Mining Effective Patterns From Text Data- A Survey

M.Thangaraj, K.Sundaravadivelu

Abstract: Text mining is one of the efficient techniques of data mining to discover and extract valuable information from textual data. The role of text pattern mining in large datasets is an interesting area of research in knowledge discovery. There are so many techniques like predictive analytics, prescriptive analytics, decision making problems that rely solely on patterns to achieve their respective goals. Extracting interesting patterns especially from documents and other unstructured data is a boon considering the data explosion due to social media. Patterns depict the inherent knowledge and other trends such as, frequent item sets, closed frequent item sets and co-occurring terms. The usage and improving the semantic value of patterns that almost matches the real world scenario is still a research gap in this field. If the patterns are long its high specificity leads to low-frequency and misinterpretation problems. Patterns are depended on individual terms which further suffer from dimensionality issues such as polysemy and synonymy. This paper aims to find out the effective algorithms for discovering customized patterns from large documents and also provides solutions to the problems mentioned above. It further compares strengths, and limitations of various algorithms using review of literature and recommend most suitable methods. The knowledge about various pattern mining techniques will be useful when implementing data models such as frequent item sets, closed sequential pattern, pattern taxonomy model and pattern deploying model.

Index Terms: Text mining, Pattern taxonomy models, Sequential, Closed Pattern mining, Trend abstraction

1 INTRODUCTION

Text analytics [20] is concerned with uncovering unstructured data and providing it a form suitable to elicit meaningful content, in other words, knowledge discovery from a large volume of data. The techniques improve that conventional Extract, Load and Transform (ETL) pipeline of data mining. For efficient exploration of knowledge contained in every data text analytics are inevitable. They form a bridge between data and all other applications that run on top of data. Manually analyzing huge volumes of data to find hidden knowledge is a tedious work and also impossible. It also helps to extract specific points and entities of interest thereby providing a personalized text mining process. The rate of data accumulation in the form of unstructured data is on an increasing trend for the past few decades. Data is in fact exploding with more and more dimensions in an exponential way. The tools that could handle such enormously growing data whether it is in structured and unstructured form is a valuable asset, as data is the new oil, data mining becomes the force that controls the data. Text mining has its applications in almost every field wherever data is generated. Some of the most famous applications of text mining are, healthcare analytics, recommender systems, social media analytics, educational data mining, government information mining, security, law etc. Pattern mining [1] is the procedure to mine effective trends present in the data. It is one of the techniques in data mining concerned with extracting most related entities from the object of analysis. The data model that is used to derive patterns of interest is called pattern taxonomy model. The patterns and their inherent relationships are exhibited in hierarchical manner. Knowledge discovery can be further enhanced if pattern taxonomy model is used to update the extracted patterns from text documents.

- *M.Thangaraj is currently the Professor and Head, Department of Computer Science, Madurai kamaraj University , Tamilnadu, India. He is an active researcher in Big Data Analytics, Social Media Analytics ,Wireless Sensor Networks and has published more than 100 papers in Journals and Conference Proceedings. E-mail: thangaraj-dcs@mkuniversity.org*
- *K.Sundaravadivelu is a part time research scholar and Assistant Professor of computer science , Department of Computer Science, Madurai kamaraj University, Tamilnadu, India. E-mail: sundarmku04@gmail.com*

These patterns could be used to analyze trends in large text documents and uses the trend information to achieve classification tasks related to text mining. There are different types of patterns present in text data, for extracting each type, a different technique is used. Some of the techniques are, mining frequent item sets, sequential pattern, association rules, closed patterns etc. Using the patterns that are extracted by various approaches and forms of updating them are open research areas. This work surveys the techniques present in pattern mining with respect to text data. Text data is taken up for this study because they are based on individual terms. Term based models provide efficient performance by using theories for term based methods using term weighting. There are also certain issues that need to be addressed like synonymy and polysemy, where a word has multiple meanings and multiple words will have similar meaning. Similar to the issue mentioned above, there are other issues in pattern mining techniques. This work aims to review the research problems and also provide solutions to them.

2. RELATED WORK

Pattern taxonomy model [1] is used in a work to attain pattern discovery where it takes semantic dataset. While using this technique, the semantic information is difficult to process and also the data was highly ambiguous and discriminated the individual terms. Patterns obtained could be cleaned with Co-occurrence matrix and Pattern Deploying based on Co-occurrence weight and absolute Support (PDCS) approach [2].The pattern co-occurrence matrix helps to clean sequential patterns, again co-occurrence weighting is used to achieve to process pattern deploying. Absolute support overcomes misinterpretation problem. The patterns are evolved to avoid problem of low frequency. The issue associated with this work is the matrix tries to find only the semantic relationships between only the patterns obtained and not in the general dataset. Another work deals with using pattern discovery to retrieve documents relevant to a particular query [3]. Pattern Taxonomy Model (PTM) is used to solve the IR problem. It suffers from the problem of achieving effective taxonomies for documents that contained multiple concepts. It also

failed to address the problem of unseen terms.[4] There is one more study where relevant features are used for discovering effective text patterns. It uses low-level terms as features using their appearance in higher level patterns and a respective specificity in the training dataset. Relevance Feature Discovery model suffers to perform when less training data is available to assess the quality of data. [5] Pattern co-occurrence matrix is also used in another approach to solve low- frequency problem. It simply processes the patterns after deriving them and clean close sequential patterns, however it considered only positive documents into consideration. D-patterns algorithm [6] is used to mine interesting patterns from text document is studied. Restructuring supports of terms in the normal forms of D-patterns present in the documents was not achievable. [7] Another work where a TF-IDF weight method based visualized pattern mining is analyzed. The system did not consider the inter-relationships in the layers to attain pattern discovery. To discover relevant features, top-k specific patterns are deployed [8]. IT summarizes the patterns in text documents to extract smaller subset of patterns that become the top-k specific patterns. This also provides wider coverage of various patterns. However, the study took only a limited features set of 11 closely related feature terms. [11] Naive bayes algorithm issued to obtain effective patterns using two processes is studied. One is, pattern deploying and another one is, pattern evolving. Here, the naive bayes algorithm is used to classify terms in the document which further enhanced pattern discovery. It used D-pattern mining techniques to achieve pattern discovery. Extracting patterns from unstructured data is a challenge in pattern discover. Pattern matching technique is used to effectively extract expression that better represent relevant words[12]. It used Regular Expression algorithm to achieve the patterns but, only small number of documents were considered. Patterns from clinical records is a open research problem due to their highly complex nature. [16] This knowledge from medical records are identified by the specific semantic nature using pattern based approach.

3. SURVEY ON PATTERN MINING APPROACHES

This work studies the various pattern mining approaches. The research directions with respect to pattern mining are specified as a flow diagram in Fig 1.Pattern mining is broadly classified into four types, term based approach, phrase based approach, pattern based approach and pattern evolving and deploying approaches. In term based approach, TF-IDF classifier and bag of methods. In phrase based method unigrams, bigrams and NP chunking are used. While in Pattern based approach, algorithms like Apriori, GSP SPADE, pattern-growth, freespan and prefix span. Finally, in pattern evolving and deploying approaches uses SPM, PTM and PCM techniques

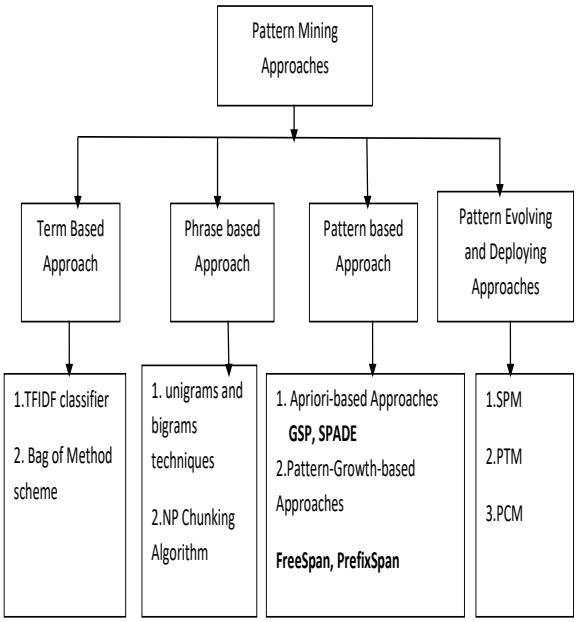

***Fig : 1** Various Approaches on Pattern Mining*

3.1Term based Approach

Many of the existing techniques in information retrieval adopt term based approaches [2]. Terms are considered as bag of words , where an entity is represented as a group of terms in information retrieval. The major drawback of this method is, the relationship among various features or words cannot be taken into account. And also there exist a semantic ambiguity in terms that are synonymous. Another peculiar problem with bag of words approach is their ability to handle homonyms. [9]Homonyms are terms that have multiple meanings. Some of the term based information retrieval approaches are, TF-IDF classifier, SVM, Rocchio Algorithms etc., All the techniques have to be improved to avoid classification errors.

3.2 Phrase based Approach

Phrase includes semantic information therefore they are far better than terms [5]. The ambiguity is reduced since they are a group of terms. Also the low frequency problem is solved with this approach however suffers from noisy data when slangs interfere with original phrases. [14]Sequential patterns are best alternative to phrase based techniques because of the statistical nature. In [23], rule based Natural Language Processing was proposed to analyze web documents to extract phrase based representations was proposed. It calculated weights for n-grams which in turn could be approximated by the use of unigram, bigram and trigram during word dependencies. [22] Unlike sequential patterns, n-grams obtain patterns without gap between words. Dealing with large dataset that contain noisy patterns is difficult problem to be solved in this regard.

3.3 Pattern based Approach

These are mainly concerned with relevance feedback and pattern discovery. [24] Its aim is to improve information retrieval of exact information specified by the user. An automatic pattern taxonomy extraction model is used in web mining for uncovering useful phrases is obtained[15]. Here, frequent sequential patterns are used to do noise pruning. It

is basically a hierarchical tree that exhibits relationships present in the data. Pattern basd methods solve the problem of overfitting, by eliminating irrelevant patterns. The work [4] implemented algorithms like, Apriori, PrefixSpan, FP-tree, SPADE, SLP miner and GST are proposed. Always the patterns are huge, containing redundant and noisy data. The major challenge with pattern discovery is dealing with large pattern set and handling noisy pattern that are irrelevant and reduce the accuracy of the data model . These pattern mining techniques have an edge over the previously discussed phrase-based technique. Pattern mining techniques are good for, data misinterpretation and pattern low frequency problems. Large patterns offers accurate knowledge about data but they still suffer from less frequent observations.

4. PATTERN EVOLVING AND DEPLOYING APPROACHES

In order to solve the problems encountered in pattern based approaches, pattern evolving and deploying approaches were proposed. [10] This updates the already obtained patterns using pattern taxonomy, deploying and evolving techniques. It cleans the obtained patterns, to achieve clear cut patterns and remove all the little noise that are very prominent in large pattern sets. The semantic meanings are also not always exhaustive in answering user queries. To solve this problem, relevance feature discovery [19] for pattern evolution technique is explored. It considers both positive and negative documents and treats them as high level features. It allows for accurate and specific patterns using term weighting. it analyzes closed sequential pattern mining to extract the appearance of terms by using high level features over low- level features. It uses specificity measure to model low level features in high level features to classify terms into positive terms, general terms and negative terms. Though these techniques involve large pattern sets in the studies, the performance remains the same without any improvement.

4.1 Problems in Pattern Discovery Approaches and their solutions:

For term based methods, there are two major problems like polysemy and synonymy. The semantic elements in data introduce ambiguity in interpretation. To avoid this, patterns are extracted instead of terms from large documents. However, this approach also has some problems like misinterpretation and low level of pattern frequency. This problem is caused by low minimum support values which lead to smaller patterns. Misinterpretation is caused due to polysemy. To avoid these problems, pre-processing techniques are enhanced to contain a word list of related terms. Inner pattern evolution is used to assign weights to terms that reduced noise in patterns obtained. One more interesting direction in pattern discovery is large patterns that are more specific to the topic but they occur in low frequencies. To avoid the problem of low frequency due to decreased minimum support pattern evolution technique is used. It also includes negative training samples to tune them to suppress noisy patterns. Misinterpretation problem used measures, confidence and support which are not suitable using discovered patterns to answer user queries. Terms with larger weights are more general and they are frequently used in both irrelevant and relevant information. So, it is not sufficient to evaluate term weights based on their distribution in documents for a specific topic. An appropriate solution is to deploy pattern deploying process to weight terms. This also improves accuracy of term weighting processes since, the patterns derived are better than whole documents. Co-occurrence weighting solves misinterpretation problems and produce effective discovery of patterns by reducing low frequency of new patterns. The keyword based approaches always have some drawbacks when compared to term based and phrase based techniques. Phrases offer more information about a term along with its semantic orientation compared to single terms. But again phrase based approaches have setbacks in the form of, substandard statistical properties of terms, frequency of patterns (phrases) is too low to produce enough patterns and also contain multiple noisy phrases where redundancy is also an issue to be noted. To solve the problem a new system was proposed which focused mainly on the knowledge discovery and the efficient use of the patterns which are discovered at the end and apply it to the field of text mining. There are multiple techniques for solving text mining problems. In Pattern taxonomy mining (PTM) models mining is done by means of closed sequential patterns existing in text paragraphs and effective deployment of them on a term space to allocate weights on most useful features. Concept-based model (CBM) was proposed to find out theories with the help of natural language processing techniques. It is concerned with verb-argument structures to obtain different concepts from sentences. These patterns can also be called as concepts, were proved to be highly effective, when compared with term based techniques. Integrating patterns from both suitable and unsuitable documents is still an issue in this regard. To solve this RFD model could be used to improve specificity of extraction and make the model more reasonable this could further be approximated by a feature clustering technique. Pattern Co-occurrence Matrix [10] applied after PTM remove the less important patterns to reduce low frequency problem to achieve effective discovery of patterns.

5. DISCUSSION

In the field of text mining, pattern extracting techniques can be used to find various interesting textual patterns that will maximize the accuracy of classification are used. Some of the techniques discussed in this study are frequent item-sets, co-occurring terms, multiple grams, and sequential patterns to construct a better representation using all these new features. This work, gives recommendations to use a particular type of pattern discovery technique based on the type of data used. The review analyzes major issues in the field of pattern discovery approaches and suggests suitable solutions. Some the recommendations of the survey are, sequential patterns are more effective than word and phrase oriented approaches for large documents. When training set is small, term based approach could be taken, and when the document is rich in natural language phrase based approaches are the best fit. Pattern based approaches may not always outperform term based approaches because of the presence low frequency and pattern misinterpretation problem. In all large patterns though the relevancy score is more, the observations are rare. This could be improved by pattern evolution.

Misinterpretation problem can be solved by including negative training samples. to tune term support and noisy patterns.

6. CONCLUSION

Pattern discovery is tool for text mining that helps to enhance the various mining tasks like, clustering, classification, prediction etc. A survey about tools of a major filed is highly valuable in achieving major tasks like, knowledge discovery, Thus, this survey about the various literature generated for pattern discovery proves useful in identifying various research gaps and probable solutions for them. The knowledge about these techniques will help both the researchers and practitioners in choosing the best technique for a particular problem. The contributions of the review states the major problems in extracting effective patterns from text data and also offers solutions for the same.

7. REFERENCES

[1] V.Aswini, S.K.Lavanya," Pattern Discovery for Text Mining", 2014 ICCPEIC.

[2] Gangarde Rupali,.Kolhe. V.L," Effective Pattern Discovery by Cleaning Patterns with Pattern Co-occurrence Matrix", 978-1-4799-3486-7/14///2014 IEEE.

[3] Shivani D Gupta, B.P.Vasgi," Implementation of pattern Discovery to retrieve relevant document using text mining",978-1-4673-7910-6/15///2015 IEEE.

[4] Li Yuefeng, Algarni Abdulmohsen, Albathan Mubarak, and Shen Yan,"Relevance Feature Discovery for Text Mining", IEEE Transactions on Knowledge and Data Engineering. 10.1109/TKDE.2014

[5] Rupali R," Pattern Co-occurrence Matrix to Reduce Low Frequency Problem and Effective Pattern Discovery", 2016 International Conference on Computing, Analytics and Security Trends. Pune, India. Dec 19-21, 2016.

[6] Dasri Yugandhara Bapurao, Nalwade Prakash Shivajirao, 2017." Effective Pattern Discovery For Text Mining Using PTM and PDM", International Conference on Intelligent Computing and Control Systems.

[7] Nan Yu," A Visualized Pattern Discovery Model for Text Mining Based on TF-IDF Weight Method", 2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics, 978-1-5386-5836-9/18 2018 IEEE DOI 10.1109/IHMSC.2018.10148

[8] Luepol Pipanmaekaporn, Yuefeng Li, Shlomo Geva," Deploying Top-k Specific Patterns for Relevance Feature Discovery", 2010 International Conference on Web Intelligence and Intelligent Agent Technology. IEEE ,DOI 10.1109/.2010.194

[9] Raphael Polig, Kubilay Atasu, Christoph Hagleitner," Token-Based Dictionary Pattern Matching For Text Analytics", 978-1-4799-0004-6/13///IEEE

[10] Ning Zhong, Yuefeng Li and Sheng-tang Wu. 2012. "Effective Pattern Discovery For Text Mining", IEEE Transactions On Knowledge And Data Engineering.

[11] Kavitha Murugeshan, Neeraj RK," Discovering Patterns to Produce Effective Output through Text Mining Using Naïve Bayesian Algorithm", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-2, Issue-6, May 2013

[12] Anujna M, Ushadevi A" Converting and Deploying an Unstructured Data using Pattern Matching", American Journal of Intelligent Systems 2017, 7(3): 54-59 DOI: 10.5923/j.ajis.20170703.03

[13] Shaikh Rijwan, Mane Nandkumar, Ankush Vyavhare, P.S.Patil," Effective Pattern Discovery for Text Mining By Pattern Deploying and Pattern Evolving", International Journal of Innovative Research and Creative Technology , Volume 1 | Issue 5,2015, ISSN: 2454-5988

[14] Bharate Laxman, Sujatha. 2014.," Improved Method For Pattern Discovery In Text Mining", International Journal of Research in Engineering and Technology.2321-7308

[15] Dipali Sonawane, Tejal. Shirole. Patil, Priyanka V. Patil and Amol K. Patil.2017 " Effective Pattern Discovery for Text Mining", International Research Journal of Engineering and Technology, Volume: 04 Issue: 04.

[16] Oleg Metsker, Ekaterina Bolgova, Alexey Yakovlev, Anastasia Funkner, Sergey Kovalchuk ," Pattern-based Mining in Electronic Health Records for Complex Clinical Process Analysis ",6th International Young Scientists Conference in HPC and Simulation, YSC 2017, 1-3 November 2017, Kotka, Finland , Procedia Computer Science 119 (2017) 197–206, Elsevier

[17] Luepol Pipanmaekaporn and Yuefeng Li. 2012. " A Pattern Discovery Model for Effective Text Mining", pp. 540–554. Springer.

[18] R. Sharma and S. Raman. 2003 "Phrase-based text representation for managing the web documents",International Conference on Information Technology Coding and Computing, 165–169.

[19] L. Pipan maekaporn, 2013 "Feature discovery in relevance feedback using pattern mining," Computer and Information Science .pp. 301–307, IEEE.

[20] Hussein Hashimi , Alaaeldin Hafez, Hassan Mathkour(2015)," Selection criteria for text mining approaches", Computers in Human Behavior (2015), http://dx.doi.org/10.1016/j.chb.2014.10.062

[21] Dursun Delen , Martin D. Crossland," Seeding the survey and analysis of research literature with text mining", Expert Systems with Applications 34 (2008) 1707–1720.

[22] S. Quiniou, P. Cellier, T. Charnois, and D. Legallois. sequential data mining techniques to identify linguistic patterns for stylistics. Computational Linguistics and Intelligent Text Processing, pages 166–177, 2012, Springer

[23] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic pattern taxonomy extraction for web mining," Web Intelligence, 2004., IEEE.

Mubarak Albathan, Yuefeng Li, Abdulmohsen Algarni, 2012." Using Patterns Co-occurrence Matrix for Cleaning Closed Sequential Patterns for Text Mining", ACM International Conferences on Web Intelligence and Intelligent Agent Technology