ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 3, pp: (170-176), Month: July - September 2016, Available at: www.researchpublish.com

SEMANTIC BASED PATTERN TOPIC MODEL

Tincy Chinnu Varghese

M.G University, Mount Zion College of Engineering, Pathanamthitta, India

Abstract: **Today in the field of information filtering topic modeling is widely used. Topic modeling helps to generate models which can discover the hidden topics in a document collection and each of these topics are represented by word distribution. There are term-based and pattern-based approaches in information filtering. Patterns are more discriminative than single words. In many pattern-based methods the patterns in the documents are considered. But the pattern which occurs multiple times in a document which has to be filtered is also given equal importance. Another problem which the existing pattern-based methods face is that the semantics of a term in a pattern is not considered. Another limitation is that the distribution of a pattern in a document is not given any importance. To overcome the above mentioned problems a new ranking method which considers the number of frequency of the patterns, distribution of patterns and semantic based pattern representation to estimate the relevance of the documents based on the user information needs is introduced. This helps to remove the irrelevant documents effectively. TREC data collection Reuters Corpus Volume 1 is used to do the extensive experiments to evaluate the effectiveness of the proposed method. The result says that the proposed model works better than the existing pattern-based topic for document modeling in information filtering.**

Keywords: **Topic Model, Information Filtering, Pattern mining, relevance ranking, user interest model.**

I. INTRODUCTION

In Today's world Information Filtering (IF) is having a major role in many areas of our daily life .Especially in the field of machine learning, web context, thematic issues as varied as voice recognition, content-based feature extraction, information retrieval and recommendations etc. Information Filtering is a method of removing redundant or unwanted information from a document using any (semi) automated or computerized methods .Traditional models for information filtering (IF) are word-based or term-based approaches. The advantage of the term-based approach is its efficient computational performance. But the disadvantage of term-based document representation is the problem of polysemy and synonymy. To improve the limitations of term-based approaches, pattern-based mining techniques have been used. By using patterns it helps to utilize patterns to represent users' interest because patterns carry more semantic meaning than terms. Several data mining techniques have been developed to improve the quality of patterns such as maximal patterns, closed patterns for removing irrelevant documents from the collection.

These data mining and text mining techniques assumes that the user's interest is related to single topic. But, in reality it is not the real case. For example, when a user asks for an article 'car', it can be related to its price, policy, market and so on, which means the user's interest can be diverse. Therefore, a model for user's interest in multiple topics other than a single topic has been introduced. Topic modeling [3][4]is one of the most popular text modeling techniques which is used now-a-days for information filtering. This technique can divide documents in a collection on the basis of the number of topics and can represents every document with multiple topics and underlying word distribution. Traditional IF models used term-based or word-based way to perform topic modeling. But it faces the problems such as polysemy and synonym.

To come across the above mentioned limitations and problems, a ranking method was introduced.The features of the proposed model include:

(1) User's needs for information is in terms of multiple topics.

(2) Each topic is shown by patterns.

(3) Patterns are generated from topic models based on the uniform distribution of patterns in the document and are arranged in terms of their ranking.

(4) The Maximum Matched Patterns along with number of frequency of patterns present in a document, helps to estimate the relevant document to the user's needs for information to filter out irrelevant documents.

(5)The open Natural Language Processing (NLP) tool also helps to take the semantic meaning of a pattern in a document.

II. EXISTING SYSTEM

Traditional IF models were done using the term-based or word-based topic models. Its computational efficiency made these models popular. But it has the polysemy and synonyms problems. To overcome this problem pattern-based topic modeling was proposed. Here patterns are considered for filtering an incoming documents to get relevant document from a collection. But these patterns cannot be applied direct for information filtering. Information filtering system carries user's interest or user information needs on the basis of their 'user profiles'. Information filtering systems gives users the information that is needed to them [6].The main goal of information filtering is the ranking of the incoming documents on the basis of its relevance. Let D be the collection of incoming documents ,then the process in information filtering is mapping the Rank(D):D-->R where rank(D) represents the relevance of the document D. Document filtering is considered as a ranking process of a document. There are several methods to model the relevance of the documents which includes term-based model [2], pattern-based model [7] [8], a probabilistic model[12].

In most popular term-based models have the problem of polysemy and synonymy and also have the limitation to express semantics. For that more semantic features like phrases and patterns should be extracted to represent the documents . But the phrase-based approaches suffers from the problem of low frequency Pattern based approach is more effective compared to other approaches [7][8]. Pattern mining is a major topic in datamining which is widely in use over years. Effective algorithms like Apriori, FP-tree etc are used to extract the frequent patterns. As the number of these frequent patterns will be huge to process sometimes For that more efficient or relevant patterns such as closed patterns ,max pattern etc are discovered. Closed patterns [10] is a condensed representation of the frequent patterns.

Topic models techniques have been included in the language model and got successful retrieval results[3], [11], that has opened up a new way for modeling the relevance of a document. LDA based document models are the main topic modeling methods. This model achieved good efficiency when compared to other models . As this was achieved by [11],due the multiple topic document model, and also because of that each topic in a topic model was represented by a collection of semantically similar words, that solves the synonymy and polysemy problem of term based document models.

III. PROPOSED METHOD

The proposed model consists of two stages: 1) A training part which generate user's interest model from a collection of training documents (user interest modeling) and 2) A filtering part that gives the relevance of the new incoming documents based on user interest model that are generated during training phase (document relevance ranking). The contributions of the proposed model to the field of IF are as follows:

1) Proposed to model users' interest on multiple topics rather than a single topic under the thought that users' information interests is diverse.

2) Integrates data mining techniques along with statistical topic modeling techniques to generate a pattern-based topic model to represent document collections. The proposed model consists of topic distributions describing topic importance of each document and pattern-based topic representations representing the semantic meaning of each topic.

3) A structured pattern-based topic representation where patterns are arranged into groups, called equivalence classes, based upon their taxonomic and statistical features. Patterns in an equivalence class is having the same frequency and represent similar semantic meanings. With this feature, the most representative patterns can be identified which will be useful for filtering relevant documents.

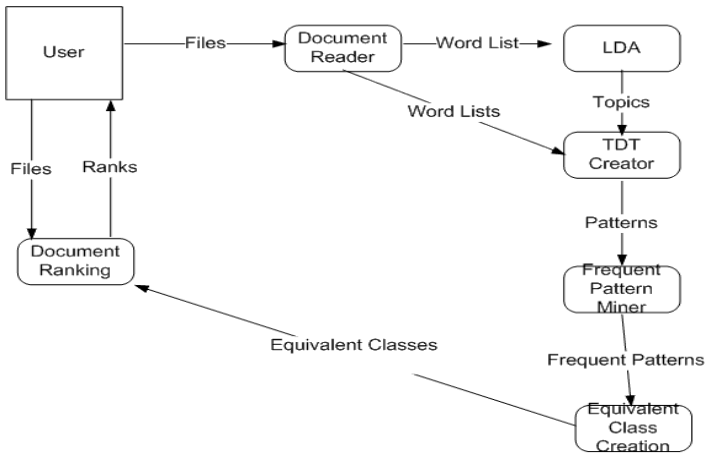
ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 3, pp: (170-176), Month: July - September 2016, Available at: www.researchpublish.com

4) A new ranking method is used to determine the relevance of new documents based on the proposed model .The Maximum matched patterns, which are the largest patterns in each equivalence class exist in the incoming documents are used to calculate the relevance of the incoming documents to the user's interest.

(5) Along with the maximum matched pattern it also calculate the frequency of the number of times the maximum matched patterns occurred in a document and also the order in which the patterns are arranged in a document to rank documents.

(6)Open NLP (Natural Language Processing) tool helps to avoid synonyms and polysemy as it takes the semantic

meanings of the patterns.The Architectural design of the proposed model is given below:



Patterns are more accurate and meaningful than words to represent topics .Also pattern-based representations gives more structural information which tells the association between words. There are mainly four steps to generate the Topic based user interest model . First two steps helps in discovering semantically meaningful

pattern to represent topics and documents,

1) Constructs a new transactional dataset(TDT) from the results of LDA model.

2)Generates pattern-based representations from the TDT to represent user needs. As Pattern-based topic representations will not be sufficient to be directly used to know the relevance of new documents to the user interests. For this the document relevance ranking is done based on the Maximum Matched Patterns.

3) Pattern equivalence classes are constructed.

4) Generates the user interest model.

3.1 Latent Dirichlet Allocation (LDA):

LDA helps to discover the topics that are hidden in the documents. In LDA, every document is viewed as a combination of various topics. LDA represents topics using word distribution and documents using topic distribution. Topic representation tells that which words or terms are important to a topic and document representation tells which topics are important to a document. LDA is a commonly used topic modeling tool.

Example result of LDA is shown in Table 1.

Table 1: Example for LDA: Word-Topic Assignments

Topic / Document	T_1 $\vartheta_{D,1}$	Words	T_2 $\vartheta_{D,2}$	Words	T_3 $\vartheta_{D,3}$	Words
D_1	0.6	w_1, w_2, w_3, w_2, w_1	0.2	$w_1, w_9, w_8,$	0.2	w_7, w_{10}, w_{10}
D_2	0.2	$w_2, w_4, w_4,$	0.5	w_7, w_8, w_1, w_8, w_8	0.3	w_1, w_{11}, w_{12}
D_3	0.3	$w_2, , w_1, w_7, w_5$	0.3	$w_7, , w_3, w_3, w_2$	0.4	w_4, w_7, w_{10}, w_{11}
D_4	0.3	$w_2, w_7, w_6,$	0.4	$w_9, w_8, w_1,$	0.3	w_1, w_{11}, w_{10}

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 3, pp: (170-176), Month: July - September 2016, Available at: www.researchpublish.com

3.1.1 Construction of a Transactional Dataset:

Let $R_{Di,Tj}$ represent the word-topic assignment to topic *Tj* in document D_i. $R_{Di,Tj}$ is a set of words assigned to topic *Tj*. For the example in Table1.1,for topic T1 in document D_1,$R_{Di,Tj} = \langle w_1, w_2, w_3, w_2, w_1 \rangle$. By constructing a set of words from each word-topic assignment $R_{Di,Tj}$ instead of using the sequence of words in $R_{Di,Tj}$,because for pattern mining, the frequency of a word within a transaction is insignificant. Let I_{ij} a set of words which occur in $R_{Di,Tj}$, $I_{ij} = \{w | w \in R_{Di,Tj}\}$, i.e. I_{ij} contains the words which are in document D_i and assigned to topic *Tj* by LDA., I_{ij} called a *topical document transaction*, which is a set of words without any duplicates. From all the word-topic assignments $R_{Di,Tj}$ to *Tj*, i= 1,…..M, we can construct a transactional dataset TD_j. Let D=$\{D_1, \ldots \ldots D_M\}$ be the original document collection, the transactional dataset TD_j for topic T_j is defined as $TD_j = \{I_{1j}, I_{2j}, \ldots I_{Mj}\}$. For the topics in D, we can construct V transactional datasets (TD_1,TD_2. .. TD_V). An example of transactional datasets is given in Table 2, which is generated from the example in Table 1

Table 2: Transactional Datasets Generated from Table 1 (Topical Document Transaction (TDT))

T	TDT	TDT	TDT
1	$\{w_1, w_2, w_3\}$	$\{w_1, w_8, w_9\}$	$\{w_7, w_{10}\}$
2	$\{w_2, w_4\}$	$\{w_1, w_7, w_8\}$	$\{w_1, w_{11} w_{12}\}$
3	$\{w_1, w_2, w_5, w_7\}$	$\{w_2, w_3, w_7\}$	$\{w_4, w_7, w_{10}, w_{11}\}$
4	$\{w_2, w_6, w_7\}$	$\{w_1, w_8, w_9\}$	$\{w_1, w_{11}, w_{10}\}$
	$\boldsymbol{TD_1}$	$\boldsymbol{TD_2}$	$\boldsymbol{TD_3}$

3.1.3 Generating Pattern Enhanced Representation:

The goal of the proposed pattern-based method is to use frequent patterns that are generated from each transactional dataset TD_j to represent T_j. In this four-stage topic model, frequent patterns are generated in this step. For a given minimal support threshold σ , an itemset X in TD_j is frequent if supp(X) >= σ, where supp(X) is the support of X which is the number of transactions in TD_j that contain X. The frequency of the itemset X is defined as $\frac{\boldsymbol{supp(X)}}{|\boldsymbol{TD_j}|}$ .Topic T_j can be represented by a set of all frequent patterns, denoted $\boldsymbol{X_{z_i}} = \{\boldsymbol{X_{i1}}, \boldsymbol{X_{i2}}, \ldots \ldots \boldsymbol{X_{i_{mi}}}\}$, where m_i is the total number of patterns in X_{Zi} and V is the total number of topics. Take TD_2 in Table 2 as an example, which is the transactional dataset for T_2. For a minimal support threshold σ = 2, all frequent pattern s generated from TD_2 are given in Table 3

Table 3 Frequent Patterns for T_2, σ = 2

Patterns	*Supp*
$\{w_1\}, \{w_8\}, \{w_1, w_8\}$	3
$\{w_9\}, \{w_7\}, \{w_8, w_9\}, \{w_1, w_9\}, \{w_1, w_8, w_9\}$	2

3.1.2 Construction of Equivalence class of pattern:

Usually, the number of frequent patterns may be large and most of them are not properly useful. Several concise patterns are used to represent useful patterns generated from a large dataset instead of frequent patterns such as maximal patterns and closed patterns. The number of these concise patterns is significantly smaller than the number of frequent patterns for a dataset. In particular, the closed pattern has drawn great attention due to its attractive features.

Equivalence Class For a transactional dataset TD, let X be a closed itemset and G(X) consist of all generators of X, then the equivalence class of X in TD, denoted as EC(X), is defined as EC(X) = G(X) U {X}.

Let EC_1 and EC_2 be two different equivalence classes of the same transactional dataset. Then $EC_1 \cap EC_2 = \emptyset$, which means that the equivalence classes are not shared with each other. All the patterns in an equivalence class have the same frequency. The frequency of a pattern indicates the statistical significance of the pattern. The frequency of the patterns in an equivalence class is used to represent the statistical significance of the equivalence class. Table 4 shows the three equivalence classes within the patterns for topic T_2 in Table .3, where f indicates the statistical significance of each class.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 3, pp: (170-176), Month: July - September 2016, Available at: www.researchpublish.com

Table 4: Equivalence Classes in TD_2

$EC_{21}(f_{21} = 0.75)$	$EC_{22}(f_{22} = 0.5)$	$EC_{23}(f_{23} = 0.5)$
$\{w_1, w_8\}$	$\{w_1, w_8, w_9\}$	$\{w_7\}$
$\{w_1\}$	$\{w_1, w_9\}$	
$\{, w_8\}$	$\{w_8, w_9\}$	
	$\{w_9\}$	

Assume that there are n_i frequent closed patterns in X_{Z_i} , which are c_{i1} ,……..c_{in_i},and that X_{Z_i} can be partitioned into n_i equivalence classes, $EC(c_{i1}), \ldots\ldots EC(c_{in_i})$. For simplicity, the equivalence classes are denoted as for $EC(c_{i1}), \ldots\ldots EC(c_{in_i})$ for X_{Z_i} or simply topic T_i .Let $E(T_i)$ denote the set of equivalence classes for topic T_i , i.e. $E(T_i) = \{EC_{i1,\ldots\ldots} EC_{in_i}\}$. In this model, the equivalence classes $E(T_i)$ are used to represent user interests which are denoted as $U_E = \{E(T_1), \ldots\ldots E(T_V)\}$.

3.1.3 Algorithm for User Profiling:

Input: a collection of positive training documents *D*; minimum support σ_j as threshold for topic T_j; number of topics *V*

Output: $U_E = \{E(T_1), \ldots\ldots E(T_V)\}$

1: Generate topic representation ∅ and word-topic assignment $T_{D,i}$ by applying LDA to *D*

2: $U_E := \emptyset$;

3: **for** each topic $T_j \in [T_1, T_V]$ **do**

4: Construct transactional dataset $\mathbf{TD}_j$ based on ∅ and $Z_{d,i}$

5: Construct user interest model $\mathbf{X}_{Z_j}$ for topic T_j using a pattern mining technique so that for each pattern *X* in $\mathbf{X}_{T_j}$,$supp(X) > \sigma_j$

6: Construct equivalence class $E(Z_j)$ from $\mathbf{X}_{T_j}$

7: $U_E := U_E \cup \{E(T_j)\}$

8: **end for**

3.2 Document Relevance Ranking based on Topics:

At filtering stage, document relevance is to filter out irrelevant documents based on the user's information needs. Identify maximum patterns in d which match some patterns in the topic-based user interest model. Then estimate the relevance of d based on the user's topic interest distributions and the significance of the matched patterns. For topic significance, let d be a document,Z_j be a topic in the user interest model .let PA_{jk}^dbe a set of matched patterns in document d for topic Z_j.Then the corresponding topic significance of Z_j can be defined as

$$\boldsymbol{sig(Z_d, d) = \sum_{k=1}^{n_j} spec(PA_{jk}^d) \times f_{jk} = \sum_{k=1}^{n_j} a\,|PA_{jk}^d|^m f_{jk}} \qquad \textbf{(1)}$$

For the incoming documents d ,we propose to estimate the relevance of d to the user interest based on the topic significance and topic distribution. The equation is as follows

$$\textbf{Rank(d)} = \sum_{j=1}^{V} \boldsymbol{sig(Z_j, d) \times \vartheta_{DJ}} \qquad \textbf{(2)}$$

By equating both these equations we get a equation which is denoted as $Rank_E(d)$ is given below

$$\boldsymbol{Rank_E(d) = \sum_{j=1}^{V} \sum_{k=1}^{n_j} |MC_{jk}^d|^{0.5} \times \delta(MC_{jk}^d, d) \times f_{jk} \times \vartheta_{D,j}} \qquad \textbf{(3)}$$

The higher the $Rank_E(d)$,more relevant is the document for the users.

3.2.1 Relevance based on Pattern semantics:

Instead of simply matching the maximum matched patterns within the incoming documents the relationships between the words are also considered. For eg: consider the pattern {pattern, topic mining} it will take all the documents including the semantics of these words.

3.2.2 Relevance based on Uniform distribution of patterns:

The patterns in an incoming document is very important. As the document are matched with the equivalent classes and given relevance according to the best matched documents .If the particular pattern which we are considering is in the title, first paragraph, conclusion the pattern will be given a special weightage during testing. If the pattern is present in a document as uniformly distributed that document will be more relevant from the collection of documents.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 3, pp: (170-176), Month: July - September 2016, Available at: www.researchpublish.com

3.2.3 Relevance based on frequency of the number of patterns:

Maximum Matched patterns in the collection of documents are considered. By Calculating the frequency of the no of times the patterns have occurred in the documents the ranking is done. The higher the number of times the pattern occurred in the document the more it is ranked higher. Frequency calculation is done along with the maximum matched pattern documents.

$$rank(d) = \sum_{j=1}^{v} \sum_{k=1}^{n_j} \left| MC_{jk}^{d\ 0.5} \right| \times \ \boldsymbol{\delta}(\boldsymbol{MC_{jk}^{d}, d}) \times \boldsymbol{f_{jk}} \times \boldsymbol{\vartheta_{D,j}} \times 0.6 + 0.2 \times uniform\ distribution + 0.2 \times ECFrequency[d]$$

3.2.4 Algorithm for Document Filtering:

Input: user interest model $U_E = \{E(T_1), \ldots \ldots E(T_V)\}$, a list of incoming document D_{in}

Output: $rank_E(d), d \in D_{in}$

1: *rank(d)*:=0;

2: **for** each d ∈ D_{in} **do**

3: **for** each topic $T_j \in [T_1, T_V]$ **do**

4: **for** each equivalence class $EC_{jk} \in E(T_j)$ **do**

5: Scan $EC_{k,j}$ and find maximum matched pattern MC_{jk}^{d} which exists in *d*

6: update $rank_E(d)$ using Equation (3):

7:$rank(d) = rank(d) + |MC_{jk}^{d}|^{0.5} \times f_{jk} \times \vartheta_{D,j} \times 0.6 + 0.2 \times uniform\ distribution + 0.2 \times ECFrequency[d]$.

8: **end for**

9: **end for**

10: **end for**

IV. EXPERIMENTAL RESULTS

The experimental results are based on the uniform distribution of patterns, based on the weight and frequency of patterns in a document. The results shows the efficiency of the new ranking method .Experimental results shows that new ranking method is more effective than the old ranking method. Fig 4.1 shows a graph which shows that the new relevance ranking shows more weight than the existing pattern model.

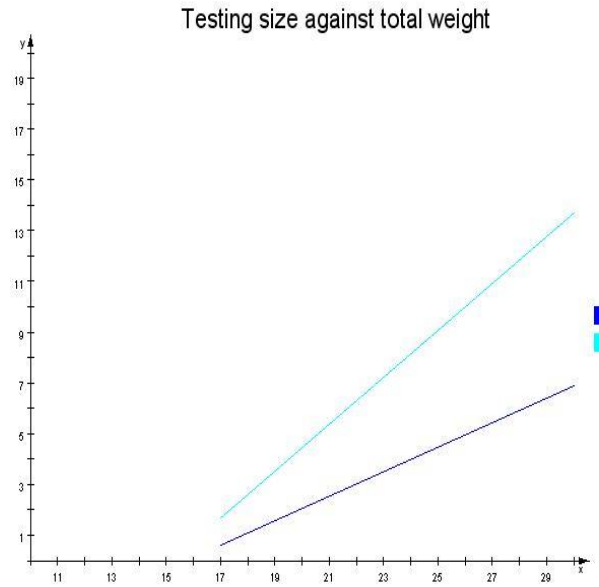


Fig.4.1 Testing size against total weight

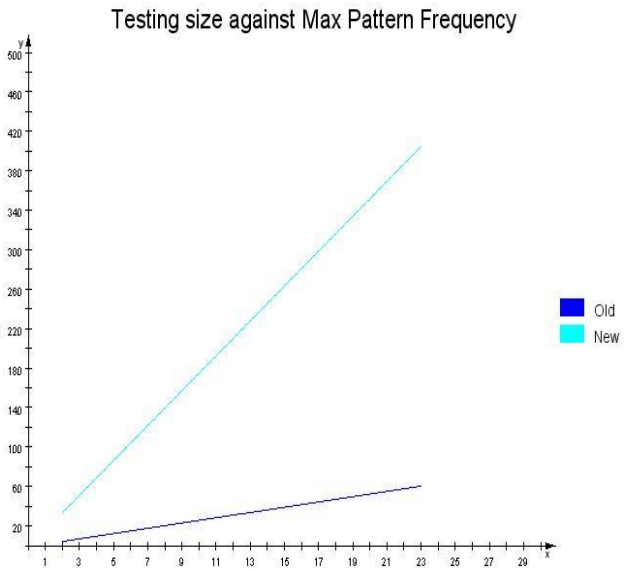


Fig 4.2 Testing size against maximum matched patterns

V. CONCLUSION

An pattern enriched topic model for information filtering including user interest modeling and document relevance ranking. The proposed model generates pattern enhanced topic representations to model user's interests across multiple topics. In the testing phase, the model selects maximum matched patterns, instead of using all discovered patterns, for estimating the relevance of incoming documents. It also considers the uniform distribution of patterns across the document, considers the semantic meaning of the pattern and also the number of frequency of patterns.It automatically generates descriptive and semantic rich representations for modeling topics by the combination of statistical topic modeling techniques with data mining techniques. The technique is used for, content-based feature extraction and modeling tasks, such as information retrieval and recommendations, machine learning and text mining.

REFERENCES

[1] S.Robertson, H.Zaragoza, and M.Taylor, "Simple BM25 extension to multiple weighted fields," in Proc. 13th ACM Int. Conf. Inform.Knowl.Manag.. 2004, pp.42-49.

[2] N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for text mining," IEEE Trans. Knowl. Data Eng., vol. 24, no. 1, pp. 30–44, Jan. 2012.

[3] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: Current status and future directions," Data Min. Knowl. Discov., vol. 15, no. 1, pp. 55–86, 2007.

[4] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2011, pp. 448–456.

[5] Y. Gao, Y. Xu, Y. Li, and B. Liu, "A two-stage approach for generating topic models," in Advances in Knowledge Discovery and Data Mining, PADKDD'13. New York, NY, USA: Springer, 2013, pp 221–232.

[6] H. S. Christopher, D. Manning, and P. Raghavan, An Introduction to Information Retrieval. Cambridge, United Kingdom.: Cambridge Univ. Press, 2009.

[7] C. Zhai, "Statistical language models for information retrieval," Synthesis Lectures Human Lang. Technol., vol. 1, no. 1, pp. 1–141, 2008

[8] A.Tagarelli and G.Karypis,"A segment-based approach to clustering multi-topics documents",Knowl.inform. Syst.,vol.34,no.3,pp.563-595,2013

[9] Haarslev, V., Lutz, C., MÄoller, R., Foundations of Spatioterminological Reasoning with Description Logics. In: Principles of Knowledge Representation and Reasoning: Proc. of the Sixth International Conference (KR'98),1998.

[10] Vishal Gupta , Gurpreet S. Lehal "A Survey of Text Mining Techniques and Applications" Journal of Emerging Technologies in Web Intelligence, VOL 1, No.1, August 2009.

[11] M. Hassel, "Exploitation of Named Entities in Automatic Text Summarization for Swedish," In the Proceedings of NODALIDA '03 - 14th Nordic Conference on Computational Linguistics, May 30-31 2003,Reykjavik, Iceland.

[12] Sang-Bum kim, Kyong-soo Han, Hae-Chang Rim, Sung Hyon Myaeng "Some Effective techniques for Naïve Bayes Text Classification" IEEE Transactions on Knowledge and Data Engineering -2006

[13] S. Alice and S. Conrad. "Page segmentation by web content clustering", In Proceedings of the International Conference on Web Intelligence, Mining and Semantics, WIMS '11, pp.1-9, New York, NY, USA, 2011.

[14] A. Zhang, J. Jing, L. Kang, and L. Zhang. "Precise web page segmentation based on semantic block headers detection", 6th International Conference on Digital Content, Multimedia Technology and its Applications (IDC), pp. 63-68, 2010.