INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD ISSN: 2455-0620 Volume - 6, Issue - 4, Apr – 2020
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value: 86.87 Impact Factor: 6.497
Received Date: 28/03/2020 Acceptance Date: 11/04/2020 Publication Date: 30/04/2020

COMPARATIVE STUDY OF TERM BASED PATTERN TAXONOMY DEPLOYING ALGORITHMS

[1]Dr. S. Brindha, [2]Dr. S. Sukumaran
[1]Assistant Professor, [2]Associate Professor
[1, 2]Department of Computer Applications,
[1]Vellalar College for Women, [2]Erode Arts and Science College, Erode, Tamilnadu, India
Email - [1] brindha.balajiee@gmail.com

Abstract: *The internet is a powerful platform as the data repository that plays a great role in storing, allocation, and recovers in sequence for information discovery. However, as there are countless, dynamic and important expansion of data, web users face big problems in terms of the relevant information needed. Accordingly, poor information precision and retrieval are part of the hottest recent research areas in today's world. Even though the capacious of information resided on the web, valuable informative knowledge could possibly be discovered with the application of advanced data mining techniques. Association mining is one way to discover frequent patterns from different data sources. In this paper, three of the foremost association rule mining algorithms used for frequent pattern discovering. This paper explains the comparisons of the algorithms are namely Apriori, FP-growth, Pattern mining algorithms used to sets of transactional databases devised from server access log file.*

Key Words: *FP-growth, Pattern Mining, Natural Language Processing, Apriori.*

1. INTRODUCTION:

Pattern mining helps on identifying system so as to explain exact patterns within the data. Of most interest is the discovery of unexpected associations, which may open new avenues for marketing or research. Another important use of patter mining is the discovery of sequential patterns. Knowledge discovery can be viewed as the process of important extraction of information from large databases, information that is completely presented in the data, formerly unknown and potentially useful for users. A fundamental step of data mining in the process of knowledge discovery in databases. In the past years, a notable number of data mining techniques have been presented to perform various tasks for knowledge. These techniques include association rule mining, sequential pattern mining as well as closed pattern mining. Most of these techniques are suggested for the purpose of developing efficient mining algorithms to discover accurate patterns within an equitable and suitable time frame. With a large number of patterns generated by using data mining approaches, how to effectively utilize and update these patterns is complicated problems, we focus on the development of a knowledge discovery technique to effectively update the discovered patterns and apply it to the field of text mining. Text mining is the discovering interesting knowledge in text documents. It is a complicated issue to discover accurate knowledge in text documents to help in users search for to fine what they want. In Information Retrieval (IR) provided many term-based methods to solve this issue, such as Rocchio and probabilistic models, rough set models, BM25 and support vector machine (SVM) based filtering models [12]. The utilization of term based methods include efficient performance as well as sophisticated theories for term weighting, which have developed over the last couple of decades from the IR communities. The term-based methods suffer from the problems of polysemy and synonymy, where polysemy implies a word has different implications, and synonmy is different words having the same implication. The semantic importance of numerous found terms is unverifiable for noting what clients need. There are two fundamental problems regarding the viability of pattern based approaches: low frequency and misinterpretation problems. A highly frequent pattern is usually a general pattern, or a specific pattern of low frequency. In the event that we diminish the minimum support, a more number of noisy patterns would be discovered. Misinterpretation implies the measures utilized as a part of pattern mining (e.g. "support" and confidence) turn out to be not relevant in using discovered patterns to answer what users want. The critical problem hence is how to utilize the discovered patterns to accurately evaluate the weights of useful patterns in text documents. Many terms with bigger weights for example the term frequency and inverse document frequency are general terms because they can be frequently used in both suitable and unsuitable information.

2. LITERATURE REVIEW:

Textual documents are increasingly added to the World Wide Web and also the electronic databases of organizations. One of the representations which are well known is known as bag of words approach that makes use of keywords. Tf*idf weighting scheme is presented in [1] for representing text. In [2] entropy weighting and global IDF are used for text representation in addition to DFIDF. For the approach bag of words various schemes were developed

for weighting [3], [4], [5]. The drawback in the bag of words is that choosing limited number of words is a problem thus it causes over fitting [6]. To reduce number of features other approaches came into existence. They include Odds ratio, Chi-Square, Mutual Information, and Information Gain [4], [6]. Though there are many representations, the choice of representation is based on the requirement, the rules of natural language [6]. Some researchers used phrases instead of words. Unigram and bigram combination is also used in the text categorization process. Phrase based approach is explored in [6]. Data mining techniques are also used as explored in [7]. There was no significant improvement in text mining when phrases are used. It suffered from lower frequency and misinterpretation problems [8]. Some insights were provided on ontology mining which is again term based [9], [2]. In [1] a technique known as pattern evolution was introduced. Algorithms such as GST [2], SLPMiner [3], SPADE [6] etc. are used for the purpose of data mining. However, discovering interesting patterns is still open to anyone to research [5], [6]. In [15] a two stage model was developed. The two stages include pattern based methods and term based methods. For text mining "Natural Language Processing" concepts are used. Recently a new model known as concept-based model came into existence [8], [9]. Conceptual Ontology Graph is also explored in order to use semantic knowledge in the discovery of patterns. This model provides effective discrimination between meaningful terms and important terms. Natural language processing (NLP) is a modern computational technology that can help people to understand the meaning of text documents. For a long time, NLP was struggling for dealing with uncertainties in human languages. This model included three components. The improvement of the concept-based model is that it can successfully distinguish among non significant terms and meaningful terms which describe a sentence meaning. The concept-based model usually relies upon its employed NLP techniques.

3. SYSTEM ARCHITECTURE

Term selectivity determines the scope of the term focuses on the topic that user's requirements. It is very complicated to compute the precision of terms because a term's specificity depends on user's perspectives of their information needs. The first definition of the specificity of terms because a term's specificity which calculated the score of a term based on its appearance in discovered positive and negative patterns. In agreement with the diffusion of terms in training set. The RFD method is able to precisely appraise term weights according to equally their specificity where the higher level features include both positive and negative patterns. It requires manually setting two empirical parameters according to hard sets. The RFD model and empirically demonstrate that the proposed specificity function is logical and the term classification are successfully approximated by a feature clustering method. Design an inclusive approach for evaluating the text document methods. That is, term evaluation giving in document. Mostly TBM suffer from the problems of polysemy and synonmy. Polysemy means a word has multiple meanings. Same as multiple words having the similar name meaning is known as synonymy. In Term Based Method every expression within document is connected by means of value recognized as weight that determines significance of term because terms contribution in document. Expression having semantic denotation is acknowledged as term as well as collection of such terms contributes meaning to document. Term based methods go through as of the troubles of polysemy as well as synonymy. Polysemy means a word has multiple meanings as well as synonymy is multiple words having the same meaning. The semantic meaning of many discovered terms is uncertain for answering what users exactly want. Information recovery provided many term-based methods like supervised as well as traditional term weighting methods to solve this challenge. The term frequency TF (t, d) is number of times term 't' occurs in document 'd'. The document frequency DF (t) is number of documents in which term 't' occurs at least once. The inverse document frequency IDF (t) is calculated from document frequency. The information of term pharse consisting of one or more specific terms together with corresponding to dependent terms. Pharse are often chosen by using word frequency counts as well as other statistical methods, possibly supplemented by syntactic procedures designed to detect syntactic relationships between governing as well as dependent pharse components[13].

$$\mathrm{IDF(t)} = \log\left(\frac{|D|}{DF(t)}\right)$$

The utilize of word grouping methods of the kind provided the uses where clauses of related words are grouped under common headings, these headings can then be assigned or content identification instead of the individual terms contained in the classes. Term relationships constructive for content identification may also obtainable by utilizing existing machine readable. |D| is total number of documents. The inverse document frequency of term is low if it occurs in many documents as well as is highest if term occurs only in one document. The value of Wi i.e. weight of term of document 'd' calculated by product as,

$$Wi = TF(ti, d) * IDF(ti)$$

This is also the central interest in most of the web personalized applications, and it has received more attention from researchers. There are two challenging issues in using pattern mining techniques for discovery relevance features in both related and unrelated documents. The first is the low-support problem. Long patterns are usually appearing

more specific for the topic, to maintain. If the smallest amount sustain is decreased, a lot of noisy patterns are discovered. The second problem is the misinterpretation problem, which revenue the measures used in pattern mining turns out to be not suitable in using patterns for solving harms. For this method, a greatly frequent pattern may be a general pattern because frequently used in both related and inappropriate documents. There are several existing methods for solving the two searches out issues within text mining. Pattern Taxonomy
Mining methods have been proposed, in which mining closed sequential patterns in text paragraphs and deploying them over a term space to weight useful features. Over the years, people have residential many full-grown term based techniques for ranking documents, information filtering and text classification. Newly, more than a few hybrid methods were planned for text classification. Pre-processing activities plays a vital role in the various applications. The present work uses three important pre-processing techniques namely stop word removal, stemming and spell Stemming algorithm which works dynamically applied for any domain. Thus our approach is designed effectively to overcome the problem of Named Entity. Also, it is necessary to cover various subject disciplines. Working with text mining applications, often hear of the term “stop words” or "stop word list" or even "stop list". There as on why stop words are critical to many applications is that, if remove the words that are very commonly used in a given language, can focus on the important words instead.

Term Frequency Inverse Document Frequency (tf-idf) is a numerical statistic which reveals that a word is how important to a document in a collection. The TF- IDF is often used as a weighting factor in information recovery as well as text mining. The value of tf-idf develops proportionally toward the numeral of period a word appears in the document, other than is counter acting through the occurrence of the word in the corpus. This can facilitate to organize the information so as to some words are normally additional common than others. Tf-IDF is effectively used for stop-words filtering in a variety of subject fields counting text summarization as well as classification. TF-IDF is the items for consumption of two information are described as a term based frequency and inverse document frequency. On the way to additional differentiate, the numeral of times every expression occurs in every article is counted as well as arithmetic them each and every one collectively.

$$TF(t)= (quantityoftimestermtdisplayinadocument)(Totalnumeraloftermsinthedocument)$$

Term Frequency (TF) is distinct as the quantity of periods a term occurs in a document. Inverse Document Frequency as well as Inverse Document Frequency is a statistical weight used for measuring the importance of a term in a text document collection[14]. IDF feature is incorporated which reduces the weight of terms so as to happen extremely repeatedly in the document set as well as increases the weight of conditions so as to take place hardly ever. Then Term Frequency-Inverse document frequency is intended for every word utilizing the formula. The frequency of the occurrence of term t in document d. TF-IDF is intended for every terms in the document through utilizing Term Frequency (tft,d) as well as Inverse Document Frequency (idf,td). The inspiration is that discovered patterns that contain supplementary semantic meaning than the terms that are selected based on a term based technique. In term based approaches, the assessment of term weights supports are based on the distribution of terms in documents. The assessment of term weights is dissimilar to the regular term-based approaches. In deploying method, terms are weighted according in the direction of their appearances in exposed closed patterns. The search engine tries to discover web pages that contained the terms “how”, “to” “develop”, “information”, “recovery”, “applications” the search engine is going to discover a lot more pages that contain the terms “how”, “to” than pages that enclose information about developing information recovery applications because the terms “how” and “to”. This is just the basic perception for utilizing stop words. Stop words are usually assumed to be a “solitary set of words”. Some applications removing all stop words right from particular word e.g. the, a, an. To prepositions e.g. above, across, before and to some adjectives e.g. good, nice and an appropriates top word list. Stop words are basically a set of commonly used words in any language. Stop words is critical to many applications is that, if remove the words that are very commonly used in a particular language, focus on the important words instead.
Determiners - To mark nouns where a determiner usually will be followed by a noun examples: *the, a,an, another.*

Step 1: *Input positive and negative documents*
Step 2: *Start*
B1= positive document
B2 = negative document
Step 3: *If (B1== value1){*
Print(“click on positive document to choose”);
Ps[T]Discover Paragraph(D) // Ps Paragraph
Loop j=2:t // t is the term
*Tfidf(t,f,d)= {tf (t,d) * idf (t,d)}// t is term, d document}*
Else {
Step 4: *If(B2==value2)*

Print("click on negative document to choose");}
If (b1==b2){
Print("select different path to decide positive as well as negative documents");}
Else {
Step 5: *Discovering Terms in documents using Term based Deploying*
Step 6: *Select both documents to compared}*
if (b1==b2) {
Wd(t)={-dc_sup(t,D-)(1+|spe(t)|) // wd Weight}
Else{ result is true positive and result is accurate term Deploying }
Step 7: *Repeat the process until the end of the document*
Step 8: *End*

Testing part discovers differences in positive/negative documents by the centroid obtained in training phase by ranking each of them. The simple method is to calculate the similarity between two documents. Each and every document is preprocessed with word stemming and words removal into a set of transactions based on its nature of document structure. System chooses one of pattern taxonomy algorithm to citation of the pattern based documents. Require that the pattern should occur in all sequences or there are a minimum number of occurrences specified by the user. In some cases the number of occurrences is not specified but it is a part of the scoring function a longer pattern with fewer occurrences is sometimes more interesting than a shorter pattern with more occurrences. A Set of Textual Documents are arranged likewisedc1,dc2,dc3….etc., and the words of the Documents Words wd2,wd3…etc.

dc1 -- wd2 wd3
dc2 -- wd4 wd5 wd6
dc3 -- wd4 wd5 wd6 wd7
dc4 -- wd4 wd5 wd6 wd7
dc5 -- wd2 wd3 wd6 wd7
dc6 -- wd2 wd3 wd6 wd7
dc7 -- wd1 wd2 wd3 wd4

Closed sequential patterns extracted from positive documents capture prices of meaningful information for describing user information needs. Frequent Words are described here like wd4,wd5,wd6 etc., Covering set of the particular document as dc2, dc3, dc4 etc.,

{wd4,wd5,wd6} {dc2,dc3,dc4}
{wd4,wd5} {dc2,dc3,dc4}
{wd4,wd6} {dc2,dc3,dc4}
{wd4} {dc2,dc3,dc4,dc7}
{wd6} {dc2,dc3,dc4,dc5,dc6}
{wd2,wd3} {dc1,dc5,dc6,dc7}
{wd2} {dc1,dc5,dc6,dc7}
{wd3} {dc1,dc5,dc6,dc7}

Factors that move the site level correctness are exact low, while the theme level accuracy is relatively high, which indicates that the algorithms can usually capture at least one correct design in an input sequence.

SID	Sequence
1	$\langle\{a,b\},\{c\},\{f,g\},\{g\},\{e\}\rangle$
2	$\langle\{a,d\},\{c\},\{b\},\{a,b,e,f\}\rangle$
3	$\langle\{a\},\{b\},\{f,g\},\{e\}$
4	$\langle\{b\},\{f,g\}\rangle$

This database contains four sequences. A sequence is an ordered list of item sets (sets of items bought together. Traditionally, sequential pattern mining is being used to discover subsequences that appear often in a sequence database, i.e. that are common to several sequences. Those subsequences are called the frequent sequential patterns. For example, in the context of our example, sequential pattern mining can be used to discover the sequences of items frequently bought by customers. This can be useful to understand the behavior of customers to take marketing decisions.

4. RESULTS:

The TBPTDM evaluated using various datasets. The TBPTDM are evaluated for single document using three metrics such as MAP, IAP, Min_sup values are displayed in the table. It is used to found the result of TBPTDM exceeds all other methods. TBPTDM gives better value compared to PTM methods.

Table 1 The Comparison Table of TBPTDM for Single Document

Methods/Metrics	MAP	IAP	Min_sup
PTM	0.19	0.12	0.20
FPM	0.10	0.14	0.18
TBPTDM	0.19	0.15	0.26

Data mining involves of extracting information from data stored in databases to understand the data and/or take decisions. Nearly the most important data mining tasks are clustering, classification, outlier analysis, and pattern mining. Pattern mining contains of defining inspiring, appropriate, and amazing patterns in databases various types of patterns can be discovered in databases such as frequent itemsets, associations, subgraphs, sequential rules and periodic patterns. The assignment of sequential pattern mining is a data mining task specialized for analyzing sequential data, to discover sequential patterns. It consists of discovering interesting subsequences in a set of sequences, where the interestingness of a subsequence can be measured in terms of various criteria such as its occurrence frequency, length, and profit. Sequential pattern mining has frequent actual presentations outstanding to the fact that data is naturally encoded as sequences of symbols in many fields such as bioinformatics, e-learning, market basket analysis, texts, and webpage click-stream analysis.

5. CONCLUSION

The method TBPTDM for patterned the document. In the proposed method there are three different types of algorithms are used to identify the values of the document. Relevant document and retrieved documents can also be classified utilizing this method. It is found that the proposed method gives the better performance results than the other two methods. The metrics such as MAP, IAP, Min sup can also be calculated to discover the values of various pattern taxonomy methods and compared those results with various datasets values.

REFERENCES:

1. Witten, I. H., Frank, E.: Data mining: practical machine learning tools and techniques. Morgan Kaufmann (2005).
2. Zida, S., Fournier-Viger, P., Lin, J. C.-W., Wu, C.-W., Tseng, V.S.: EFIM: A Highly Efficient Algorithm for High-Utility Itemset Mining. Proc. 14th Mexican Intern. Conf Artificial Intell., pp. 530-546 (2015).
3. Fournier-Viger, P., Lin, C.W., Duong, Q.-H., Dam, T.-L.: PHM: Mining Periodic High-Utility Itemsets . Proc. 16th Indust. Conf. Data Mining, 15 pages (2016).
4. G. K. Gupta, Introduction to Data Mining with Case Studies, PHI 2006.
5. Dr.S.Brindha,Dr.S.Sukumaran "Relevance Pattern Discovery for Text Classification using Taxonomy Methods",IJSART-Volume4 Issue 11, ISSN Online: 2395-1052, Pp:321-326.November 2018.
6. Yanchang Zho, R and Data Mining: Examples and Case Studies, Elsevier, December 2012.
7. V. Van Laer and L. De Raedt. How to Upgrade Propositional Learners to First Order Logic: A Case Study. In [16], pages 235–261, 2001.
8. S. Wrobel. Inductive Logic Programming for Knowledge Discovery in Databases. In [16], pages 74–101, 2001.
9. Y. Yang, S. Slattery, and R. Ghani. A study of approaches to hypertext categorization. Journal of Intelligent Information Systems, 18(2-3):219–241, 2002.
10. B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In Proc. of UAI-02, pages 485–492, Edmonton, Canada, 2002.
11. Dr. S .Brindha, Dr.S.Sukumaran "A Comparative Study on Pattern Mining Techniques" in International Conference on Computer, Electrical, Electronics, Management and Information Technology, March 11, 2016. This Paper was published in International Journal of Advanced Research Trends in Engineering and Technology (IJARTET), Volume 3, Issue 15, March 2016.
12. Dr. S. Brindha, Dr.S.Sukumaran "A Survey on Classification Techniques for Text Mining"in3rd International Conference on Advanced Computing & Communication Systems January 22, 2016. IEEE Publisher Electronic ISBN: 978-1-4673-9206-8 Print on Demand (PoD) ISBN: 978-1-4673-9207-5 INSPEC Accession Number: 16359465 DOI: 10.1109/ICACCS.2016.7586371 10th October 2016.

13. Dr. S .Brindha, Dr.S.Sukumaran "The Comparison of Term Based Methods Using Text Mining" in International Journal of Computer Science and Mobile Computing (IJCSMC), Volume 5, Issue 9, (ISSN 2320-088X) September 2016.
14. Dr. S. Brindha, Dr.S.Sukumaran "Pattern Document Weight Discovery For Text Classification Mining" in IEEE Conference on Communication and Electronics Systems (ICCES 2016), ISBN:978-1-5090-1065-3), 22nd October 2016.
15. Dr. S. Brindha, Dr.S.Sukumaran "Pattern Taxonomy Term Based Model for Text Document Classification" in International Journal of Engineering Development and Research (IJEDR 2017), 2017 IJEDR | Volume 5, Issue 1 | ISSN: 2321-9939, March 2017.

About the Authors:

Dr. S. Brindha received B.Sc degree in Science from Bharathiyar University. She done her Master Degree in Information Science and Management in Periyar University and she awarded M.Phil Computer Science from the Bharathiyar University. She received the Ph.D degree in Computer Science from the Bharathiar University. She has 5 years of teaching experience and 6 years of Technical Experience in Hash Prompt Softwares Pvt. Ltd. At present she is working as Assistant Professor of Computer Applications in Vellalar College for Women, Erode, Tamilnadu, India. She published around 13 research papers in International Journals and Conferences Her Research area includes Data Mining and Image Processing, Pattern Taxonomy Mining.

Dr. S. Sukumaran graduated in 1985 with a Degree in Science. He obtained his Master Degree in Science and M.Phil in Computer Science from the Bharathiar University. He received the Ph.D degree in Computer Science from the Bharathiar University. He has 28 years of teaching experience starting from Lecturer to Associate Professor. At present he is working as Associate Professor of Computer Science in Erode Arts and Science College, Erode, Tamilnadu, India. He has guided for more than 55 M.Phil and 10 Ph.D Research Scholars in various fields. Currently he is Guiding 6 Ph.D Scholars. He is a member of Board studies of various Autonomous Colleges and Universities. He published around 75 research papers in National and International Journals and Conferences. His current research interests include Image Processing and Data Mining.