# Enhanced Biomedical data modeling using unsupervised probabilistic machine learning technique

**Syed Rizwana, Kamala Challa, Shaik Rafi, S.Sagar Imambi,**

***Abstract--- Text mining approaches uses feature similarity techniques or distributed keyword searching techniques. But machine learning techniques develop a statistical model to categorize documents by learning from vast amount of medical documents available at pubmed. It is unsupervised techniques. The proposed algorithm enhances the traditional document clustering techniques.and generate accurate and reliable model. We experimented the algorithm with 1000 document data set It showed the significant improvement over other traditional algorithms.***

***Keywords: Machine learning algorithms,LDA, unsupervised probabilistic model***

## 1. INTRODUCTION

In biomedical field, volumes of information is available and it became difficult to get information on a particular topic. It is impractical to cover all the data and get insight of it. In the last decade text mining, an interdisciplinary field that uses techniques from data mining and natural language processing became very popular to deal with unstructured data. These techniques extract key words from unstructured data and represent them in bag of words or vector format, before applying any analytical algorithms. But the key words may not represent the real meaning i.e conceptual meaning of the documents. There are many research instances for text mining based on keywords[2], but very few are available on semantic concepts of words. Citation relations are identified by Song and Kim [4] who collected full-text articles from PubMed Central. They discovered the knowledge structure and understand the importance of the bioinformatics field. They used MESH indexing as keywords.

Classification and clustering are the two most biomedical data analysis techniques. Traditional document clustering technique adopts unsupervised learning algorithm for learning clustering model. Bag of words was a common way to represent the features space of documents[6]. Every document is represented by weighted features calculated by various feature weighting techniques. The feature weighting and selection methods are mostly based on similarities of features[3]. But these methods do not represent the conceptual meaning of document. As a result, the two words with same meaning may be considered as two different features, which intern deteriorate the performance of learning model. Machine learning and statistical methods are offering solutions to these problems. In this paper we are proposing the unsupervised learning technique, which internally identify semantic distribution of words in documents. Instead of similarities between the words of document, probabilistic distribution of words according to the biomedical topics is measured in our algorithm.

## 2. SURVEY OF LITERATURE

*2.1 Topic modeling:*

A topic model is a probabilistic generative model in the area of computer science with center of attention on textual content mining and informative retrieval. In addition to text mining it has positive purpose within the fields of system vision, population genetics and social networks. It acts greater than clustering process. Topic model is starts with latent semantic indexing (LSI). Established on LSI, probabilistic latent semantic analysis [2] is a authentic topic model. Latent Dirichelt allocation (LDA) [1], is more complete probabilistic generative model,extension of PLSA. These models are used in text evaluation for identifying unsupervised topics in a corpus of document. Topic models are introduced in to the fields of biological/biomedical text mining and clinical informatics since their superiority in analysing of big scale datasets. Extracting hidden knowledge and relations from the biological microarray datasets become a great challenge now a days. Topic model can model a biological model in terms of hidden “Topics”, which can identify the underlying biological meaning more systematically.

To address the challenge of BOW, topic models are able to identify and group words based on the semantic meaning of words. Instead of representing features space with words from documents, this model projects each document into topic space. It maps semantically equal terms in to same feature space. It decreases the noise of similarity measure and dimensions of feature space[3]. When clustering of document is based on semantic meaning it outperforms the text mining techniques. Topic modeling is not only identify semantic meaning of documents, but it can be applied to image modeling[6]

**Revised Manuscript Received on March 10, 2019.**

**Syed Rizwana,** Asst.Professor CSE Department, Narasaraopeta Engineering College, Narasaraopet, Guntur(Dt), Andhra Pradesh, India (rizwana.nec@gmail.com)

**Kamala Challa,** Asst.Professor CSE Department, Narasaraopeta Engineering College, Narasaraopet, Guntur(Dt), Andhra Pradesh, India (kamalachalla@gmail.com)

**Shaik Rafi,** Asst.Professor CSE Department, Eswar College Of Engineering,Kesanupalli Narasaraopet,Guntur(Dt),Andhra Pradesh, India (simambi@gmail.com)

**S.Sagar Imambi,** Associate Professor,CSE Department, KLUNIVERSITY,Vaddeswaram, Andhra Pradesh, India (shaikrafi123@yahoo.com)

*2.2 LDA MODELING:*

The data modeling is dependent on the quality of input data. Here comes LDA modeling, where categorized feature vectors are used as input instead of directly extracted features from documents. Song M et al. proposed system which automatically execute topic analysis on the entire document., when categorized feature vectors are given as input. [5].

LDA can also practiced for feature learning method and to represent the documents from the corpus. Chen, Sih-Huei, et al proposed an early-warning system to detect the crime activity intention using latent Dirichlet allocation (LDA). Their collaborative representation classifier (CRC) used to find related topic for a given document [10].

A Location-aware Topic Model (LTM) is proposed to (i) mine the common features of songs that are suitable for a venue type in a latent semantic space and (ii) represent songs and venue types in the shared latent space, in which songs and venue types can be directly matched. It is worth mentioning that to discover meaningful latent topics with the LTM, a Music Concept Sequence Generation (MCSG) scheme is designed to extract effective semantic representations for songs. An extensive experimental study based on two large music test collections demonstrates the effectiveness of the proposed topic model and MCSG scheme.[11]

## 3. BIOMEDICAL DATA MODELING

Biomedical data modeling process include data collection, preprocessing, and key phrase extraction to feed input into topic models. Topic modeling is an unstructured data analysis method, where each term in document chooses a topic assignment and the term from corresponding topic. The main assumptions of topic modeling are

- M documents in the repository are assumed to be related to one or more topics from the set of K topics.
- By using Gibbs's principle choosing a topic randomly form the distribution over the topic(P(t|d)), randomly choosing a term from the distributed over the vocabulary (P(w|t)).

The terms used in this algorithm are
*d denotes label of document,*
*t represents topic,*
*w represent a word and*
*N is the number of words in the document d.*
*Extract-Tpocis()*

*BEGIN*
for each document dϵ{1,…..,M}:
Randomly assign topics to words in the document
For each term w in document d
- Compute *p(t|d)* – proportion of words in document d which are assigned to topic t.
- Compute *p(w|t)* - ratio of words assignments to topic t

Reassign the term w a new topic t' by choosing argmax(*P(t'|d) *P(w|t').*
Return the probability of topic t' generated term w.

*END*

A graphical model can also reflect the generative process of documents, as show in below figure 1.

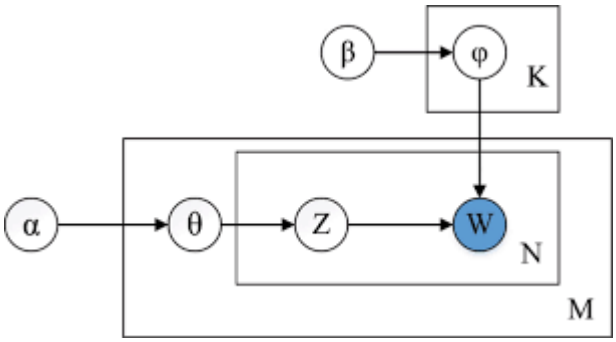


**Fig 1. LDA model**

## 4. METHODOLOGY OF BIOMEDICAL DOCUMENT CLUSTERING

Assume that a set of M clusters with probabilistic distribution function ($pd_1$,$pd_2$…. $pd_m$) and their probabilities are ($p_1$,$p_2$…. $p_m$)

Probability of a topic t generated by cluster $C_i$ is $P(t|C_j)=p_i * pd_i(t)$

Probability of topic t generated by set of clusters C is $P(t|C) = \sum_{i=1}^{k} p_i pd_i\,(t)$

As topics are assumed to be generated independently for a data set D=($t_1$,$t_2$,…$t_n$) then

$P(D|C) = \prod_{i=1}^{n} P(t_i|\,C) = \prod_{i=1}^{n} \sum_{j=1}^{m} p_j pd_j$

The main step is to find a set of m probabilistic clusters such that P(D|C) is maximized. We may assume initial the probability distribution as normal.

## 5. EXPERIMENTAL RESULT

*5.1 Experimental procedure :*

1)Extracting documents:

To generate the topics from the data set we, choose the documents from pubmed i.e 2010 to 2018 by giving query string as "health risks factors". We selected first 1000 documents and the topics are "cardio","diabetic", "neurology", "pregnancy".

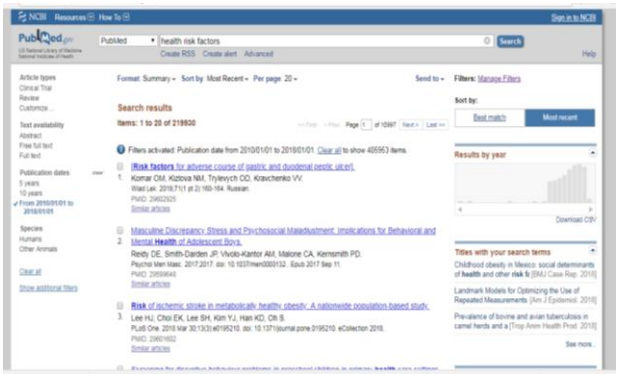

**Fig 2. Searching for biomedical documents in PubMED.**

2) Document modeling: We conducted experiment using the above algorithm to generate the topics from the 1000 document set. The Documents retrieved from pubmed are shown in fig 3.

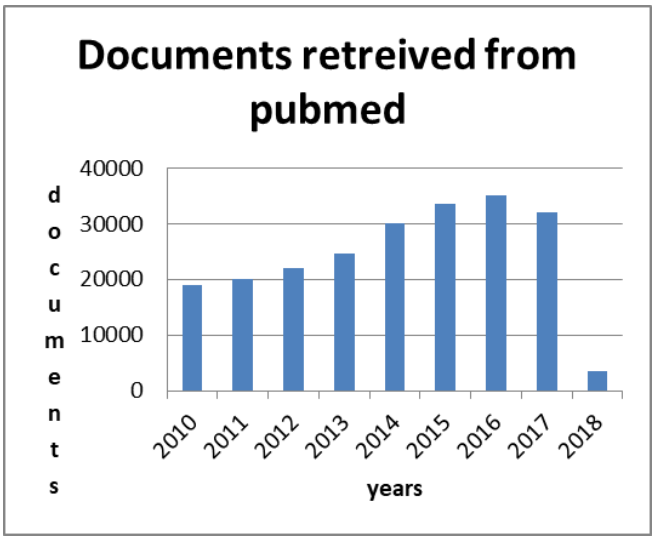


**Fig 3. Retreived Pubmed Doucments vs Year**

The distribution of the terms in document set is represented in the above graph. The top terms retrieved from the documents are listed in the table

| | | | |
|---|---|---|---|
| • Pregnency | • Diabetic terms | • Cardiology | • Neurology |
| • Family history. | • Endocrine gland | • posterior cerebral artery | • APNEA |
| • Blighted ovum | • Edema | • rubral tremor | • ANTI-COAGULANT |
| • Premature | • dilated eye exam | • slow wave sleep | • ATAXIA |
| • Race. | • adhesive capsulitis | • sleep paralysis | • BRADYKINESIA |
| • Blood clot | • ACE inhibitor | • stria medullaris | • CATHETER |
| • Smoking. | • Anemia | • vestibulo-ocular response | • CT scan |
| • High cholesterol. | • fasting blood glucose test | • uncus | • CEREBELLUM |
| • High blood pressure. | • glycemic index | • vacuolar myelopathy | • GALACTORRHE |
| • Sedentary lifestyle | • hypoglycemia | • transient ischemic attack | • CRANIECTOMY |
| • Placenta | • insulin | • trochlear nerve | • DIPLOPIA |
| • Thyroid | • gangrene | • vascular | • EPENDYMA |
| • Fetus | • | • Sedentary lifestyle | • Sedentary lifestyle |

**Table 5.1 Top words**

*5.2 Evaluation Measure of algorithm.*

This paper measures text similarity and clustering effect with a clustering analysis of text, adopting F Metric, Precision Ratio and Recall ratio. F Metric is a balance index for information retrieval combining Precision Ratio and Recall ratio. Test results prove that the similarity computing method proposed in this paper is feasible. Table5.2 shows the precision recall

| category | precision | Recall | Accuracy |
|---|---|---|---|
| cardio | 0.824 | 0.7285 | 85.8% |
| diabetic | 0.812 | 0.7125 | 91.2% |
| Neurology | 0.785 | 0.7245 | 89.8% |
| Pregnancy | 0.867 | 0.745 | 87.8% |

**Table 5.2 Result**

## 6. CONCLUSION

In this paper, we proposed a probability based approach to categorize the 1000 medical documents retrieved from pubmed. This algorithm utilizes the probability distribution of topics across the documents. LDA model also identifies the frequently used terms in the documents specific to the categories like Diabetic, Neurology, Pregnancy an cardio.

## REFERENCES

1. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research*3.Jan (2003): 993-1022.
2. Hofmann, Thomas. "Unsupervised learning by probabilistic latent semantic analysis." *Machine learning* 42.1-2 (2001): 177-196.
3. Abualigah, Laith Mohammad, et al. "Text feature selection with a robust weight scheme and dynamic dimension

reduction to text document clustering." *Expert Systems with Applications*84 (2017): 24-36.

4. Song M, Kim SY. Detecting the knowledge structure of bioinformatics by mining full-text collections. Scientometrics. 2013;96(1):183–201.
5. Park, Kiejin, and Minkoo Kang. "A Development of Automatic Topic Analysis System using Hybrid Feature Extraction based on Spark SQL." *International Journal of Applied Engineering Research* 12.16 (2017): 5472-5480.
6. Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 2, pages 524{531. IEEE, 2005.
7. Jun Zhu, Li-Jia Li, Li Fei-Fei, and Eric P Xing. Large margin learning of upstream scene understanding models. Advances in Neu
8. http://www.diabetes.org/diabetes-basics/common-terms/common-terms-s-z.html
9. https://www.webmd.com/baby/pregnancy-glossary
10. Chen, Sih-Huei, et al. "Latent dirichlet allocation based blog analysis for criminal intention detection system." *Security Technology (ICCST), 2015 International Carnahan Conference on*. IEEE, 2015.
11. Cheng, Zhiyong, and Jialie Shen. "On effective location-aware music recommendation." *ACM Transactions on Information Systems (TOIS)* 34.2 (2016