www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 1 January 2015, Page No. 10180-10183

Different Methods of Classification of Documents A Survey Paper

Sneha K.Dehankar[1], K. P.Wagh[2] and P.N.Chatur[3]

[1]PG Scholar, Government College of Engineering,
Department of Computer Science and Engineering, Amravati.
INDIA dehankar.sneha@gmail.com

[2]Assistant Professor, Government College of Engineering,
Department of Information Technology, Amravati. INDIA
kishorwagh2000@yahoo.com

[3]Associate Professor, Government College of Engineering,
Department of Computer Science and Engineering, Amravati. INDIA
chatur.prashant@gcoea.ac.in

Abstract: ***The task of web mining is to classify documents automatically many algorithms have been developed to deal with automatic text classification The most common techniques used for this purpose include Apriori algorithm. In this survey, the various models using Apriori algorithm are explained to classify the text of documents, saving time and increasing accuracy of the web searching. It increases the data accessibility on the web giving the results faster. All the methods explained in this survey are executed to get the result accuracy for retrieving the web pages faster and getting the documents classified easily based on the probability calculated. The easiest algorithm used for classification of documents is Apriori algorithm used. The result of the survey of various methods used for classification shows that the search results are classified according to the classes they belong to. It is done on the basis of content matching of documents.***

Keywords: Apriori algorithm, Naïve Bayes classifier, Association rule, Genetic algorithm

1. Introduction

Web mining is based on the application of techniques of data mining as classification, clustering and association to search patterns from web. Web mining can be classified into web structure mining, web content mining and web usage mining. A web search engine is an application designed to seek out helpful information on World Wide Web.

The various techniques of Apriori Algorithm are used to classify the documents by using various data mining methods. The very first method used is preprocessing of the text in which the document is pre-processed and the stop words are removed for ex. a, an, the, like etc. Then the stemming is performed for removing the forms of the verbs like ing, ed forms. The next step is to classify the documents according to the classes, may be they are predefined classes or not. The documents are classified according to their contents. The classification is done by using various methods of data mining. In this survey, the very popular and easy to implement method is used called as Apriori algorithm.

2. Related Work

The various methods of classification of documents are used to classify the documents on the basis of their contents. If the contents of documents are matched, then the document is classified into the class it belongs to. The main and common method used for the classification of documents is Apriori algorithm. It is used to calculate the frequent itemset and then the different methods are applied on that result, such as to calculate probability of the frequent itemset obtained from the Apriori algorithm [3], the Naive Bayes classifier is used. It calculates the posterior probability of the itemset and then classifies it according to the class it belongs to. The frequent itemsets are generated using the Apriori algorithm based on the minimum support value.The frequent item is that one which is frequently accessed by the user. For ex.in retail sector the frequently purchasing habit of customer is considered and only the items which are having more demands are considered as the frequent items. Suppose the customer buys bread then he will also buy butter and milk too. So the frequent items are bread, butter and milk in this case. Now from these frequent items then the association rules[1] are generated. If the support value is greater than the minimum support value defined, then only the item is said to be frequent item else not.If the confidence value of the rules, generated from the frequent items is greater than the minimum confidence value then it is considered as mining of association rules.If the value of minimum confidence is increased then it also filters the rules and gives results more accurately.

3.Improved Apriori Algorithm

In this paper, the use of Apriori algorithm [1]is used to find out the useful hidden pattern of frequent items for the customer benefit in the retail sector. If the customer buys bread and butter then he will also buy the milk too. So in this case the

milk is considered as the hidden pattern of frequently used items by customer. Let the two items be A and B which customer usually buys. Support value for this will be given as Support = Tuples containing both A and B / Total no of tuples Confidence = Tuples containing both A and B/Tuples containing only A

The order of customer may vary it may be like, if he first buys butter and then bread or first milk and then bread and butter. But in association rule,[6]the order matters. In this use of Apriori algorithm is only to find the hidden pattern of frequent itemsets purchased by the customer for improving the business application. In traditional Apriori, most of the time is wasted for scanning the whole database searching on the frequent itemsets.In this the whole transactions are scanned for getting the frequent itemsets which increases the time required for scanning the database and also increases the complexity of the algorithm. The algorithm makes many searches in database to find frequent itemsets where k itemsets. In the first,the algorithm scan database to find frequency of 1-itemsets that contains only one item bycounting each item in database. The frequency of 1-itemsets is used to find the itemsets in 2-itemsets which in turn is used to find 3-itemsets and so on until there are not any more k-itemsets are used to generate k+1-itemsets. The main limitation is costly wasting of time to hold a vast number of candidate sets with much frequent itemsets, low minimum support or large itemsets.So it is costly and wasting of time of candidate generation. It will check for many sets from candidate itemsets, also it will scan database many times repeatedly for finding candidate itemsets.For overcoming from this drawback of Apriori algorithm the use of improved Apriori algorithm [2] is preferred.

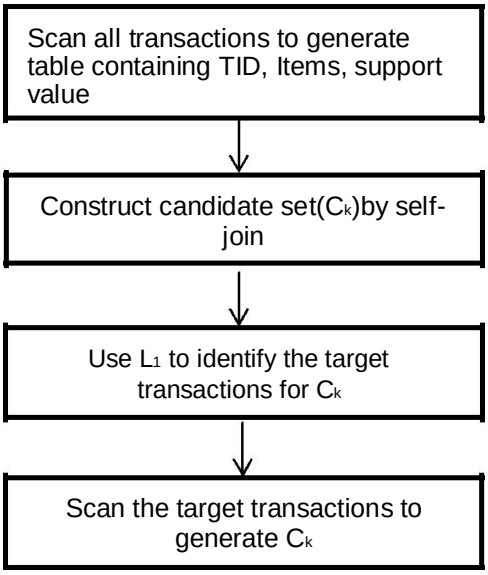

Figure 1: Steps for C_k generation

Firstly scan all transactions to generate L1 table which contains the items, their support count and Transaction ID where the items are found and then use L1 later as a helper to generate L2, L3 ... Lk. While to generate C_2, we make a self-join L1 * L1 to construct 2-itemset C (x, y), where x and y are the items of C_2. Before scanning all transaction records to count the support count of each candidate, use L1 to get the transaction IDs of the minimum support count between x and y, and thus scan for C_2 only in these specific transactions.The same thing for C_3, construct 3-itemset C (x, y, z), where x, y and z are the items of C_3 and useL1 to get the transaction IDs of the minimum support count between x, y and z, then scan for C_3only in these specific transactions and repeat these steps until no new frequent itemsets are Identified. It will save the time required to scan the whole database only some specific part of database will be scanned. Suppose we are having the transactions from T1 to T9 having items from I1 to I5 in it.

Table 1: TID and Items

T id	Item
T1	I1, I2, I5
T2	I2, I4
T3	I2, I4
T4	I1, I2, I4
T5	I1, I3
T6	I2, I3
T7	I1, I3
T8	I1, I2, I3, I5
T9	I1, I2, I3

The support value for each item is then given by:

Table 2: Items with support value

Item	Support value
I1	6
I2	7
I3	5
I4	3
I5	2

Let the minimum support value be 3.so item I5 will be deleted. First the frequent items are calculated and the frequent 1 itemset is generated then frequent 2 itemset is generated and atlast frequent 3 itemset is generated.In traditional Apriori algorithm,the all transactions T1 to T9 are scanned to find the frequent 2 itemset consisting of I1 and I2.But in improved Apriori ,split the item (I1, I2) into I1 and I2 and get the minimum support between them using L1 table which consist of items, support value ,transaction ids of items. Here I1 has the smallest minimum support. After that search for itemset (I1, I2) only in the transactions T1, T4, T5, T7, T8 and T9 because item I1 is found in these transactions only. Thus it reduces the search area of database reducing the time and complexity too.The main difference in traditional and improved Apriori algorithm is found in this work. The difference in transactions to be scanned for traditional and improved are: 1-itemset 45 and 45 2-itemset 54 and25 3-itemset 36 and14 Total 135 and 84 respectively. In this paper, an improved Apriori is proposed through reducing the time consumed in transactions scanning for candidate itemsets by decreasing the number of transactions to be scanned. Whenever the k of k-itemset increases, the gap between improved Apriori and the original Apriori increases from view of time consumed and whenever the value of minimum support increases, the gap between improved Apriori and the original Apriori decreases from view of time consumed.

4.Apriori Algorithm With Naïve Bayes Classifier

The most common method used for classification of web documents is Apriori Algorithm. For classifying the documents into different classes Naive Bayes Classifier is used.

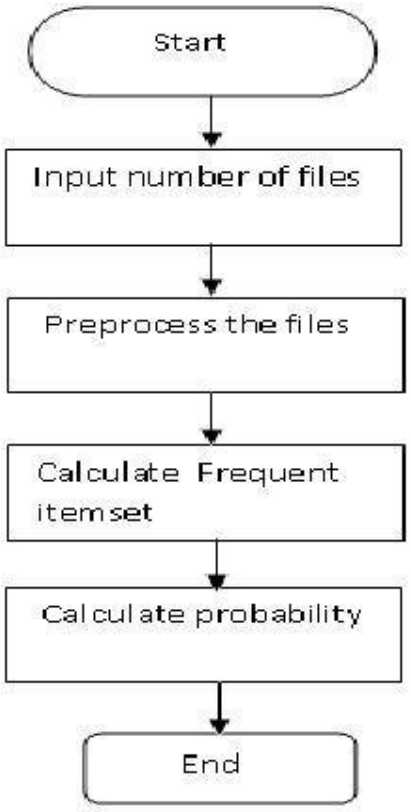

Figure 1: Flowchart of work

Apriori Algorithm[3] finds interesting association among a large set of data items by finding the frequent itemsets using association rule mining. After calculating the frequent itemsets the preprocessing of the document is done to remove the stopwords. Then the documents are classified based on their contents into the predefined classes. The predefined classes are defined first to classify the documents according to their contents. Then, use Naive Bayes Classifier to calculate probability of keywords among a large data itemsets to classify document based on highest probability. First the preprocessing of the document is done in which all the unwanted words are removed such as like,a, the or etc. These words are called as stopwords. After removing the stopwords, the frequent items are calculated using the Apriori algorithm. The frequent items are that words which are occurred frequently in the document. The number of occurrences of the word is counted called as the support value of the item. The minimum support value is predefined and if the support value of the item is greater than the predefined minimum support value then it is taken as the frequent item otherwise not. . In this the main two steps are used to calculate the frequent items. The first is the join step and other is prune step. If the occurrence of the word is found to be greater than minimum support value then it is considered as the frequent itemset belonging to the frequent wordset else prune step is applied which removes the word.Once the frequent itemset is generated then from that rules are formed by using the confidence value called as association rule mining.

When the frequent itemsets are calculated using the algorithm then the final categorization of document is done based on the probability value calculated by using naive bayes classifier. Then the document is classified into the predefined class to which its contents are matched. In this the prior probability is calculated it means the number of words found in the document to the total number of words in the document. After that the word count is calculated which is the number of occurrence of the word in the document in the particular class which are predefined first. In these respective classes the documents are classified based on the probability calculated. After that the word likelihood is calculated for the respective classes. It is calculated as word count divided by the prior probability calculated for that class. The very last step is then calculation of posterior probability. For this the document is divided into matrix format containing 0 and 1 values. If the frequent keyword is found then 1 is put into the feature vector else 0 is placed. The various document models are used for calculation of probability. Bernoulli document model is used in this paper. The posterior probability is calculated by using this model. In this the formula makes use of prior probability, word likelihood calculated previously. It is calculated for all the predefined classes and the it is then checked, whether which probability value is greater than the other one.If the probability value comes greater for the first class than as compared to the other one then that document is classified into the first class because its probability calculated is greater than all the remaining classes.In such a method all documents are classified into the respective classes based on their content matching.

5. Conclusion

In this survey paper, the different efficient techniques for web page classification are introduced. Techniques require more or less data for training as well as less computational time of these techniques.It fastens the access of web document searching over the web to save the time and complexity of web pages.

References

[1]Jiao Yabingò Research of an Improved AprioriAlgorithm in Data Mining Association Rulesò International Journal of Computer and Communication Engineering, Vol. 2, No. 1, January 2013

[2] Mohammed Al-Maolegi, BassamArkokò An Improved Apriori Algorithm For Association Rulesò International Journal on Natural Language Computing (IJNLC) Vol. 3, No.1, February 2014

[3] Keyur J. Patel1, Ketan J SarvakaròWeb Page Classification Using Data Miningò, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 7, July 2013

[4] S.M.Kamruzzaman,FarhanaHaider,AhmedRyadhHasan ñText Classification Using Data MiningòICTM 2005.

[5] Chowdhury Mofizur Rahman and Ferdous Ahmed Sohel and ParvezNaushad and S M Kamruzzaman, ñ Text Catagorisation using Association Rule and Naivey Bayes classifierò, CoRR2010

[6]Jugendra Dongre1 GendLalPrajapati s. V. Tokekar ,ò The Role of Apriori Algorithm for Finding the Association Rules in Data Miningò, 978-1-4799-2900-9/14/$31.00 2014 IEEE, 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)

[7] Nandagopal, S., 2V.P. Arunachalam and 3S. Karthik ñMining of Datasets with an Enhanced Apriori Algorithmò Journal of Computer Science 8 (4): 599-605, 2012 ISSN 1549-3636 © 2012 Science Publications

Author Profile

S. K. Dehankar hasreceived her B.E. degree in Information Technology from Shri Sant Gajanan Maharaj College of Engineering, Shegaon Maharashtra, India in 2012.Currently pursuing M.Tech degree in Computer Science and Engineering from Government College of Engineering, Amravati, Maharashtra, India. Her research interest includes data mining. At present she is engaged with web document classification using data mining techniques.

Prof. K. P. Wagh has received his Diploma in Computer Engineering from Government Polytechnic Jalgaon. BE(CSE)from Government college of Engineering Aurangabad. ME (CSE) from Walchand College of Engineering Sangli. Presently he is working as Assistant Professor in Information Technology Department at Government College of Engineering,Amravati,Maharashtra.

Dr. P. N. Chatur has received his M.E degree in Electronics Engineering from Government College of Engineering Amravati, Maharashtra, India and Ph.D. degree from Amravari University. He has published twenty papers in international journals. His area of research includes Artificial Neural Network, Data Mining, Data Stream Mining and Cloud Computing. Currently, he is Head of Computer Science and Engineering & Electronics Engineering Department at Government College of Engineering Amravati, Maharashtra, India. At present he is engaged with large database mining analysis and stream mining.