Volume 3, Issue 6, June 2018 ISSN (online): 2456-0006

International Journal of Science Technology Management and Research

Available online at: www.ijstmr.com

Topic Modeling and Recommendation Generation Using Bag-of-Discriminative-Words (BoDW)

Prashant C. Shah
Computer Engineering Department
SVIT COE
Nashik,India

Prof. K. N. Shedge
Computer Engineering Department
SVIT COE
Nashik,India

Abstract : In the topic analysis Bag of word is important technique. The important words present in a document, either represents a fact or sentiment related to the topic. The fact represents the objective labels and sentiment represents the subjective labels. The meaning of word varies with respect to the topic. Every word in a document does not have equal degree to represent a topic. The subjective and/or objective degree of a discriminative word varies with respect to the topic. The system extracts bag of discriminative words from a document based on objective and subjective selection of variables. The LDA and regression techniques are used to filter the discriminative words in a document. Based on objective and subjective analysis recommendations are generated by analyzing subjective score with respect to the user query.

Keywords: Bag-of-Discriminative-Words, Latent Dirichet Allocation, Objective and Subjective Classification, Recommendation, Topic Modelling.

I. INTRODUCTION

In data mining, information retrieval is a key task. Bag of words technique is used to extract information form a large text documents contents. The words representing the text document are the frequent words exist in a document. This is a orderless collection of different words This bag of words technique is used in document classification technique.

In information retrieval, topic modeling is again an emerging branch. It extracts hidden semantic structure in a text document. In topic modeling the one or more topics present in a document are extracted. The topic modeling technique uses bag of words technique to model topic specific words.

The two most important technique in topic modeling are : probabilistic latent semantic analysis PLSA and latent Dirichlet allocation-LDA. The PLSA technique extracts the hidden semantics from data using different words. This technique projects the documents in low dimensional space with the help of words representing latent topic. The every word has multinomial distribution over a fixed vocabulary. It uses probabilistic generative process. The LDA technique inherits the properties of PLSA and applies an extra generative process. LDA defines the topic proportion covered by each document.LDA has ability to process multiple documents.

The words present in a document are either represents a fact or the sentiment related to the fact. Fact represents the objective and sentiment represents the subjective label of a document. The word has different meaning in different context of topic for example the word :order Hemiptera represents a objective information in biology. The order Hemiptera is a kind of insects,. But in case of software domain the same word has a meaning bug and represents a subjective description. The subjective description contains the negative feedback of objective software.

Better analysis of a document can be performed with the help of subjective and objective discriminative power of words. The system uses discriminatively objective-subjective LDA to generates the bag-of-discriminative words BoDW. With the help of objective and subjective representation of a document recommendations are generated for user with respect to the query.

.

II. LITERATURE SURVEY

The supervised LDA[2] model is the traditional topic modeling technique. It is a supervised extension of tradition latent dirichlet allocation. SLDA model can be merged with generalized linear model [3] to generate various types of results such as multiclass label generation, real, discrete and non-negative values for labels. Using generalized linear model a multiclass SLDA model is proposed.

TLDA[4] is another technique based on the SLDA technique used to bridge the gap between language present in multiple document. TLDA model uses a binary notation for a word to represent the word is technical or not. This technique uses cosine regression model. To identify the subjective and objective sense of a document Mei et al. [5] proposes a new technique to model multiple topics in documents along with the sentiments. This technique is called as topic sentiment mixture.

I. Titov and R. McDonald[6] proposes a new technique that finds the sentiment words related to each topic. This technique is called as Multi-grain LDA. The techniques[5][6] are unsupervised techniques and these techniques represents topic as well as sentiment in a document. These models are not predictive like traditional SLDA model.
The joint sentiment topic -JSTmodel and revised version of JST[7] : R-JST[8] are two techniques designed to identify sentiments in a given topic. These techniques extract the sentimental polarities for every word in a document.

All the SLDA based technique and SLDA model are used for classification and regression. But SLDA technique is not capable to identify both object and sentiment identification at a time. The LDA model with visual feature extraction technique is also implemented for multimedia data analysis. Gaussian multinomial LDA (GM-LDA) technique[9] is used to analyze images and its annotations. Each iamge is a collection of multiple topics. dosLDA[1] technique is used to extracts subjective and objective labels from multiple topic documents. This technique uses LDA as well as regression model. Microblog recommendations are generated using Latent Dirichlet Allocation topic model[10]. Microblog is a browser based platform. User uses this platform for data communication and sharing. Based on the user interest recommendations are generated to the user using LDA technique.

III. ANALYSIS AND PROBLEM FORMULATION

For a collection of documents, retrieve a vocabulary of size V. Extract one objective label for each document and extract one subjective label. The subjective label value is binary i.e. positive or negative. Based on the objective information and subjective analysis of collection of a documents a recommendation need to be generated based on the user preferences.

IV. SYSTEM OVERVIEW

The system is divided in 5 modules:

A. Define Constants

For execution of latent dirichlet allocation and Bernoulli distribution the constants are required. These constants are evaluated using expectation-maximization. Expectation-maximization EM technique is a statistical technique. It is used to find parameter values from an equation which can not be solved directly. This technique find maximum likelihood. This is an iterative process where expectation and maximization are calculated in each iteration.

B. Latent Dirichlet Allocation-LDA

Latent Dirichlet allocation model find the topic proportion of a document by analyzing the document content. The input dataset contains mixture of various topics. A single document can cover one or more topics. The LDA technique assigns one or more topic to the document by analyzing its content.

C. Multinominal Distribution

In Multinomial distribution topic distribution in a single document and word distribution among multiple topics are evaluated. The Multinomial distribution provides the probability of word belong to the topic and also the topic probability of a document.

D. Regression

The objective labels and subjective labels are extracted from high probable words found in Multinomial distribution process. To check whether the word is discriminative or not, a regression process is used. A binary selector model is used to assign words to identify objective labels and subjective labels.

E. Recommendation

Based On the objective label and its subjective analysis, the recommendation result can be generated. User fire a query related to specific objective i.e. product. Based on the average total subjective score the products are recommended to the user in a sorted manner.
Following fig. 1 shows the system architecture

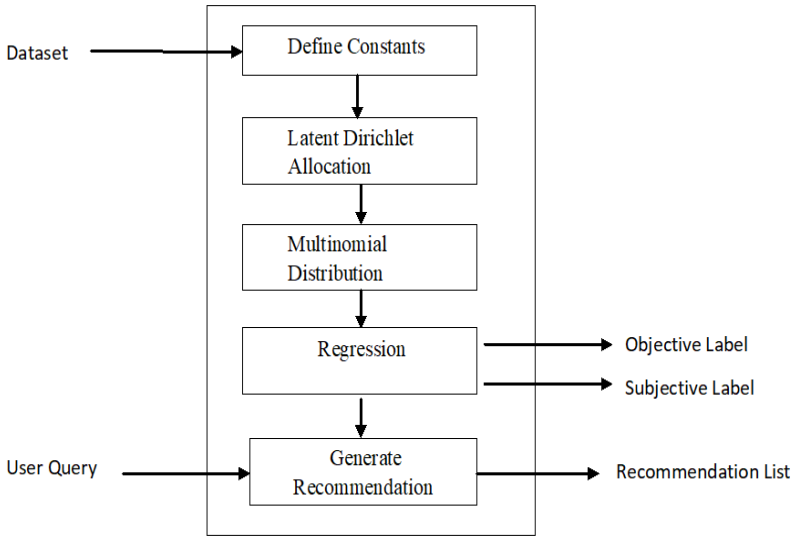

Fig. 1. System Architure

A. Algorithms

dosLDA Algorithm
Input: D: Dataset
Output:OL: Objective label
SL: Subjective Label
Processing::
1) Calculate bernoull and dirichlet constants using expectation-maximization
2) Calculate topic proportions using latent dirichlet allocation
3) sample each topic assignment using multinomial distribution among topics
4) sample each word using using multinomial distribution among words
5) sample each binary selector in terms of objective power using bernoulli distribution
6) sample each binary selector in terms of subjective power using bernoulli distribution
7) OL : draw objective label using logistic regression
8) SL:draw subjective label using logistic regression.

Rec-DOS-LDA Algorithm
Input: User Query
Output: Recommended product list
Processing:
1) Apply dosLDA Algorithm
2) Match objective label with user query
3) Match subjective label with user query
4) Calculate Positive and negative product score with matched label score
5) Calcualte average score
6) Arrange products in descending order
7) Generate recommendation list

B. Mathematical Model

System S can be defined as
S= fI, O,Fg Where,
I = fI1, I2g, Set of inputs
I1= Text dataset
I2 = User Query

O= fO1,O2,O3,O4g Set of outputs
O1= Objective Label
O2= Subjective Label

O3 = Word Vocabulary
O4 = Recommendation List

F =fF1,F2,F3,F4,F5,F6,F7,f8,f9g Set of Functions
F1 = Data preprocessing
F2 = expectation-maximization
F3 = latent dirichlet allocation
F4 = multinomial distribution among topics
F5 = multinomial distribution among words
F6 = bernoulli distribution
F7 = Objective label identification
F8 = Subjective label identification
F9 = recommendation Generation

V. IMPLEMENTATION

A. Experimental Setup:

The system is implemented in java using jdk.17 on windows system with 8 gb ram and i5 processor.

B. Dataset:

Text dataset[11]: The text dataset is a multi-domain text dataset along with the sentiment representation. This dataset contains reviews of multiple products present on amazon.com

C. Performance Metric:

- Accuracy of recommendation: Based on the user query the generated recommendations are analyzed using precision and recall.
- Time: The time of execution is evaluated for topic modeling and recommendation generation.

D. Implementation Status:

The Multi domain Text dataset contains 4 type of products: dvd, book, electronics, kitchen and house wares . The review of each type is mentioned in xml document. The dataset is prelabeled dataset with positive and negative labels. The labels are not used for processing. These labels are useful to evaluate the system accuracy. For data preprocessing, XML is parsed and reviews text are extracted. The extracted text is then processed using stemmer and stopword algorithm. The word frequency is calculated and words with least occuance i.e. exist less than 10 times are removed from the data. The remaining words are treated as a vocabulary.
Following table I shows the Processing time for each type of product reviews.

Sr. No.	Dataset	No. of Reviews	Vocabulary Generation Time(in Minute)
1	Book	2000	4.32
2	DVD	2000	4.91
3	Electronics	2000	5.03
4	Kitchen And house wares	2000	4.81

CONCLUSION

The proposed system uses supervised topic model dos-LDA to analyze text documents. After analysis, system extracts discriminative bag of words having subjective and objective sense. The system generates one objective and one subjective label for each document. The document analysis result helps to generate recommendation as per the user preferences. The subjective analysis of a document is used to generate multiple recommendations to the user. In future system can be extended for multimedia content analysis using visual features and text annotations using dos- LDA.

ACKNOWLEDGMENT

First of all my special thanks to head of Department of Computer Engineering, SVIT, Chincholi, Nashik . Prof. K. N. Shedge, principal Dr. S. N. Shelke for their kind support and suggestions. It would not have been possible without the kind support. We would like to extend our sincere thanks to all the faculty members the department of Computer Engineering for their help. We are also thankful to colleagues for moral support and encouragement. At the end, We are very much thankful to all for direct and indirect help.

REFERENCE

1. Yueting Zhuang, Hanqi Wang, Jun Xiao, Fei Wu, Yi Yang, Weiming Lu, and Zhongfei Zhang, ”Bag-of-Discriminative-Words (BoDW) Representation via Topic Modeling”, IEEE Transactions on Knowledge and Data Engineering, Vol. 29, Issue: 5, pp. 977 - 990 May 2017.
2. J. D. Mcauliffe and D. M. Blei, ”Supervised topic models,” in Advances in Neural Information Processing Systems 20, 2008, pp. 121-128
3. P. McCullagh, ”Generalized linear models,” European Journal of Operational Research, vol. 16, no. 3, pp. 285-292, 1984.
4. S. Yang, S. P. Crain, and H. Zha, ”Bridging the language gap: Topic adaptation for documents with different technicality,” in Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, 2011, pp. 823-83.
5. Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, ”Topic sentiment mixture: Modeling facets and opinions in weblogs,” in Proceedings of the 16th International Conference on World Wide Web, 2007, pp. 171- 180.
6. I. Titov and R. McDonald, ”Modeling online reviews with multigrain topic models,” in Proceedings of the 17th International Conference on World Wide Web, 2008, pp. 111-120.
7. C. Lin and Y. He, ”Joint sentiment/topic model for sentiment analysis,” in Proceedings of the 18th ACM Conference on Information and Knowledge Management, 2009, pp. 375-384.
8. C. Lin, Y. He, R. Everson, and S. Ruger, ”Weakly supervised joint sentiment-topic detection from text,” IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 6, pp. 1134-1145, 2012.
9. Jianyong Duan, Yamin Ai and Xia li, ”LDA topic model for microblog recommendation”, Asian Language Processing (IALP), 2015 International Conference on April 2016
10. D. M. Blei and M. I. Jordan, ”Modeling annotated data,” in Proceedings of the 26th Annual International ACM SIGIR Conferenceon Research and Development in Informaion Retrieval, 2003, pp. 127-134
11. J. Blitzer, M. Dredze, and F. Pereira, ”Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification,” in Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007, pp. 440-447.