ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Pattern Discovery in Text Mining Using Text Patterns and Clustering

Manjiri M. More [1], Prof. Archana S. Vaidya[2]

PG Student, Department of Computer Engineering, Gokhale Education Society's R. H. Sapat College of Engineering, Nashik, India[1]

PG Coordinator, Department of Computer Engineering, Gokhale Education Society's R. H. Sapat College of Engineering, Nashik, India[2]

ABSTRACT: Topic modeling has been generally acknowledged in the regions of machine learning and data mining, etc. It was proposed to produce statistical models to gather numerous topics in a collection of documents. A principal supposition for this strategy is that the documents in the collection are about one topic. Topic modeling, for example, LDA (Latent Dirichlet Allocation), was proposed to make statistical model to show multiple topics in a collection of documents, and this has been broadly utilized as a part of the fields of information retrieval. However, its adequacy in information filtering has not been very much assessed. Patterns are always thought to be more discriminative than single terms for describing documents. Determination of the most illustrative and discriminative patterns from the huge amount of discovered patterns becomes critical. To manage the limitation and issues, a novel information filtering model is proposed. Proposed model incorporates user information needs that are produced as far as different topics where every topic is appeared by patterns. Patterns are produced from topic models and are sorted out as far as their statistical and taxonomic feature and the most discriminative and representative patterns are proposed to the document relevance to the user's information needs to filter out unessential documents. Experiments are conducted to find effectiveness of Maximum matched Pattern-based Topic model with Dimensionality Reduction (MPBTM-DR) by using TREC data collection Routers Corpus Volume 1.The results shows that MPBTM-DR model significantly outperforms term-based model.

KEYWORDS: Topic modeling, Text mining, Document relevance, Information filtering, Information retrieval. Latent dirichlet allocation

I.INTRODUCTION

Topic modeling is a type of data mining, a method for discovering patterns in a corpus. Numerous data mining techniques have been proposed for mining useful patterns in text documents. Pattern mining based methods have been utilized to use patterns to represent user's interest and have accomplished a few enhancements in effectiveness since patterns convey more semantic significance than terms. To create user interest document representation by utilizing information filtering i.e. remove or erase undesirable document and make user interested document is principle point of information filtering. Numerous information filtering model are term based [1], [2] and pattern based [3], [4]. In term based information filtering term weighting is done yet it experiences issue of polysemy and synonymy. To overcome the weak point of term-based methodologies, pattern mining based technique have been utilized to make utilization of patterns to speak about users interest and have gotten a few changes in effectiveness since pattern convey more semantic importance than terms. These whole models expected that user just concerned in single topic in the field on text mining and information filtering for instance one news article discussing an "Apple" in that it might be connected numerous topics that are price, fruit, smart phone etc. Once in a while new document might land at time and the user interest might change. So here we are concentrating on user interest model on numerous topics not single topic. There are diverse strategies utilized for topic representation that are PLSA [5] (Probabilistic Latent Semantic Analysis) and LDA [6], [7], [8] (Latent Dirichlet Allocation). Which generate multiple topics and distribution of topics. However, there are two issues if we specifically utilize topic models for information filtering. The one issue is that the topic distribution itself is inadequate to speak about documents because of its limited number of dimensions. The second

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

issue is that the word based topic representation is confined to distinctively speak about documents which have diverse semantic content since numerous words in the topic representation are duplicate general words.

In this paper, we propose to choose the most discriminative and representative patterns, which are known as Maximum matched patterns, to speak about topics instead of utilizing frequent patterns. A new topic model, called MPBTM-DR (Maximum matched Pattern based Topic Model with Dimensionality Reduction) is proposed for document representation and relevance ranking. The patterns in the MPBTM-DR are all around organized with the goal that maximum matched patterns are efficiently and effectively used to represent and rank documents.

The association of this paper is as per the following. Section 2 audits the literature survey. In section 3, we present details of proposed work. Section 4 presents mathematical model. Section 5 contains experimental setup and results. Finally paper concludes in section 6.

II. LITERATURE SURVEY

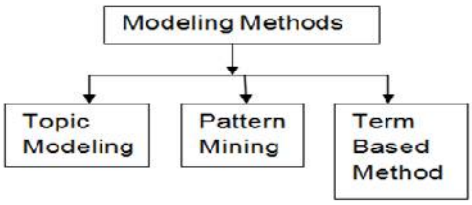

Fig 1: Three technical categories of modeling.

H.Zaragoza, S.Robertson, and M.Taylor in [1] demonstrates that term based methodology gives efficient computational execution, and also theories for term weighting. But it makes issue of polysemy and synonymy.

M. Ester,F. Beil,and X. Xu in [2] has proposed a pattern mining based procedures which is utilized to make practical and effective utilization of patterns to represent users interest and effectively get a few upgrades in effectiveness since pattern convey more semantic significance than terms.

A. Y. Ng,D. M. Blei, and M. I. Jordan, in [3] has displayed a topic modeling strategy which is a standout among the most liked probabilistic text modeling strategies and has been immediately acknowledged by machine learning and text mining . It consequently classifies documents in a collection by various topics and speaks each document with numerous topics and their respective distribution .It has ambiguity.

D. M. Blei and C. Wang in [7] proposed Probabilistic topic modeling which separate long term user interests by examine content and representing it in terms of latent topics decided from user profiles. The significant documents are controlled by a user-specific topic model that has been separated from the users information needs. These topic model based applications are all connected to long term user extraction needs and related to the task of this paper. But, there is a nonappearance of explicit discrimination in the greater part of the language model based methodologies and probabilistic topic models. This weakness shows that there are still a few gaps between the present models and what we have to accurately model the significance of documents.

H. D. Kim, Y. Lu , C. Zhai and , D. H. Park, in [8] has proposed frequent patterns which are pre-created from the original documents and after that embedded into the original documents as a part of the input to a topic modeling , for example, LDA(Latent Dirichlet Allocation). The resulting topic representations incorporate both individual words and pre-generated patterns.

S. Mukhopadhyay, J. Mostafa,W. Lam, and M. Palakal,in [9] has concentrated on a multilevel strategy to intelligent information filtering. A filtering model is recommended that partitions the overall task into subsystem functionalities and highlights the need for numerous adaptation techniques to cope with uncertainties. Proposed filtering framework executed taking into account the model, utilizing traditional methodology as a part of information retrieval and artificial intelligence. These techniques incorporate document representation by a vector space model, document classification by means of unsupervised learning, and user modeling through reinforcement learning. Framework filters information based on content and a user's specific interest. The users interest are automatic learned with just limited user intervention as ideal relevance feedback for documents.

J. Xu,Y. Cao, H. Li, T.Y. Liu,Y. Huang, and H.W. Hon in [10] has considered adapting ranking SVM (Support Vector Machine) to document retrieval. In this framework the document filtering is dealt with as a classification task or

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

a ranking task. Techniques, for example, Naive Bayes and SVM appoint binary decisions to documents (irrelevant or relevant) as a special type of classification.

A. McCallum,X. Wang, and X. Wei [11] has proposed a topical n-Gram model which consequently and at the same time finds topics and extracts topically appropriate phrases. It has been consistently coordinated into the language modeling based IR task. Compared with word representation, phrases are more discriminative and convey more concrete semantics. Since phrases are less arguable than words, they have been broadly assessed as text representation for text retrieval, yet few studies here have demonstrated important enhancements in effectiveness.

D. H. Park, H. D. Kim, C. Zhai and Y. Lu in [12] demonstrate that the topics in the PBTM model are represented by patterns only.

G. Karypis, and A.Tagarelli in [13] has concentrated on a segment based way to clustering multi-topic documents. They tended to the issue of multitopic document clustering by leveraging the natural composition of documents in text segments, which bear one or numerous topics on their own.

Y. Li, N. Zhong, and S.T. Wu in [14] successful pattern discovery for text mining is proposed. Author concentrated on another and effective pattern discovery technique which incorporates the method of pattern deploying and pattern evolving, to improve effectiveness of utilizing and upgrading produced patterns for finding relevant and interesting information. In this work, an efficient pattern discovery strategy has been concentrated on to minimize the low-frequency and misinterpretation issues for data mining. The proposed approach utilizes two methods, pattern deploying and pattern evolving to refine the found patterns in text documents.

Our contribution is the use of dimension reduction techniques. Dimension reduction refers to the procedure of converting a set of data having vast dimensions into data with low dimensions ensuring that it conveys similar information concisely. Here we use missing values and low variance methods for dimensionality reduction. While exploring data, if we encounter too many missing values, we impute missing values because it would not have lot more details about data set. Also, it would not help to improve the power of model. Low variance method is useful where we have a constant variable (all observations have same value,10) in our data set. Power of model is not improved, because it has zero variance. In case of vast number of dimensions, we should drop variables having low variance compared to others because these variables will not explain the variation in target variables. In our contribution we also use document similarity clustering technique to recommend similar document to user as per his/her interest.

III. PROPOSED WORK

In proposed framework user's interest with numerous topics are considered. The proposed model Maximum Matched Pattern-based Topic Model with Dimensionality Reduction made up of topic distributions describing topic preferences of every document or the document collection and pattern-based topic presentations demonstrating the semantic meaning of every topic. User's interest includes various aspects identifying with different topics. The most motivating contribution of topic modeling is that it consequently arranges documents in a collection by various topics and represents to each document with various topics and their relating distribution.

A. Proposed System Architecture

In our proposed framework we have utilized MPBTM-DR model for information filtering

1) MPBTM-DR produce users interest model for various topic not for single topic.
2) Proposed model present topic preferences of collection of document and pattern based topic representation representing to the specific significance of every topic.
3) In this model pattern are partitioned into groups gathering called as equivalent classes in view of their characteristics and statistical features .In all present class pattern have same frequency and essential meaning with this structured representation .The majority part of representative pattern are recognized and this pattern utilized for information filtering.
4) When we get very much structured pattern based topic model then we proposed new ranking technique for document filtering. At the point when new document are coming in gathered documents then again maximum matched pattern are utilized to calculate importance of the new approaching document to the users interest.

In our proposed model, patterns are utilized to represent corpus and documents, which tackle the synonymy issue, as well as manage the low frequency issue of phrases. Frequent patterns are pre-generated from the original documents and afterward embedded into the original documents as a major aspect of the input to a topic model, for

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

example, LDA. The subsequent topic representations contain both individual words and pre-generated patterns. Our proposed model MPBTM-DR is not quite the same as the other model [18] as in the topics in the MPBTM-DR model are appeared by patterns only. Patterns in the model are very much organized so maximum matched patterns are distinguished and used to estimate. Collection of documents is given as input in the block diagram. Initially pattern improved LDA is utilized to discover most significant patterns which represent maximum matched patterns. Develop another transactional dataset from the LDA model which brings about document collection D. At that point, produce pattern based representations from the transactional dataset to show user needs of the collection D. When we get number of patterns some are helpful and a few patterns don't have any significance. We are discovering most useful and long patterns from the arrangement of pattern. At that point most ranked pattern is utilized for information filtering. After that relevance ranking of document is done. Yield of the framework is most applicable documents from the collection of documents.

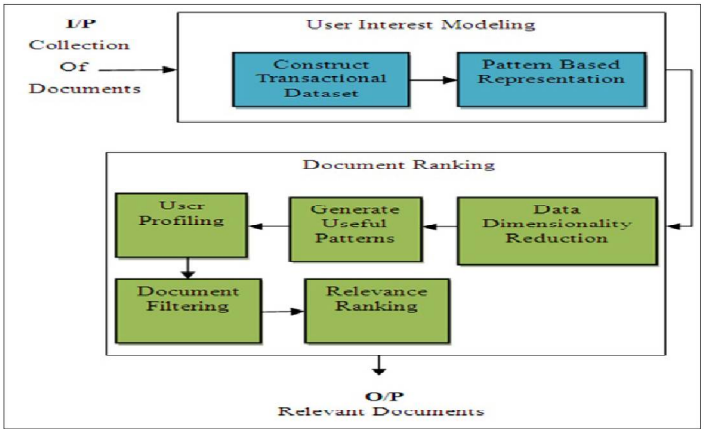

Fig 2: Proposed system (MPBTM-DR) Block Diagram

3.1 Pattern Enhanced LDA

Pattern is normally characterized as an arrangement of related terms or words. Patterns convey more semantic importance and are more understandable than individual words. It assumes an essential part in numerous data mining tasks coordinated toward finding interesting patterns in datasets. Pattern based representations are more meaningful and more accurate to represent topics than word-based representations.

Table 1: Example Results of LDA: Word-Topic Assignments

Topic Document	Topic Id 1 words	Topic Id 2 words	Topic Id 3 words
Document Id 1	compute,image,software, image,compute	compute,language,sructure	visual,logic,logic
Document Id 2	image,engine,engine	visual,structure,compute,structure	compute,model,learn
Document Id 3	image,compute,visual,graphic	visual,software,software,image	engine,visual,logic,model
Document Id 4	image,visual,environment	language,structure,compute	compute,model,logic

3.1.1 Construction of Transactional Dataset

The motivation behind the proposed design based strategy is to find related words (i.e., patterns) from the words allocate by LDA to topics.

Let $R_{di,Zj}$ represent to the word-topic assignment to topic Z_i in document d_i. $R_{di,Zj}$ is a sequence of words relegated to topic Z_j. We build an transaction dataset Γ_j for each word-topic assignment $R_{di,Zj}$ to Z_j , j=1,...,V and i=1,...,M here V is the number of topics and M is the number of documents. Let D= $\{d_1,...,d_M\}$ be the accumulation of documents, the transactional dataset Γ_j for topic Z_j is characterized as $\Gamma_j = \{\Gamma_{1j}, I_{2j},..., I_{Mj}\}$ where I_{ij} is known as topical document transaction which contains words with no repeated words. I_{ij} contains the words which are in document d_i and

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

allocated to topic Z_j by LDA. For example from Table 1 for topic (Z_1)in Document (d_1) word-topic assignment is{compute, image, software, image, compute}after eliminating the repeated transaction dataset Γ_1 is generated for topic (Z_1) and {compute, image, software} be the topical document transaction (TDT).

Table 2: Transactional Datasets Generated from Table 1

T	TDT	TDT	TDT
Document Id 1	{compute,image, software}	{compute,language, structure}	{visual,logic}
Document Id 2	{image,engine}	{visual,structure, compute}	{compute,model, learn}
Document Id 3	{image, compute, visual, graphic}	{visual,software, image}	{engine, visual,logic, model}
Document Id 4	{image,visual, environment}	{language,structure, compute}	{compute,model, logic}
	Γ_1	Γ_2	Γ_3

3.1.2 Topic based Pattern Representation

Frequent patterns are created in this stage. Table 3 demonstrates the frequent patterns produced for (Z_2).The frequent patterns are created from each transactional dataset.

Table 3: The Frequent Patterns for Computer (Z_2), $\sigma = 2$

Patterns	Support
{compute},{structure},{compute, structure}	3
{visual}{structure, language},{language},{compute, language},{compute, structure, language}	2

The frequency of the itemset X is defined as $\frac{\text{Support}}{|\Gamma_j|}$.....(1)

Give σ minimum support threshold, then an itemset X in Γ_j is frequent if support(X) $\geq \sigma$, here support(X) is the support of X which is the number of transaction in Γ_j having X.

3.1.3 Construction of Pattern Equivalence Class

The number of frequent patterns in topics is very large and a large portion of the patterns are not discriminative to speak about particular topic. As a conclusion, these topic representations are not sufficient to represent to the documents accurately. That implies the pattern based representation that represent to the user interest is not exact or adequate to be utilized to decide the importance of new documents. So the importance of the new documents is evaluated taking into account the most distinctive and representative patterns. These patterns are the more particular and representative patterns. Rather than frequent patterns, closed patterns are useful for topic representation and the numbers of these patterns are smaller than the number of frequent patterns for a dataset [16], [17].

1. Closed Itemset [15]: For a transactional dataset, an itemset X is a closed itemset if there exists no itemset X such that (1) $X \subset X'$ (2) support(X) = support(X′).
2. Generator [15]: For a transactional dataset G, let X be a closed itemset and T(X) consists of all transactions in Γ that contain X, then an itemset g is said to be a generator of X iff $g \subset X$, T (g) = T(X) and support(X) = support(g).
3. Equivalence Class[15]: For a transactional dataset , let X be a closed itemset and G(X) consist of all generators of X, then the equivalence class of X in Γ, denoted as EC(X),is defined as EC(X) = G(X) ∪{X}.

Each pattern in an equality classes have the same frequency. The frequency of a pattern demonstrates the statistical outcomes of the patterns. Table 4 demonstrates the Equivalence Classes in (Z_2).

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Table 4: The Equivalence Classes in (Z_2)

$EC_{2,1}$	$EC_{2,2}$	$EC_{2,3}$
{compute, structure}	{compute, structure , language}	{visual}
{compute}	{compute, language}	
{structure}	{structure,language}	
	{language}	

Every topic may be having a set of equivalent classes. These equivalent classes are useful to represent the user interest model. Let $E(Z_i)$ is the set of equivalent classes of topic Z_i and the user interest model for number of topics (V) are represented as $U_E = \{E(Z_1), E(Z_2), ..., E(Z_V)\}$. The proposed Information filtering model is presented in two algorithms: user profiling (creating user interest models) algorithm and document filtering (relevance ranking of incoming documents) algorithm

A. Algorithm (User Profiling)

Input: set of documents D, number of topics V minimum support σ_j as threshold for topic Z_j.
Output : user interest model, $U_E = \{E(Z_1), E(Z_2), ..., E(Z_V)\}$.

1. Load collection of input documents D
2. Perform stemming and remove stop words
3. Dimensions reductions based on user profile attribute
4. Apply LDA and generate word-topic assignment for V number of topics
5. $U_E = \{\}$
6. for each topic Z_j do
7. Construct the Transaction Dataset Γ_j
8. Generate all frequent patterns whose support(X) $\geq \sigma$ using pattern mining techniques
9. Construct equivalent classes $E(Z_j)$
10. Construct user interest model $UE = \{E(Z_1), E(Z_2)...E\ Z_V)\}$
11. End for

User profiling algorithm produces pattern based topic representations to represent the user's information needs. Algorithm uses functions such as construction of transactional dataset, construction of user interest model and construction of equivalence class.

3.2 Document Relevance Ranking

Significance of the documents is assessed based on the user interest model to filter out irrelevant documents. The maximum matched patterns in the equivalent classes are helpful to assess the relevance of the new incoming documents to the user interest. In light of the relevance of the documents the new documents will be ranked.

3.2.1 Relevance Based on Equivalent Class Frequency

The count of the pattern present in the incoming document is also considered while estimating the relevance. The more number of presences of patterns, relevance of the document is more. Therefore document relevance is evaluated utilizing the equation

$$\textbf{Rank}_E(d) = \sum_{j=1}^{V} \sum_{k=1}^{nj} |MC_{jk}^{d}|^{0.5} * \partial(MC_{jk}^{d}, d) * f_{j,k} * \upsilon_{D,j} \quad \text{.....(2)}$$

Where V is the quantity of topics and f_{jk} represents the frequency of equivalent class. Represents the maximum matched patterns to the equivalent classes is the topic distribution. Uniform distribution represents estimated uniform distribution and EC frequency is the Equivalent class frequency.

B. Algorithm (Document Filtering)

Input: User interest model, collection of new input documents Din

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Output: ranked documents
1. Rank (d) =0
2. for each document d do
3. for each topic Zj do
4. for each equivalence classes do
5. Scan all the equivalent classes and find out the maximum matched pattern
6. for each maximum matched pattern which is exist in d
7. Calculate rank using equation 1
8. Check the distribution of pattern in the document d
9. End for
10. Calculate the equivalent class frequency
11. Update Rank (d) using the equation (1)
12. Rank $_E$ (d) = rank (d) + $|MC^d_{jk}|^{0.5}$ * δ (MC^d_{jk}, d) *$f_{j,k}$ * $\vartheta_{D,j}$
13. End for
14. End for
15. End for

The Document filtering algorithm ranks the incoming documents in view of the significance of the documents to the user's needs. Algorithm scans the documents to find maximum matched pattern and update the ranking of documents.

IV. MATHEMATICAL MODEL

S= {S, E, I, O, Fs, DD, NDD, Øs}
S: Initial state: User asks for information about a product. The user probably wants to find documents that contain information about different aspects of the product.
E: End state: User information needs are produced regarding relevant documents.
I: Input: A set of positive training documents and minimum support for topic.
O: Output: Ranking of documents.
Fs: Functions = {f_1, f_2, f_3, f_4, f_5, f_6}
f_1 → Construct transaction dataset,
f_2 → Construct user interest model,
f_3→ Construct equivalence class.
f_4 → Data dimensionality reduction
f_5 → Calculate rank
f_6 → Update rank
DD: Deterministic Data: Ranked documents
NDD: Non- Deterministic Data:
Øs: Processing will not be done

V. EXPERIMENTAL SETUP AND RESULTS

A. Experimental Setup

The Reuters Corpus Volume 1 (RCV1) dataset contains a numerous topics and a lot of information. 100 sets of documents are utilized as a part of TREC [20] filtering track. In this paper, to separate from the term topic in the LDA model, the term collection is utilized to refer to a set of documents in the TREC dataset. The implementation environment for the proposed system uses the windows operating system, Java SE Development Kit 7 and Eclipse Juno version.

B. Results

Cluster Purity Results is shown in the fig 4.Euclidean distance, Cosine similarity, extended jacard coefficient and MPBTM-DR model are used for analysis.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

1. Euclidean distance is calculated by using formula below where, p = (p_1, p_2,..., p_n) and q = (q_1, q_2,..., q_n) are two documents .If answer is close to zero means document are more similar.

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

……..……..……...……(3)

2. Cosin similarity is calculated by using formula below where, two vectors of attributes, *A* and *B*, the cosine similarity, *cos(θ)*, is represented using a dot product . If answer is close to one means document are more similar.

$$\mathbf{similarity} = \cos(\boldsymbol{\theta}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

……...…….…….. (4)

3. Extended Jaccard coefficient is calculated by using formula below, where d1 and d2 are two documents. If answer is close to one means document are more similar.

Extended Jaccard coefficient=$\frac{d1*d2}{d1d1+d2d2-d1d2}$

..... (5)

4. To evaluate effectiveness of MPBTM-DR model. Two measures are used: F1 and MAP (Mean average precision)

F1 is calculated by F1=$\frac{2*precision*recall}{precision+recall}$

..... (6)

MPBTM-DR model outperform well compared with other term based methods. Here total 1740 documents are considered for analysis. Document contains total 10 topics. Result of recommendation is shown in fig 3. Figure shows recommended document ID, Rank value and document contents. Most ranked document appeared at first and least rank document appeared at last.

Table 5: Topic with Selected Words in That Topic

TopicId	Word list
1	Compute,engine,software,visual,conference,graphic,associate,proceed,environment, image
2	Program,language,work,generate,compile,include,implement,run,file,structure
3	Algorithm,learn,intelligence,theory,model,robot,mathematics,problem,artificial,logic
4	Project,design,system,develope,software,time,professor,techniques,testing
5	Compute, science,universe,depart,professor,phone,fax,office,engine,inform
6	Research,group,interest,universe, graduate,work,student,public,member,laboratory
7	Inform,postscript,object,project,time,data,database,version,paper,list
8	System, parallel, network, compute, distribute, perform, operation, architecture, ieee, application
9	Page,home,link,web,mail,site,www,java,interest,picture
10	Assign,class,homework, office, lecture, hour, due,solution,read,exam

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Show Rank Document :: Show

Document Id	Rank Value	Content
82	84	course introduct oper system section fall instructor marvin solomon offic compu
154	76	project vision touch guid manipul group mit artifici intellig lab nonlinear system le
47	74	course introduct artifici intellig professor david leak mail leak indiana offic hour t
90	72	course seminar read multimedia digit audio comput music network new group uc
55	69	faculty jame laru laru wisc associ professor comput scienc depart comput scienc
182	69	faculty richard fateman professor comput scienc univers california berkelei fate
29	65	faculty david assist professor depart comput scienc john hopkin univers baltimo
135	65	faculty mari vernon professor comput scienc industri engin comput scienc depar
83	64	project cornel robot vision laboratori web cornel robot vision laboratori match m
95	63	faculty mok associ professor faculti fellow comput scienc electr engin massachu
178	63	faculty yanni ioannidi yanni wisc research interest databas manag system scien

Fig 3: Recommendation Result

Table 6: Cluster Purity Results

Top'k'documents	EuclideanDistance	Cosin Similarity	Extended Jacard Coeff	MPBTM-DR model
3	0.33	0.33	0.33	1
5	0.4	0.6	0.6	0.8
7	0.28	0.71	0.71	0.85
9	0.33	0.77	0.66	0.77
11	0.27	0.72	0.54	0.81
13	0.3	0.61	0.61	0.76
15	0.26	0.6	0.66	0.73

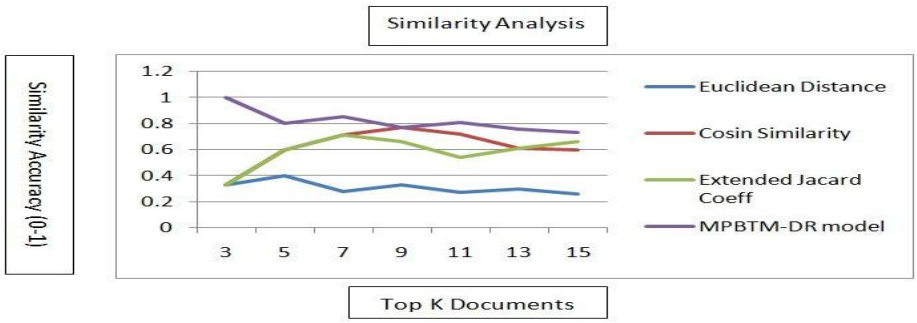

Fig 4: Document Purity Details

VII. CONCLUSION

Modelling strategies are extremely useful in information filtering, document ranking, information retrieval and recommendations. Here paper exhibits an inventive pattern improved topic model for information filtering including user interest modelling and document importance ranking. The proposed MPBTM-DR model creates pattern upgraded topic representations to demonstrate user's interest over

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

numerous topics. In the filtering task, the MPBTM-DR chooses maximum matched patterns, for assessing the significance of incoming documents.

ACKNOWLEDGEMENT

We are happy to express our feelings of appreciation to all who rendered their important direction to us. We might want to express our gratefulness and on account of Prof. Dr. P. C. Kulkarni, Principal, G. E. S. R. H. Sapat College of Engg., Nashik. We are likewise appreciative to Prof. N. V. Alone, Head of Department, Computer Engg., G. E. S. R. H. Sapat College of Engg., Nashik. We thank the unknown analysts for their remarks.

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. 4249.
2) F. Beil, M. Ester, and X. Xu, Frequent term-based text clustering, inProc. 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.,2002, pp. 436442.
3) Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal, Mining frequent patterns with counting inference, ACM SIGKDD Explorations Newslett., vol. 2, no. 2, pp. 6675, 2000.
4) H. Cheng, X. Yan, J. Han, and C.-W. Hsu, Discriminative frequent pattern analysis for effective classification, in Proc. IEEE 23rd Int. Conf. Data Eng., 2007, pp. 716725.
5) R. J. BayardoJr, Efficiently mining long patterns from databases, in Proc. ACM Sigmod Record, 1998, vol. 27, no. 2, pp. 8593.
6) X. Wei and W. B. Croft, LDA-based document models for ad-hoc retrieval, in Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2006, pp. 178185.
7) C. Wang and D. M. Blei, Collaborative topic modeling for rec- ommending scientific articles,"in Proc. 17th ACM SIGKDD Int. Conf.Knowl. Discov. Data Min., 2011,pp. 448-456
8) D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent dirichlet allocation,"J.Mach. Learn.Res., vol. 3, pp. 993-1022, 2003.
9) J. Mostafa,M.Palakal, and W.Lam, A multi- level approach to intelligent information filtering: Model, system,and evaluation,ACM Trans. Inform. Syst.,vol. 15,no. 4, pp. 368-399, 1997.
10) T. Hofmann, Probabilistic latent semantic indexing, in Proc. 22nd Annu. Int. ACM SIGIR Conf. on Res. Develop. Inform. Retrieval, 1999, pp. 5057
11) X. Wang, A. McCallum, and X. Wei, Topical n-grams: Phrase and topic discovery, with an application to information retrieval, in Proc. 7th IEEE Int. Conf. Data Min., 2007, pp. 697702.
12) H. D. Kim, D. H. Park, Y. Lu, and C. Zhai, Enriching text representation with frequent pattern mining for probabilistic topic modeling, in Proc. Am. Soc. Inform. Sci. Technol., 2012, vol. 49, no.1, pp. 110.
13) A. Tagarelli and G. Karypis, A segment-based approach to clus- tering multi-topic documents, Knowl. Inform. Syst., vol. 34, no.3,pp. 563595, 2013.
14) N. Zhong, Y. Li, and S.-T. Wu, Effective pattern discovery for text mining, IEEE Trans. Knowl. Data Eng., vol. 24, no. 1, pp. 3044, Jan 2012.
15) Yang Gao, Yue Xu, and Yuefeng Li, Pattern-based Topics for Document Modelling in Information Filtering,in Knowledge and Data Engineering,2015.
16) M. J. Zaki and C.-J. Hsiao, CHARM: An efficient algorithm for closed itemset mining. in Proc. SDM, vol. 2, 2002, pp. 457473.
17) Y. Xu, Y. Li, and G. Shaw, Reliable representations for association rules, Data Knowl. Eng., vol. 70, no. 6, pp. 555575, 2011.
18) Y. Gao, Y. Xu, and Y. Li, Pattern-based topic models for informa- tion filtering, in Proc. Int. Conf. Data Min. Workshop SENTIRE,2013, pp. 921928.
19) Ms Manjiri M. More, Prof. Archana S. Vaidya,A Survey Paper On Modeling Methods for Information Filtering And Relevance Rank- ing Of Documentsin International Research Journal of Engineering and Technology (IRJET) Volume: 02 Issue: 09 — Dec-2015
20) S. E. Robertson and I. Soboroff, The TREC 2002 filtering track report, in Proc. TREC, 2002, vol. 2002, no. 3, p. 5.
21) https://www.knime.org/files/knime\seventechniquesdatadimreduction.pdf

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

BIOGRAPHY

M.M.More Is pursuing the Masters in Computer from G.E.S. R. H. Sapat College of Engg., Nashik under Pune University. She has pursued her Bachelor's Degree in Information Technology from Cummins college of engineering for women, pune under Pune University.

A.S.Vaidya Is currently working as an Assistant Professor in the Computer Engineering Department of GESs R. H. Sapat College of Engineering Management Studies and Research ,Nashik (INDIA).She received her Master's Degree in Computer Engineering from V.J.T.I.,Mumbai University (INDIA) in 2010 and Bachelors Degree in Computer Science and Engineering from W.C.E.,Sangli Shivaji University(INDIA) in 2002. She has an academic experience of 13 years (since 2002).