Volume 8, No. 5, May-June 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

Probabilistic latent feature discovery model and Multi label content categorization in e-learning Using r package

Dr.S.Arulselvarani
M.Sc.,M.Phil.,Ph.D.
Guest Lecturer,
Computer Science Department,
Govt. Arts College,
Trichy, Tamil Nadu, India

Abstract:In an E-learning environment, users have access to huge online documents of various learning materials, hence finding the suitable learning content becomes harder. In this research paper, author uses high level machine learning approach using R packages to propose a method namely probabilistic latent feature discovery approach and multi label content categorization in e-learning. Probabilistic latent feature discovery model is a generative model for multi label e-learning content categorization, which has significant, effects of both accuracy and efficiency. Latent Dirichlet Allocation (LDA) is an effective probabilistic approach to develop a topic models depends on a formal generative model of a document, viable and efficient algorithm in e-learning text modeling. Authors propose a generative LDA-based model within the information retrieval approach, and estimate it on an e-learning environment, training the learning documents via Gibbs sampling. The predictive distribution LDA fit model is used to predict new words. The experimental results on e-learning, multi label content categorization demonstrate the accuracy and effectiveness of the proposed research approach.

Keywords: e-Learning, Information Retrieval, Topic modeling Generative process,LDA, Latent feature discovery, content categorization.

1.INTRODUCTION

The increasing popularity and usefulness of e-learning approach has created the requirement for the personalized learning content search model which can be used to optimize the useful learning system for the learner. The learning content search model is a type of information filtering used to identify the group of topics from the learning documents that are relevant to the learner. The multi label content categorization model provides facility to learners about the learning content they might desire to examine. E-Learning is an innovative way which complements the traditional learning system. It enable people to learn new technologies at anytime and anywhere. It includes educational training, the delivery of information and guidance from the facilitator. An information retrieval system is an application that stores and manages information on text documents. The objective of information retrieval approach is to give users with those documents that will convince their information need.

A balancing method of finding a suitable text document over the web is text keyword search. This is an influential approach, but it is also limited. Forming queries for finding new user learning document can be complicated as learner may not know what to look for; search is mainly based on learning text content, and search is only good for directed searching, while many researchers would also like a “feed” of new and interesting document [1].

There are several learners, learning text documents available in the recent electronic structure. Such documents represent a considerable amount of text data that is easily accessible. Finding correct text document value in this large collection needs association, much of the recent work of categorizing the text documents can be mainly automated through various machine learning algorithms using supervised and unsupervised data mining approach.

The accuracy and performance of such systems very much influence their usefulness [2]. With the existing machine learning algorithms, a number of new and innovative methods are involved in the computerization of text content classification in a supervised learning environment [3]. The task of mining approach is to dynamically classify documents into predefined multi label classes based on their content. Many advanced computational algorithms have been designed to deal with automatic text document supervised learning approach based classification [4]. The most common modern computational intelligence techniques used for this purpose include Association Rule Mining, Implementation of Naïve Bayes Classifier, Genetic Algorithm, and Decision Tree.

Mining of text document means the application of machine learning algorithms to collections of documents consisting of words and sentences. The text mining approach includes standard classifier learning, clustering, and pattern recognition. Supervised classifiers for text documents are very useful for many applications. Major uses for document classifiers include email spam detection and personalization of articles. Most classifiers for documents are designed to categorize according to learners subject matter. In many recent applications of multi class classification, a single document can fit into more than one group, so it is correct to predict text more than one class label. This task is explicitly called multi label text document categorization.

In the recent past, generative approach based topic model has become more popular in some text document related tasks. Topic Model supposes documents and huge text corpus composed of various topics and then the documents can be thought of bag of collection of topics. Probabilistic

Topic models can solve the problem successfully about text terms dependency.

The topic is viewed as Probability distribution which implies semantic coherence about words. For example, a topic related to fruit would have high probabilities of the words “orange”, “apple”, and even “juicy”. Wallach [5] demonstrated the “bag of topics” to best in performance to “bag of words” in unigram and bigram schemas.

There are several topic modeling research in the past ,such as Latent Semantic Analysis [6] and the Probabilistic Latent Semantic Indexing (PLSI)[7]. Two of the most popular topic modeling algorithms are Probabilistic Latent Semantic Analysis (PLSA) [8] and Latent Dirichlet Allocation(LDA) [9]. Both methods are related to generative models where each document is a mixture over a rigid set of essential topics, where each topic is characterized as a distribution over various words. The topic probabilities can be indirectly inferred by maximizing the log-likelihood of the data to be generated. One limitation of these two approaches is that they fail to consider the intrinsic geometrical structure of the multidimensional data space [10].

In order to overcome the drawbacks of previous topic models such as PLSI which is a MAP/ML estimated LDA model under a uniform Dirichlet distribution [11], this paper addresses a variant of LDA and an extension of Language Model [12], which is a novel model for text categorization as we know. This generative model represents words set of each category with a mixture of topics assumed independent as many state-of-the-art approaches did, and extends these approaches to estimate maximum a posteriori of category, language model parameters by assuming that variance parameters would be multinomial and Dirichlet parameters of a category language , moreover, [13] LDA model is applied to find scientific document topics. Recently, in the field of machine learning, there has been renewed research interest in combining these two paradigms into a hybrid framework to gain both merits [14, 15].

Markov Chain Monte Carlo (MCMC) algorithms for approximate inference is widely used as an inference method for a variety of topic models [30,31,32]. In the MCMC context, the usual procedure is to integrate out the mixtures and topics, a procedure called collapsing—and just sample the latent variables. The methods have been tested on 100 Chinese standard text images[33].The two standard inference methods namely variational bayes and collapsed gibbs have gained great accomplishment in learning LDA, as tested by hopeful test results on four standard document datasets[34].

Variational inference [16, 17] focuses on the balanced estimates of standard procedures for potentially biased, but computationally well-organized algorithms whose arithmetical convergence are easy to evaluate. Under assumptions of ‘‘bag of words’’ [27] and ‘‘bag of topics’’, some well known topic models are proposed, such as mixture unigram model [28] and finite mixture model [29]. Due to the progress of Web technologies and human activities in online participation, content generation has become easier than before. Text articles appear everywhere on the Web, such as, news website, blogs, search engine, and so on. One of the most important characteristic lies in their dynamics and quantity, which lead to great challenges and effort in dealing with a text stream [23], [24]. It is also required to provide reasonable topics for many kinds of topic analysis tasks, such as opinions recognition, topic propagation and topic evolution [25], [26]. Hence, the automatic topic discovery on the Web becomes necessary to meet these requirements.Document import to R environment, text corpus handling, document preprocessing, metadata management, and creation of bag of words, term-document matrices. Our focus is on the main aspects of getting started with probabilistic latent feature discovery model and multi label content classification in R an in-depth description of the modern text mining platform offered by tm package [18]. An introductory article on text mining in R was published in R News [19] The R package topic models currently provides an interface to the code for fitting an LDA with Gibbs sampling. Text corpus topic models are constructed using standard package [20]. R, an environment for various statistical calculation and generating the graphs [21], the Comprehensive R Archive Network features two packages for topic model fit: lda and topic model. The lda package in R [22] provides collapsed Gibbs sampling method for posterior probability of the latent features.

2. METHODOLOGY

The following research contributions are provided in this paper:

In this study, we focus on the probabilistic approach based latent feature discovery of the learning documents and text document categorization using R package. We conduct topic modeling based analysis and infer the hidden topics to understand the learning contents of individual users and to score their interestingness.

1. We propose a novel method for text document dataset preparation using the R package (tm). Before building the model we scan the input text dataset content and perform various operations over text document and find out various undesirable characters and topic terms such as a bag of words representation.
2. We also explore the vector model which contains learning document and terms in the form of rows and columns as an object. This representation provides the association between terms present in the learning document.
3. We model textual contents in learning document dataset as probabilistic latent feature structure using the LDA Latent Dirichilet Allocation ,where learning documents are viewed as a collection of topics. The machine learning algorithm is applied to the topics and validates the model using log-likelihood values. The predictive distribution LDA fit model is used to calculate a predictive distribution of latent words. Description of various learning topic results from a latent feature discovery model with visual representation.
4. We conduct experiments on a real dataset.From the probabilistic latent feature discovery model, each learning document has a topic associated with it. These resultant features can be used to obtain similarity between the documents and topic distribution. The results prove our multi label model content categorization is more effective than the existing work.

3.GENERAL ARCHITECTURE

3.1. The General architecture of probabilistic latent features discovery model and multi label content categorization

The application presented shows a learning topic representation of information is based on multi label content categorization technique in the R environment. The topics represented are used for the learners to search the relevant documents. This research work, presents a novel method to construct a probabilistic latent feature discovery model and a multi label content categorization. The general architecture is presented in Fig 1. Give a document input data, a number of steps are followed by the e-learning document's topic representation. The first step is the e-Learning document load to R environment for document corpus creation. Dataset preprocessing, document term vector construction, tf-idf computation, probabilistic latent feature discovery and learning content categorization represents the second step. At this stage information cleaning techniques can be successfully applied to the document such as white space removal, lower case conversion, stop words removal, stemming according to the weight values (tf-idf).The next step is to perform multi label content categorization to group the documents based on the model generated by latent feature discovery. The final stage is to represent the learning topics through the predictive fit model as the testing phase.

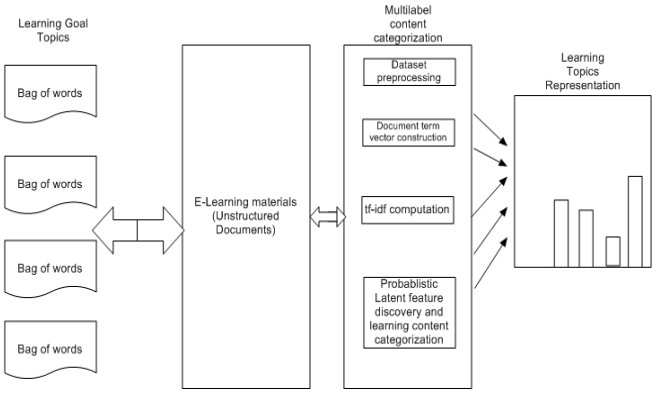

Fig1.Genera1architecture of probablistic latent feature discovery and multilable content categorization In e-learning

3.2.e-Learning documents Bag of Words Presentation:

The bag-of-words model is a simplified representation used in natural language processing and information retrieval (IR). In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity.

In order to do probabilistic latent feature discovery over unstructured e-Learning text documents first we represent various words in the documents. The primary task is to identify the set of all words used at least once in at least one document (Vocabulary). Vocabulary is a collection of word items of size n. Once the terms has been fixed, each document is represented as a vector with integer values of length n. If this vector is y then its k th component yk is the number of occurrences of the word k in the document. The length of the document is m=y1+y2+…yn.

3.3. Dataset Preprocessing

Dataset preprocessing has been an often neglected but important step in the data mining process. The unstructured learning content is highly susceptible to noise, missing values, and inconsistency. In order to help improve the quality of the data and, consequently, of the search results raw data is preprocessed so as to improve the efficiency and ease of the learners search process. To start learning document analysis usually this stage includes various operations (removal of unwanted characters such as white space, lowercase text character conversion and punctuation). This process is mainly called noisy character data removal. In R package supports tm library includes various text processing methods to do the various operations that can be applied to the learning text document corpus.

3.4. Constructing document term vector

After dataset preprocessing, we can construct document term vector associated with learning text documents. It is the simple vector having occurrences of all terms occurring in the list of documents. In a document-term vector, rows represent documents in the collection and columns represent to terms. The document term matrix contains sparse and non sparse values (0 or non-zero value).

3.5. tf-idf weight computation:

Term frequency–inverse document frequency, is a computational statistic that replicates how important a word is to a document in a text corpus. In Information retrieval the tf-idf is used as a weight calculating factor. If the tf-IDF value increases, then the number of times a word appears in the document also increases. The frequency of the word in the corpus is larger indicates some words are more common than the other words. Tf-idf term weighting methods are used by several search engines to score and rank the document relevance given by the learners query. Tf –idf is used in various areas including text document classification and summarization techniques.

D – Collection of documents (Corpus)

t-term, d-document, t_{i-} i th term , d_j –j th document , $t_i d_j$ belongs to D. The total occurrences of t_i in d_j is referred as term frequency.

The idf (inverse document frequency) for a term t_i is calculated by $idf_i = \log(|D|/|\{d:t_i \in d\}||D|/|\{d:t_i \in d\}|)$

|D| is the collection of documents in the text corpus

$|\{d:t_i \in d\}|$ is the number of documents in which the term occurs. The method to the calculated tf-idf is defined as tf_{ij}*idf $_I$. The larger weight term, which are less appearance in the collection of documents.

3.6. Probabilistic latent feature discovery model and multi label content categorization

3.6.1 Probablistic multinomial distribution

Once dataset representation is completed, the next immediate step is to select a model for learning documents. A representation is a way of training an entity as a data structure. A probabilistic model is an idea of a set of entities. Given a training set of learning documents, the parameter values of a probabilistic model that create the training documents have high probability value. Then given a test document, we can estimate its probability according to the model. If the probability is high it indicates more documents are similar in testing and training set. Mathematically, probabilistic multinomial distribution is

Prob(a;b)= $(m!/\prod_{i=1}^{n} a_i !)(\prod_{i=1}^{n} b_i^{a_i})$ (1)

a-vector contain non-negative intergers
b-parameter real valued vector.
a,b vectors have length n.
data point is a document containing m words.

b_i is the probability of word I while a_i the count of word i. The word i appears in the document contributes an amount b_i to the total probability(b_i^ai).
Equation (1) contains two factors, first one represent the multinomial coefficient (number of different word sequences that provide the same counts) and the second one represents the probability of any individual member of the equivalence class of a.
log probabilities :

log prob(a;b)= $\log m! - [\sum_{i=1}^{n} \log a_i !] + [\sum_{i=1}^{n} a_i * \log b_i]$

(2)

The probabilities of the individual documents refuse exponentially with length m, and then the above step is required.
Given a set of training learning documents, the maximum likelihood estimation of the i th parameter is

$$b_i = (1/T) * (\sum_a a_i)$$

T – Normalizing constant ,which is the sum of the sizes of all training documents($\sum_a \sum_i a_i$).

3.6.2 Multi label categorization generative process:
The generative process is a parameterized distribution, learning based on maximum likelihood. In order to organize the documents from a collection of documents by using unsupervised learning method.
A generative method for a single document is

Step1: Set a multinomial distribution using parameter c of length W.(used to set up the probability distribution)
Step 2: For each word in the document draw a word w according to c.(to produce the observed training data).

For a collection of documents for multi label groups the generative process is:
Step 1: Set a multinomial d over groups 1 to L
For group number 1 to group number L
Set a multinomial with parameter vector c_k
Step2: For document number 1 to document number Q
Draw a group e according to d
For each word in the document
Draw a word v according to $c_{c.}$
e-integer between 1 and L
The value of the e is latent for each document.
The global probability distribution is

F(a)= $\sum_{k=1}^{L} d_k f(a; c_k)$

a-document
c_k is the parameter vector of kth multinomial.
L-number of components in the mixture model
$f(a;c_{k)}$ is the distribution of component number k.
d_k- scalar variable.

3.6.3 Multi label content categorization by topics:
Multi label content categorization of topics is performed by using Latent Dirichlet Allocation (LDA) generative process method. Each learning document is referred as a combination of multiple distinct topics. By applying LDA features to fit different learning document dataset, we can easily explore the information about the topic present in the document and resulting multi label categorization.
The LDA generative process model is as follows:
Input:Dirichlet distribution with parameter vector d of length L
Input:Dirichlet distribution with parameter vector f of length W

For topic number 1 to topic number L
Draw a multinomial with parameter vector c_k according to f
For document number 1 to document number Q
Draw a topic distribution , a multinomial b according to d
For each word in the document
Draw a topic e according to b
Draw a word v according to $c_{e.}$
e- Integer between 1 and L for each word.
The above steps are used to construct term distribution for each topic model, the proportion of topic distribution for each learning document distribution and word association with topic. In this model LDA is a bag of words representation.

3.6.4 Predictive distribution LDA Fit model
Modeling documents is to fit a topic model to the text document corpus. To generate topic description that is subject to predictive distribution LDA fit model to compute the predictive distribution of new words.
The predictive Probability is : $prob_d(w)$=

$$\sum_i (a_{d,i} + b)(c_{w,i} + d)$$

b : The scalar value of the Dirichlet hyperparameter for topic proportions
d : The scalar value of the Dirichlet hyperparamater for topic multinomials
$a_{d,I}$: matrix where each entry is a numeric proportional to the probability of seeing a topic (row) conditioned on document

$c_{w,I}$: matrix where each entry is a numeric proportional to the probability of seeing the word (column) conditioned on topic (row)

3.6.5 Learning the Documents:
For machine learning, the training data input are the words in all the documents. The distributions(priori) d, f are assumed to be known and fixed, as are the number L of topics, the number M of documents and the cardinality W of the vocabulary(words dictionary).Learning has two main goals: to infer the document specific multinomial b for each document and to infer the topic distribution $c_{k.}$ The collapsed Gibbs sampling learning algorithm is used to infer the latent value e for each word occurrence in each

document. This algorithm starts with assigning a random topic to each word in each document. Then each iteration step the algorithm resembles a new topic.After a large number of iterations, a model tends to meet the topic assignment. The predictive distribution LDA fit model used to calculate a predictive distribution of new words. This fit model is useful for making useful predictions about held-out word. Finally Log-Likelihood model validation is used to compare two models having different parameters based on their log-likelihood. A model with high likelihood is considered as appropriate. Model validation is used to measure the model performance and also ensure that the topic model is able to generalize from the training document in a useful way.

4. EXPERIMENTS AND RESULTS

We do some experiments to probabilistic latent feature discovery model and multi label content categorization on predictive distribution LDA fit to evaluate the effectiveness of the proposed methods. The experiments are done on the publicly available real world datasets Cora collection of 2410 scientific documents.

Initially the input dataset is loaded into R. The dataset contains documents and topics. In the experiments, the datasets are first preprocessed by removing the standard list of stop words (using tm package in R) before feeding to the selective predictive distribution LDA fit model. The multi label content categorization is obtained from generative model LDA via collapsed Gibbs sampler. The model LDA is built with the following input parameters

i) The vocabulary associated with the corpus.
ii) Number of iterations for the gibbs sampling(100 iterations)
iii) The third parameter describing the term distribution for each topic (p(term|topic)=0.1)
iv) The next parameter describing the topic distribution for each document (p(topic|document)=0.1)
v) The last parameter indicates the log-likelihood to determine the convergence of LDA model .

The LDA model result with following features

1. a list of vectors represent topic association of each term in each document

```
result$assignments[[1]]
 [1] 3 5 5 4 2 2 4 2 2 5 2 2 5 2 5 2 2 2 5 5 2 2 2 2 2 2 5 5 5 2
6 5 5 5 5 5 3 5
[39] 6 2 2 2 2 5 0 5 5 5 2 5 5 2 5 4 2 2 5 5 5 5 2 5 4 6

 result$assignments[[2]]
 [1] 7 8 8 0 7 7 8 7 7 8 8 7 8 8 7 2 8 8 7 8 8 8 7 8 8 7 8 8 8 7
7 7 8 7 7 7 8 7
[39] 7
```

2. A matrix represents the number of times in each document was associated with each of the topics. Rows represent topics and columns are documents as shown inTable.1. From the result table document 1 would have 36 terms associated with topic 3, and 45 terms with topic 6.

Table 1: result$document_sums[,1:10]

Topic/ Document	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	1	1	44	56	1	0	24	0	0	0
2	0	0	0	5	0	0	18	0	0	0
3	36	2	0	19	42	4	0	0	1	21
4	2	0	6	0	15	0	3	13	80	0
5	5	0	16	0	0	0	0	0	0	0
6	45	0	23	26	0	0	3	11	0	0
7	3	0	0	0	0	0	1	0	0	0
8	0	20	30	0	3	0	8	0	0	0
9	0	23	0	0	2	31	0	0	11	2
10	0	0	0	8	0	0	0	0	0	0

The log-likelihood convergence of LDA model graphically presented in Fig2. From the curve plot of the log-likelihood, It is more convenient to work with natural logarithms. It has been observed that the log - likelihood converge slowly after 40 iterations only.

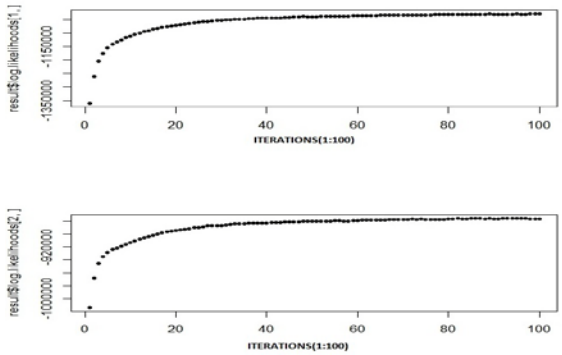

Fig2: Log-likelihood values convergence plot

The topics distribution graph of LDA model is graphically presented in Fig 3.The plot for the topics association for the 100 first documents of our dataset is visualized graphically. The stacked bar chart gives a clear perspective view of document distribution in the topic space. The chart shows the weight of each topic for documents in the dataset.

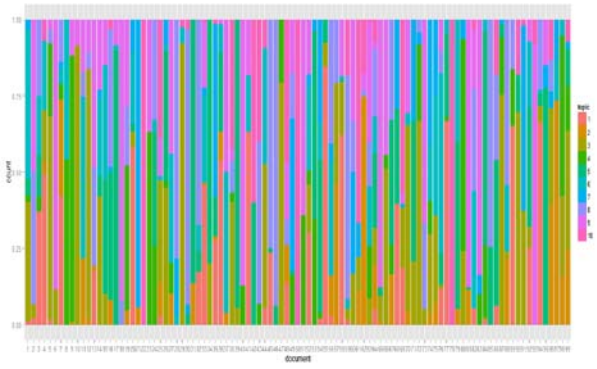

Fig3: Topics distribution in documents

The following document similarities between documents shown in Table2 are exploring the document similarity and relations in order to obtaining the topic association of documents. The resulting matrix is a symmetric matrix.

Table 2: Document similarity matrix outcome

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	1.00	0.04	0.33	0.50	0.59	0.08	0.09	0.53	0.04	0.62
2	0.04	1.00	0.35	0.05	0.14	0.75	0.19	0.00	0.10	0.14
3	0.33	0.35	1.00	0.78	0.08	0.00	0.73	0.32	0.10	0.00
4	0.50	0.05	0.78	1.00	0.29	0.04	0.73	0.26	0.00	0.29
5	0.59	0.14	0.08	0.29	1.00	0.16	0.07	0.26	0.35	0.94
6	0.08	0.75	0.00	0.37	0.16	1.00	0.00	0.00	0.14	0.22
7	0.09	0.19	0.73	0.74	0.07	0.00	1.00	0.13	0.09	0.00
8	0.53	0.00	0.32	0.26	0.26	0.00	0.13	1.0	0.76	0.00
9	0.04	0.10	0.10	0.00	0.35	0.13	0.09	0.76	1.00	0.03
10	0.61	0.14	0.00	0.29	0.93	0.22	0.00	0.00	0.03	1.00

The distance between the documents is obtained through the Euclidean distance function to find out the distance values between the document and it is shown in table 3.

Table 3: Distance between the documents

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0.00	64.4	68.3	61.6	47.7	63.6	63.3	51.1	97.4	47.8
2	64.4	0.00	57.4	70.7	50.5	21.6	39.3	34.9	83.3	34.6
3	68.3	57.4	00.0	42.3	72.1	68.7	48.2	57.3	96.0	64.4
4	61.6	70.7	42.3	00.0	67.5	71.3	47.1	63.0	103.6	62.5
5	47.7	50.5	72.1	67.5	00.0	50.1	52.9	43.6	77.4	26.0
6	63.6	21.6	68.0	71.3	50.1	00.0	42.2	35.5	82.5	33.6
7	63.3	39.3	43.2	47.1	52.9	44.2	00.0	33.6	83.8	37.7
8	51	34	57	63.	43	35	33	00	68.	27
9	97.4	83.3	96.0	103.6	77.4	82.5	83.8	68.7	00.0	82.9
10	47.8	34.6	64.0	62.5	26.0	33.6	37.7	27.1	82.9	00.0

The high usage terms for each topic result are shown in Table.4. The sorted list of topics associated with corresponding term output have been generated using the LDA model.

Table 4:Topic description generated by LDA model

	1	**2**	**3**	**4**	**5**
1	"learning"	"theory"	"bayesian"	"algorithm"	"network"
2	"reinforcement"	"logic"	"data"	"learning"	"neural"
3	"control"	"learning"	"markov"	"bounds"	"networks"
4	"robot"	"revision"	"distribution"	"queries"	"learning"
5	"agent"	"representation"	"models"	"polynomial"	"input"
6	"agents"	"belief"	"model"	"efficient"	"visual"
7	"system"	"inductive"	"estimation"	"instruction"	"recurrent"
8	"environment"	"knowledge"	"chain"	"number"	"units"
9	"actions"	"examples"	"mixture"	"execution"	"recognition"
10	"decision"	"system"	"methods"	"model"	"training"

	6	**7**	**8**	**9**	**10**
1	"search"	"genetic"	"design"	"learning"	"research"
2	"problem"	"population"	"system"	"training"	"report"
3	"function"	"fitness"	"reasoning"	"decision"	"university"
4	"optimization"	"evolutionary"	"knowledge"	"classification"	"technical"
5	"algorithm"	"crossover"	"case"	"data"	"grant"
6	"optimal"	"evolution"	"learning"	"algorithm"	"science"
7	"problems"	"programming"	"planning"	"feature"	"supports"
8	"algorithms"	"neural"	"adaptation"	"methods"	"department"
9	"genetic"	"evolve"	"similarity"	"algorithms"	"national"
10	"solution"	"programs"	"cases"	"accuracy"	"part"

"

After obtaining the sorted list of topics, the predictive distribution LDA fit model is applied to compute the new terms in the test sample. Topic description generated by predictive distribution LDA fit model output shows in Table 5.

Table 5: Topics description generated by prediction distribution LDA fit model

	[,1]	**[,2]**

[1,]	"problem"	"learning"
[2,]	"algorithm"	"paper"
[3,]	"function"	"system"
[4,]	"search"	"algorithm"
[5,]	"algorithms"	"design”
[6,]	"model"	"problem"
[7,]	"problems"	"knowledge"
[8,]	"method"	"data"
[9,]	"paper"	"approach"
[10,]	"results"	"methods"

5. CONCLUSIONS AND FUTURE WORK

In this paper , we have presented the probabilistic latent feature discovery model and multi label content categorization in elearning.This model helps peer learners to search the appropriate learning contents from the categorized documents. The new effective system is built to extract topics from the unstructured documents using mining methods. The methodology involves several steps like document import R environment, document corpus handling, dataset preprocessing, bag of words representation,document term vector construction,tf-idf calculation, probabilistic latent feature discovery ,multi label content categorization and topic representation of e-learning documents.Topic representation allows the peer learners querying with the original e-learning documents. To test our research , we have selected dataset with different topics. Topic description generated by LDA model and predictive distribution LDA fit model.We have calculated documents similarity and distance between the documents. Finding an appropriate word distribution is essential in topic modelling in learning documents. In terms of understanding a huge number of documents with multi label content categorization.The predictive distribution LDA fit model can minimize the complexity in finding the new terms or topics..As a future work we are planning to modify the sampling model using predictive link distribution for enabling the link between the documents.This research work further can be extended into dynamic topic detection from corpus.

REFERENCES

1. Chong Wang and David M. Blei.Collaborative Topic Modeling for Recommending Scientific Articles.KDD '11 the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, CA, USA — August 21 - 24, 2011
2. Jason D. M. Rennie, “Improving Multi-class Text Classification with Naive Bayes ,” 2001, Massachusetts.Institute of Technology, http://citeseer.ist.psu.edu/cs
3. Yang Y., Zhang J. and Kisiel B, “A scalability analysis of classifiers in text categorization,” ACM SIGIR'03, 2003.
4. Canasai Kruengkrai , Chuleerat Jaruskulchai, “A Parallel Learning Algorithm for Text Classification,” TheEighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002), Canada, July 2002.
5. H. Wallach, “Topic modeling: beyond bag-of-words,”Proceedings of the 23rd International Conference on Machine Learning, (2006)
6. S. Deerwester, S. Dumais, G. Furnas, T. Landauer and R. Harshman, “Indexing by latent semantic analysis,” Journal of the American Society for Information Science, **41(6)**, 391–407 (1990).
7. T. Hofmann, “Probabilistic Latent Semantic Indexing,”Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 50–57 (1999).
8. T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Develop- ment in Information Retrieval, ACM, New York, NY, USA, 1999, pp. 50–57.
9. D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, The Journal of Machine Learning Research 3 (2003) 993–1022.
10. D.Cai, X. Wang, X. He, Probabilistic dyadic data analysis with local and global consistency, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, New York, NY, USA, 2009.
11. M. Girolami and A. Kaban, “On an equivalence between PLSI and LDA,” Proceedings of the 26th annual international ACMSIGIR conference on Research and development in information retrieval, 433–434 (2003).
12. J. Ponte and W. Croft, “A Language Modeling Approach to Information Retrieval,” Proceedings of the21st annual international ACM SIGIR conference onResearch and development in information retrieval 275–281 (1998).
13. T. Griffiths and M. Steyvers, “Finding scientific topics,” Proceedings of the National Academy of Sciences, **101** 5228–5235 (2004).
14. J. Lasserre, C. Bishop, T. Minka, Principled hybrids of generative and discriminative models, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, 2006.
15. Bosch, A. Zisserman, X. Munoz, Scene classification using a hybrid generative/discriminative approach, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2008) 712.
16. Jordan, M., Ghahramani, Z., Jaakkola, T. and Saul, L. (1999). Introduction to variational methods for graphical models. Machine Learning 37 183–233.
17. Wainwright, M. and Jordan, M. (2003). Graphical models, exponential families, and variational inference. Technical Report 649, Dept. Statistics, U.C. Berkeley.
18. Feinerer. An introduction to text mining in R. R News, 8(2):19{22, Oct. 2008. URL http://CRAN.R-project.org/doc/Rnews/.
19. Feinerer, K. Hornik, and D. Meyer. Text mining infrastructure in R. Journal of Statistical Software, 25(5):1{54, March 2008. ISSN 1548-7660. URL http://www.jstatsoft.org/v25/i05.
20. Feinerer I (2011). tm: Text Mining Package. R package version 0.5-5., URL http://CRAN. R-project.org/package=tm.
21. R Development Core Team (2011). R: A Language and Environment for Statistical Computing.R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/
22. Chang J (2010). lda: Collapsed Gibbs Sampling Methods for Topic Models. R package version 1.2.3, URL http://CRAN.R-project.org/package=lda
23. Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). Information disusion through blogspace. In: Proceedings of international conference on world wide web (pp. 491–501).
24. Morinaga, S., & Yamanishi, K. (2004). Tracking dynamics of topic trends using a finite mixture model. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, Seattle, USA, (pp. 811–816).
25. Mei, Q., Ling, X., Wondra, M., Su, H., & Zhai, C.X. (2007). Topic sentiment mixture: Modeling facets and opinions in weblogs. In: Proceedings of international conference on world wide web
26. Zeng, J. P., Zhang, S. Y., Wu, C. R., & Ji, X. W. (2009). Modelling the topic propagation over the internet. Mathematical and Computer Modelling of Dynamical Systems., 15(1), 83–93.
27. Salton, G., & McGill, M. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.
28. Nigam, K., McCallum, A., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. Machine Learning, 2/3, 103–134.
29. Morinaga, S., & Yamanishi, K. (2004). Tracking dynamics of topic trends using a finite mixture model. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discov ery and data mining, Seattle, USA, (pp. 811–816).

30. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In Proceedings of the Conference on Uncertainty in Artificial Intelligence, volume 20, pages 487–494, 2004.
31. Li and A. McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. In Proceedings of the International Conference on Machine Learning, volume 23, pages 577–584,New York, NY, 2006. ACM.
32. Chemudugunta, P. Smyth, and M. Steyvers. Modeling general and specific aspects of documents with a probabilistic topic model. In Advances in Neural Information Processing Systems 19, pages 241–248. MIT Press, Cambridge, MA, 2007.
33. Yanwei Wang;Tsinghua,Xiaoqing Ding;Changsong Liu.Topic Model Adaption for Recogniotion of Homologous offline Handwritten Chinese Text image,IEEE transactions on Volume: 19, Issue:12, 2014.
34. Jia Zeng;Cheung. W.K;Jiming Liu, Learning Topic Models by Belief Probagation, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 35,Issue:5,2013.