Cluster based Text Mining

Nikita R. Andhrutkar[1] Prashant M. Yawalkar[2]
[1]PG Student [2]Associate Professor
[1,2]Department of Computer Engineering
[1,2]MET's Institute of Engineering Nashik, Maharashtra (Affiliated to SPPU) India

Abstract— In text mining, discovery of relevance features (RFD) is the challenging task as term-based approach is used for it. A term based approach suffered from the problems of polysemy and synonymy. Pattern based method is also introduced in previous systems which perform better than term based approach. In proposed system, relevance feature discovery model is introduced which discovers positive and negative patterns from given dataset of text documents. Pattern taxonomy mining i.e. PTM along with n-gram is proposed for pattern discovery. The proposed model describes relevance feature in three groups such as, positive, negative and general. F-Clustering algorithm describes the feature clustering approach in which set of positive documents DP+ and set of negative documents DP- get sorted. Term based classification is given as a contribution of proposed work.

Key words: Text Mining, Text Feature Extraction, Text Classification

I. INTRODUCTION

To discover useful features from text documents relevance feature discovery i.e. RFD model is used. It discovers the relevant and irrelevant features involved into given document dataset. To find the relevant and irrelevant text is challenging task in the domain of information analysis in both empirical and theoretical perspective. Such problems are occurred into the domain of data mining, machine learning, IR systems and web intelligence communities. From study of existing systems there are two major problems have been observed, first is the low-support problem and the other is the problem of misinterpretation. In low support problem is one in which long patterns are more specific but there appearance in documents is very low or frequent. Misinterpretation is used in pattern mining which is not suitable in using patterns for problem solving. Several methods have been introduced previously to solve challenging issues in text mining. In Pattern taxonomy mining (PTM), closed sequential patterns in text paragraphs get mined using natural language processing techniques (NLP). A concept-based model (CBM) is proposed in [4] for verb argument structure to discover structure of topics in given sentences which depicts an importance as well as effectiveness of each text . In proposed system, RFD model is used which demonstrates the conservatives and the language classification through feature clustering method. To evaluate the performance of proposed work, comprehensive approach is also designed. The proposed technique is innovative to discover and classifying low-level terms based on both their appearances in the higher-level patterns as well as their specificity in a training set.

Two types of algorithms have proposed for identification of closed sequential patterns such as, Fclustering() and WFeature() algorithms. WFeature() feature describes relevance features for specified topic and to distinguish topic from other topics, specific terms are used whereas, FClustering() algorithm describes feature clustering process. In the process of feature clustering, set of positive and set of negative documents get combined i.e. DP^+ and DP^- . Based on the above analysis, we can develop a clustering method to group terms into three categories automatically for each topic by using the specificity function.

II. RELATED WORK

F. Song et al. discussed about a language model. It is based on range of data smoothing having great-Turing estimate, functions of curve-fitting and model combinations for information retrieval. It is conceptual simple and intuitive model. It simply extended to arrange the presumption of phrases containing pairs as well as triples of words. This approach is developed for smoothing of data which can be easily fit into proposed language model. It is also known as the general framework for language based information retrieval [1].

F. Sebastiani discussed text categorization (TC). Test categorization the classification of texts. From document indexing based on a controlled vocabulary, to document filtering, automated metadata generation it can be utilized in different areas. Knowledge engineering constructs the set of rules on how to classify documents in specified categories. In the field of text categorization researchers detects challenging applications in which dataset of thousands of documents and categorized by tens are widely used. [2].

S. Tang Wu et al. have suggested two approaches based on pattern deploying approach. They made investigation of their performance on Reuter's dataset RCV1. Information retrieval technique such as, PTM i.e Pattern Taxonomy Model is discussed by them. PTM is beneficial data mining technique in text mining domain. PDM i.e. Pattern Deploying Method and PDR i.e. Pattern Deploying with Relevance Function introduced for effective discovery of patterns. In PDM, sequential patterns expanded into feature space, also relations between patterns described as "is-a" relation in PTM. PDR uses relevance functions and utilized a probabilistic method for estimation of weight of term. SPMiner algorithm is proposed to retrieve a set of frequent sequential patterns. To enhance the effectiveness of pattern based method they shown pattern refinement as key improvement [3].

S. Tang Wu et al, extracts descriptive frequent sequential patterns by deducting meaningless ones. They proposed a pattern taxonomy extraction. A pattern based model with frequent sequential pattern instead of keyword based concept is also introduced by them for pattern extraction. The problem of mining sequential patterns from text documents solved with the proposed technique. PTM illustrates relationship between extracted patterns from the collection of text. It also helps to prune meaningless patterns from pattern taxonomy. Basically, PTM is tree like structure. They were applied PTM for user profile filtering task in

which non-relevant document that incoming according to profile of user get filter out. Once topic or the patterns are gained using PTM, centroid i.e. feature vector is used to handle representation of area of topic. In this extracted patterns from training set are represented [4].

S. Shehata make analysis of text features and concept based similarity measures. Traditionally, concept of clustering is used to club similar types of documents into cluster. Documents are splited into different clusters as per similar topics of each document in text clustering approach. Clustering of documents represents the classification and clustering of documents. Text documents term frequency is computed to find important terms. A similarity based concept is implemented to determine similarity among documents. To extracts the concept from document the system scans the documents. Concept-based mining technique is developed using sentence-based concept analysis, document-based concept analysis, the corpus-based concept analysis and the concept-based similarity measure. Using semantic structure of the sentences better quality is achieved in clustering. And the proposed concept based algorithm performance is better than the existing single term based approaches [5].

S. Shehata et al, represented concept-based model. It analyzed the terms on sentences as well as document level than analysis of documents. Proposed model contains concept-based statistical analyzer, conceptual ontological graph representation and concept extractor. Each sentence of the text document is marked automatically with the help of PropBank notations. Both verb and arguments are considered as terms. In sentence, there may have one argument but more than one verb. A concept-based statistical analyzer, representation of COG and concept extractor is also included in proposed model. To maintain sentence semantics concept based statistical analyzer is used [6].

D. Metzler et al proposed a robust query expansion technique. It is used for information retrieval based on Markov random model. The proposed technique is also referred as,"Latent concept expansion". At the time of expansion, it provides the mechanism for term dependencies. They have evaluated technique against the relevance model. For multi term expansion LCE is utilized to perform single or multi-term expansion. The technique produced better attributes, well construct and topically relevant multi-term expansion concepts. Lastly, they discussed that LCE is capable of capturing syntactic dependencies [7].

G. Ifrim accomplish the task of weaken a-priori required knowledge about database as well as tokenization result with the character length. Gradient ascent is used for accomplishment of the task in the space of all 'n'-grams. They discussed about bag of word representations used for categorization of text. A typical type of pre-processing such as, stemming or removal of words is used to provide training to the text. But it required detailed knowledge of text language for categorization. Their contribution is for Structured Logistic Regression i.e. SLR which incorporates the best features of 'n'-grams for variable length. To increase logistic regression likelihood of the training data they have developed a coordinate-wise gradient ascent technique. It inherits the structure of n-gram feature space [8].

C. D. Manning did the study of information retrieval system. IR mainly identifies the usual documents of unstructured nature. IR hides the problems which satisfies the core definition. Generally, unstructured data consists of the data that is not clear but the fact is no data is unstructured. IR covers the supporting users in filtering collection of documents. The proposed task is similar to arrange or sort the books into book-shelf as per topic [9].

R.K.Pon, et al, introduced MTT i.e. multiple topic tracking. It is suitable for news recommendation article for the users having various interests and which is also dynamic over a time. MTT mainly handles the multiple interests' profiles for discovery of interesting articles for individual user given feedback. Performance of iScore is enhances due to MTT. It focuses on new topics and track of all articles as TDT. It significantly earns the better performance. To more specific operating parameters are analyzed for case study. Rocchio algorithm is filtering based approach to represent the topics and text documents as vector. The smaller clusters can also manage by MTT.

R.K.Pon, et al, introduced MTT i.e. multiple topic tracking. It is suitable for news recommendation article for the users having various interests and which is also dynamic over a time. MTT mainly handles the multiple interests' profiles for discovery of interesting articles for individual user given feedback. Performance of iScore is enhances due to MTT. It focuses on new topics and track of all articles as TDT. It significantly earns the better performance. To more specific operating parameters are analyzed for case study. Rocchio algorithm is filtering based approach to represent the topics and text documents as vector. The smaller clusters can also manage by MTT. There analysis suggests that MTT can relatively well suitable for positively tagged articles[10].

S.Zhu et al, discussed about classification problems. In this, classifier has to assign single document to the different category. It is also called as multileveled classification. Many classification methods such as, Naïve Bayes, Logistic regression, SVM etc developed for the problem of single-labelled classification. In multi-labelled classification multiple data categories not be privileged and each data point also belongs to multiple categories simultaneously. A maximum entropy method is proposed by them for multilabelled classification. In this category labels are absolutely considered in the model. The proposed approach is beneficial for correlation classes when there exists strong relationship between classes. The document corpus of Reuters-21578 is used in the phase experimental setup. Ten-fold cross validation used for in all methods for optimal regularization parameters [11].

T. Joachims et al, developed an approach which utilizing click through data for the purpose of training. Support Vector Machine approach is used for retrieval of learning function. The proposed approach can easily grasp the Meta search function for information retrieval. Click through data search is the triplet(q,r,c) of q where, q is the query, r is the ranking function and c is the set of links. In clickthrough data is recorded with fewer overheads. There are strong dependencies between all the parameters included in triplet. Ranking r is depends on query q whereas, c is depends on both r and q [12].

M. J. ZAKI et al, represented SPADE. It is abbreviated as, Sequential Pattern Discovery using Equivalence classes. It is novel algorithm for rapid discovery of Sequential Patterns. It repeatedly scans databases as well as uses the complicated hash functions for flat locality.

Proposed algorithm decomposes the many problems by using combinatorial properties. The task of sequence mining extracts the set of attributes. It is more challenging to discover most frequent sequences from huge databases. There are some key approaches for frequent sequences discovery such as, use of vertical-id list for database format, use lattice-theoretic approach for decomposition of original search space. Decoupled the problem of pattern search by proposing two different strategies such as, frequent sequence enumeration within sub-lattice [13]

Y. Li,et al., discussed about the problem of existing text mining and text classification techniques. All are adopted language-based approaches. They analyze that previous techniques suffered from the problems of polysemy and synonymy. Also they demonstrate that effective tools are required to adequately utilize high scale patterns. They have proposed relevance feature discovery (RFD) to find relevance features present in the text documents. They addressed two problems in text mining such as, low-level support and pattern mining. Continued with RFD model they have implemented WFeatures and FClustering algorithms. FClustering algorithm describes the feature clustering process and discovers the set of patterns whereas; WFeature algorithm is used for computations of weight of classified terms [14].

III. Problem Formulation

A. *"Discovery of relevance feature for mining text"*

There are several existing techniques available for text mining which is based on term based approach such as, PTM, PDR, LCE, PDM etc. But they suffered from some challenging issues such as, low level support, polysemy and synonymy in which same word in different context has multiple meanings, large number of noisy patterns. From literature survey analysis, we analyzed that there is need of such technique which discovers the positive and negative patterns from the text documents.

IV. System Architecture

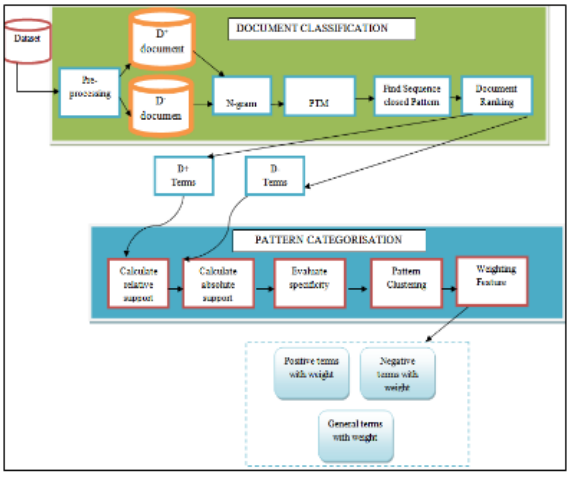

Fig. 1: System Architecture

A. *Dataset Pre-Processing*

For system testing, Reuters-21578 dataset is used. It contains set of text document gathered from Reuters' 1986 newswire. All the documents are assembled and indexed with categories. It is in DTD format. In the pre-processing phase, main categories such as, cricket, football, cooking likewise gets extracted. All these are forwarded to discover relevant features given as below:

1) Relevant feature discovery

From identification of extracted categories, its relevant features are determined. For example, consider categories cricket and basketball for such categories of similar domain relevant words get discovered from dataset.

2) PTM analysis

In proposed system, PTM model is used to determine closed pattern sequence in text paragraphs from relevant document. It is the best solution than the existing pattern based models. For closed sequence pattern discovery, stemmer and stopword algorithms are used to extract important phrases. In stemmer algorithms, variant forms of a word are reduced to a common form and in stopword algorithm words such as, the, is, at, which, and on are reduced.

3) Term frequency identification

In this phase, extracted phrases and words are taken as an input and frequency of each word is evaluated. In term frequency identification, we wish to determine which document is most relevant to the query and simply avoiding documents that do not contain query terms in it. To further distinguish them, we might count the number of times each term occurs in each document; the number of times a term occurs in a document is called its term frequency.

4) N-gram analysis

It is language models. In terms of n-grams, the Trigram model outperforms the Bigram and Unigram models. The performance of the Trigram model is very good and has similar results as PTM.

N-gram also extracts sequential patterns with a specified number of words and with no gaps between the words. It is usually selected based on the sliding window technique and the probability of a n-gram= w_1, w_2, w_3,… ……. W_n, is calculated using the following equation:

$$P(w_1, w_2, w_3,\ldots \ldots. W_n)=P(w_1)P(w_2| w_1,\ldots \ldots. W_{n-1})$$

In proposed work, according dataset analysis, dictionary of relevant words and dictionary of their common features is generated.

As compared to existing system, proposed system performance is better due to n-gram analysis is used along with PTM analysis.

V. Algorithms

A. *FClustering Algorithm*

– Input: Extracted features < T; DP+; DP- > and operation spe.
– Output: 3 sections of terms T+, G and T-.
– Method

1) G =Ø; T+=Ø ; T-=Ø ;
2) foreach ti ε T do
3) if ti ε f ={ t|t ε P; P ε DP+}
4) then T-= T-U{ti};
5) foreach ti ε T –T- do {
6) ci ={ti}
7) maxspe(ci) =minspe(ci)= spe(ti); }
8) let m = |T- T-|;
9) let C = {c1, c2… cm} and minspe(c1)>….> minspe{cm};

10) while (|C| > 3)
11) assume k = 1 and mind =dif(c1; c2);
12) for i = 2 to m-1 do
13) if dif(ci; ci+1) < mind
14) then {k = i; mind =dif(ci, ci+1)}
15) assume ck= ckU ck+1;
16) if minspe(ck+1) < minspe(ck)
17) then minspe(ck) = minspe(ck+1);
18) if maxspe(ck+1) > maxspe(ck)
19) then maxspe(ck)= maxspe(ck+1);
20) for i = k + 1 to m- 1 do // delete ck+1 from C
21) assume ci = ci+1;
22) if |C|= 1 then T+= c1
23) else if |C|= 2 then {T+= c1; G = c2}
24) else {T+= c1; G = c2; T- =T-U c3};

B. *WFeature algorithm*

– Input: A renewed training set, {D+, D-}; Extracted features < T, DP+; DP- >; and the initial term weight function w.
– Output: A term weight function.
– Processing

1) assume n =|D+|;
2) T1= {t|t ε p, p ε D+};
3) foreach t ε T do
4) if t ε T1
5) then sup(t)= d_sup(t, D+);
6) else sup(t)=d_sup(t, D+);
7) foreach t ε T do
8) assume (T+, G, T-= FClustering(T, DP+, DP-, spe());
9) foreach t ε T+ do
10) w(t) = sup(t)*(1 + spe(t));
11) foreach t ε T- do
12) w(t) = sup(t)-|sup(t)* spe(t)|

VI. MATHEMATICAL MODEL

S is the relevance feature discovery system such that
S = {I, F, O}
I is the input to the system
F is system functions
O is Systems output

I: {I1, I2, I3}, Set of input data	– I1= User Login – I2= Upload dataset – I3= Preprocessing request
F: {F1, F2, F3, F4, F5, F6, F7, F8, F9, F10, F11}	– F1= User login – F2=Upload dataset – F3=Request for preprocessing – F4= Apply stemmer algorithm – F5= Apply stopword algorithm – F6= Apply PTM – F7= Used 'n'-gram and tf*tdf technique – F8= Generate sequence pattern – F9=Rank documents – F10=Generate document set i.e. D+ – and D- documents – F11=Calculate relative support
O:{O1 }	O1=Word dictionary

Table 1: Mathematical Model

VII. EXPERIMENTAL SETUP

A. *Experimental Setup*

Desktop based system is generated using jdk 1.7. Mysql is used to store database. System is implemented and tested on windows platform with i5 processor and 8 gb ram.

B. *Dataset*

Reuters-21578 corpus [15]: It is gathered and labeled by Carnegie Group, Inc. and Reuters, Ltd. It is very large dataset. It contains set of text document gathered from Reuters' 1986 newswire. All the documents are assembled and indexed with categories.

The dataset is in DTD format. The format of dataset contains various tags such as, NEWID, DATE, TOPICS, TITLE, DATELINE, BODY, etc.

The system is tested on various topics with different file sizes. As dataset size increases more words are extracted and better precision is achieved.

Following graph shows the number of terms extracted based on various data sizes for PTM technique.

Avg. Training Documents	T+	T-	G
40	685	395	99
60	430	410	119
80	392	449	102
100	372	314	174

Table 1: Dataset Word Extraction

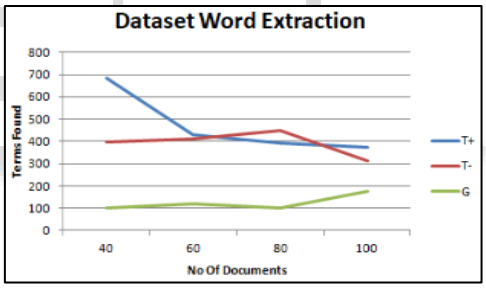

Fig. 1: Graph of word extraction

T+, T- and G terms are extracted using PTM and N-Gram hybrid technique. Following graph represents the number of terms extracted using PTM and NGram hybrid technique. More terms are extracted using PTM and NGram hybrid technique.

Avg. Training Documents	T+	T-	G
40	946	1143	215
60	1120	863	165
80	645	754	141
100	980	750	195

Table 2: 'N'-Gram Evaluation

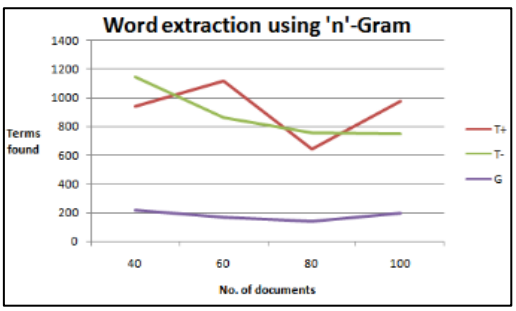

Fig. 2: Graph of N-gram evaluation

Figure 2 represents the N-gram evaluation for given number of documents such as, 40, 60, 80 and 100. With the N-gram technique more précised results are achieved.

In given graph, X-axis represents number of documents and Y-axis represents terms found such as, T+, T-, G.

Avg. Training Documents	T+	T-	G
40	1024	2560	198
60	930	970	119
80	743	745	102
100	834	721	174

Table 3: 'N'-Gram + Ptm Hybrid Evaluation

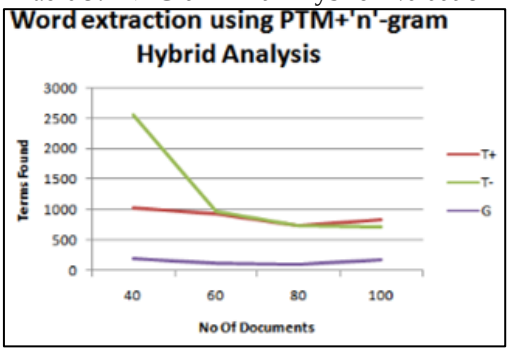

Fig. 3: Graph of PTM + N-gram hybrid evaluation

After matching the score of term and its occurrence in extracted dataset relevant and irrelevant terms are classified. Based on this classification result precision is calculated. Following graph shows the precision evaluation for topic dataset. More precision is achieved using PTM and N-Gram hybrid technique.

Avg. Training Documents	PTM T+	PTM T-	PTM G	PTM & Ngram T+	PTM & Ngram T-	PTM & Ngram G
40	0.62	0.74	0.8	0.69	0.81	0.84
60	0.73	0.81	0.92	0.81	0.86	0.95
80	0.88	0.87	0.95	0.93	0.92	0.96
100	0.89	0.91	0.94	0.95	0.94	0.97

Table 4: Avg. Precision

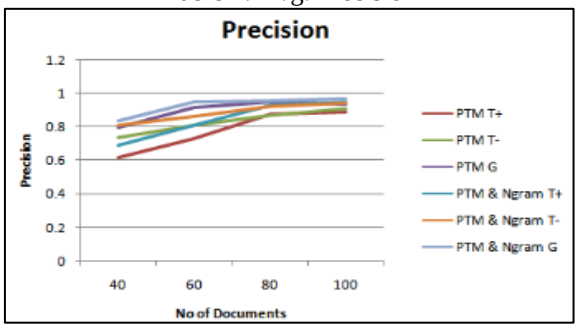

Fig. 4: Graph of avg. precision

Avg. Training Documents	Precision	Recall	FB	MAP
40	0.72	0.86	0.88	0.79
60	0.82	0.76	0.79	0.79
80	0.9	0.86	0.86	0.88
100	0.9	0.8	0.88	0.85

Table 5: 'P' Value Measure In Existing System

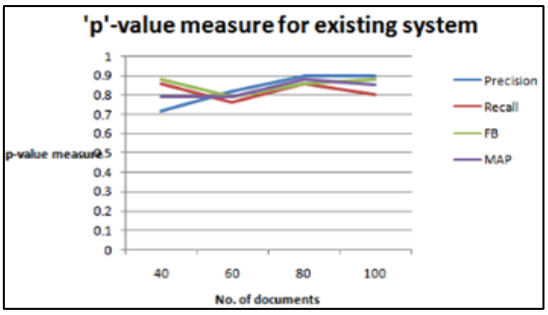

Fig. 5: Existing system 'p'-value measure

In table 5, existing system 'p'-value measure is shown. In this we have calculated, Fβ-measure value, MAP and b/p. The definitions of these parameters are as follow:

F-beta measure is the function which describes recall and precision of system together. Beta is the parameter which denotes the precision and recall equally weighed.

F_β is denoted by $F_1=(2PR)/(P+R)$.

- MAP: Mean Average Precision: This combines the average precision of each topic. The average precision of each topic is evaluated based on the precision of each relevant document.
- b/p: break-even point: This is the p/r –precision recall curve. Larger intersection point value shows the better performance. In case of existing system intersection point is: 0.74

Avg. Training Documents	Precision	Recall	FB	MAP
40	0.77	0.81	0.78	0.8
60	0.81	0.83	0.81	0.83
80	0.84	0.82	0.82	0.85
100	0.92	0.82	0.86	0.88

Table 6:'P' Value Measure in Proposed System

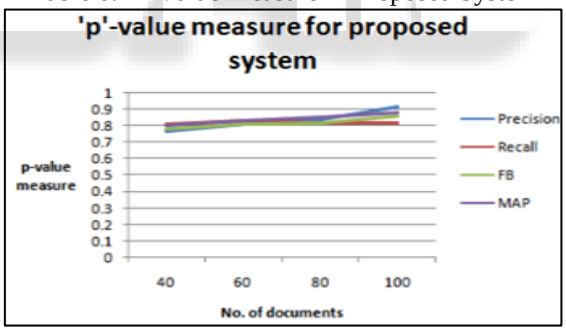

Fig. 6: Proposed system 'p'-value measure

In table 6, 'p'-value measures for proposed system is given. As per observations, proposed system is more précised due to extraction of more results and their accuracy.In case of proposed system, intersection point i.e. b/p is: 0.8. The value of interesection is greater as compared to existing system due efficient precision value of proposed system.

VIII. Conclusion

In proposed system RFD model is implemented to discover the positive, negative patterns from input text dataset. There two algorithms utilized in for feature clustering such as, Fclustering and Wfeature algorithms. Ngram & PTM hybrid pattern extraction technique extracts more terms and achieves more precision than PTM technique. From result table and analysis, it is seems that with the proposed approach more précised as well as accurate results are generated due to

hybrid 'n'-gram+PTM analysis. Previously, 0.74 average precision is outputted due to less results & less accuracy. But in case of proposed system it is 0.8 which is more due to more efficiency and accuracy in system results. With 'n'-gram & PTM analysis more terms and achieve more precision than PTM technique.

REFERENCES

[1] F. Song and W. B. Croft, "A general language model for information retrieval," in Proc. ACM Conf. Inf. Knowl. Manage., 1999, pp. 316321
[2] F. Sebastiani, "Machine learning in automated text categorization," ACM Comput. Surveys, vol. 34, no. 1, pp. 147, 2002
[3] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic pattern taxonomy extraction for web mining," in Proc. Int. Conf. Web Intell., 2004, pp. 242248.
[4] S.-T. Wu, Y. Li, and Y. Xu, "Deploying approaches for pattern refinement in text mining," in Proc. IEEE Conf. Data Mining, 2006, pp. 11571161
[5] S. Shehata, F. Karray, and M. Kamel, "Enhancing text clustering using concept-based mining model," in Proc. 2nd IEEE Conf. Data Mining, 2006, pp. 10431048
[6] S. Shehata, F. Karray, and M. Kamel, "A concept-based model for enhancing text categorization," in Proc. ACM SIGKDD Knowl. Discovery Data Mining, 2007, pp. 629637.
[7] D. Metzler and W. B. Croft, "Latent concept expansion using Markov random fields," in Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2007, pp. 311318.
[8] G. Ifrim, G. Bakir, and G. Weikum, "Fast logistic regression for text categorization with variable-length n-grams," in Proc. ACM SIGKDD Knowl. Discovery Data Mining, 2008, pp. 354362.
[9] C. D. Manning, P. Raghavan, and H. Schtze. "Introduction to Information Retrieval". Cambridge, U.K.: Cambridge Univ. Press, 2009.
[10] R. K. Pon, A. F. Cardenas, D. Buttler, and T. Critchlow, "Tracking multiple topics for finding interesting articles," in Proc. ACM SIGKDD Knowl. Discovery Data Mining, 2007, pp. 560569
[11] S. Zhu, X. Ji, W. Xu, and Y. Gong, "Multi-labelled classification using maximum entropy method," in Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2005, pp. 10411048.
[12] T. Joachims, "Optimizing search engines using clickthrough data," in Proc. ACM SIGKDD Knowl. Discovery Data Mining, 2002, pp. 133142.
[13] M. J. Zaki, "Spade: An efficient algorithm for mining frequent sequences," in Mach. Learn. J. Spec. Issue Unsupervised Learn., vol. 42, pp. 3160, 2001.
[14] Y. Li, A. Algarni, M. Albathan, Y. Shen, and M.A. Bijaksana, "Relevance Feature Discovery for Text Mining," in IEEE Trans. Knowl. Data Eng., vol. 26, no. 6, pp., Jan. 2015.
[15] https://datahub.io/dataset/reuters-21578/resource/a80d95de-4b9f-4bcda78b-2f1ee844de26

