, A. Kumar, B. Pant, R.H. Goudar, “Big Data: Mining of Log File through Hadoop”, IEEE International Conference on Human Computer Interactions (ICHCI'13), at Saveetha University, Chennai, India, 23,24 August 2013.

[25] B. Kotiyal, A. Kumar, B. Pant, R.H. Goudar, “Classification Technique for Improving User Access on Web Log Data”, Proceedings of International Conference on Advanced Computing, Networking, and Informatics, India, June 2013.
[26] B. Kotiyal, A. Kumar, B. Pant, R.H. Goudar, “A Novel Concept For Mining Negative And Positive Rule Through Association Based K-Map”, Proceeding of International Conference on Mathematical Techniques in Engineering Applications,Oct 24-25, GEU, Dehradun, India (ICMTEA2013).
[27] BamshadMobasher, Honghua Dai, Tao Luo, Miki Nakagawa, “Effective Personalization Based on Association RuleDiscovery from Web Usage Data”, WIDM01 , 3rd ACM Workshop on Web Information and Data Management,Atlanta, Georgia, USA, November 9, 2001.
[28] C. Xu,C.Du, G.F.ZhaoandS.Yu, “A NovelModelforUserClicksIdentificationBasedonHiddenSemi-Markov”,Elsevier,Journal ofNetworkandComputerApplications, Vol. 36, pp. 791–798, 2013.
[29] B.Santhosh Kumar, K.V.Rukmani, “Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms”, Int. J. of Advanced Networking and Applications, Vol.01, Issue6, pp. 400-404, 2010.
[30] Reference for ECE320, “Quine-McCluskey's Method”.

Authors’ Profiles

Mrs Bina Bhandari is currently working as an Assistant Professor, Department of School of Computing, Graphic Era Hill University. Her area of interest incudes Data Mining, Big Data Analytics, Cyber Security and Cloud Computing. She has published a book on cyber security. She has also published papers in International Conferences and Journals on Data Mining and Big Data.

Dr. R H Goudar, currently working as an Associate Professor, Dept. of Computer Network Engineering, Visvesvaraya Technological University, Belagavi. He has 13 years of Teaching Experience at Professional Institutes across India. He worked as a faculty Member at International Institute of Information Technology, Pune for 4 years and at Indian National Satellite Master Control Facility, Hassan, India. He published over 145 papers in International Journals, Book Chapters and Conferences of High Repute. He has guided over 140 M.Tech Dissertation and 04 ongoing Ph.D Students. Dr R H Goudar has received various awards like Outstanding Faculty Award, Research Performance Award, Young Faculty Award from VIFA, and Young Research Scientist Award from VGST Karnataka. He has received research grants from AICTE, UCOST and VGST, Karnataka. He has received over 754 citations for the work in subjects of Interest includes Semantic Web, Cloud, Big Data, Network Security and Wireless Sensor Networks.

Mr. Kaushal Kumar is currently working as an Assistant Professor, Department of Electronics & Communication Engineering, Graphic Era University. His area of interest includes Wireless Sensor Networks, Image Processing, Wireless Communication, Digital Electronics and Digital signal processing.. He has published papers in International Conferences and Journals on Wireless Communication, Wireless Sensor Networks

How to cite this paper: Bina Bhandari, R. H. Goudar, Kaushal Kumar,"Quine-McCluskey: A Novel Concept for Mining the Frequency Patterns from Web Data", International Journal of Education and Management Engineering(IJEME), Vol.8, No.1, pp.40-47, 2018.DOI: 10.5815/ijeme.2018.01.05