A Survey on Topic Modeling in Text Mining

A.Saranya[1] V.Vennila[2] S.Suganya[3] G.S.Rizwana Banu[4]
[1,2,3,4]Department of Computer Science Engineering
[1,2,3,4]K.S.R College of Engineering, Tamilnadu, India

Abstract— Text Mining has become an essential research area. Text Mining is the innovation by computer of new, earlier unknown information, by spontaneously extracting information from dissimilar written resources. In this paper, a Survey of TextMining technique for Selective supervised Latent Dirichlet Allocation have been presented. Although ssLDA receives the universal framework of sLDA where many forms of response (such as real and categorical responses) variables can be shown, we focus on the case where the response variable is definite in this paper. In this paper, according to the following two considerations: First, most of other topic models, ssLDA views documents as discrete data which consist of word counts, while documents are treated as directional data in STM. Second, the topics in ssLDA and STM are generated from Dirichlet and vMF distributions, respectively. Due to Dirichlet distributions, ssLDA can set a topic assignment (topic indicator) and adjust the weight for each individual word that STM fails to do. This paper explores existing research highlights and provides various needs of significant research in these topics.

Key words: Data Mining, Text Mining, Topic Modeling, Document Relevance, Information Filtering, Latent Dirichlet Allocation, Supervised Learning

I. INTRODUCTION

Data Mining, also commonly known as Knowledge Discovery in Databases (KDD), discusses to the nontrivial extraction of implied, previously unknown and hypothetically useful material from data in databases. Though data mining and knowledge discovery in databases are usually treated as substitutes, data mining is essentially part of the knowledge discovery process. Data mining originates its name from the comparisons between penetrating for valuable information in a large database and mining rocks for a manner of valuable ore. Both suggest any selecting through a large amount of material or inventively searching the material to exactly pinpoint where the values reside. Text mining is related to data mining, excluding that data mining tools are considered to handle structured data from databases , but text mining can effort with unstructured or semi-structured data sets such as emails, full-text documents and HTML files etc. In opic Modelling consider Natural Language Processing (NLP) and Machine Learning (ML), a focus to prototypical is a type of statistical model for determining the intellectual "topics" that arise in a collection of documents. Spontaneously, given that a document is about a specific topic, one would assume particular words to seem in the document more or less frequently. A topic model detentions this intuition in a mathematical framework, which allows tentative a set of documents and determining, based on the data of the words in each, what the topics capacity be and what each document's balance of topics is.

II. LITERATURE SURVEY

A. Probabilistic Word Selection via Topic Modeling

According to Yueting Zhuang, Haidong Gao, Fei Wu, Siliang Tang, Yin Zhang, and Zhongfei Zhang [1], The system achieves document classification using Bernoulli distribution based word selection with discriminatory property values. We propose Selective supervised Latent Dirichlet Allocation (ssLDA) is used to boost the forecast performance of the supervised probabilistic topic models. The Bernoulli distribution is parameterized by the discrimination power of the word for its allotted topic. As a result, the document is signified as a "bag-of-selective-words" instead of the probabilistic "bag-of-topics" in the topic modeling province or the flat "bag-of-words" in the traditional natural language processing domain to form a new perspective. The system supports word selection and classification on single labeled documents. Probabilistic topic models, which use probabilistic generative approaches to discover the latent semantics embedded in the documents, have recently attracted more and more attention in modeling and analyzing textual or image documents. The traditional (unsupervised) LDA is a Bayesian multinomial mixture model, where topics are mixture components and topic proportions are mixing proportions. As discussed before, ssLDA selects each word as a strongly or weakly discriminative one in a document by a per-word binary selection variable (with a discriminative value). ssLDA is expert of acquiring the analytical representation of a document taking the latent semantics fundamental the words as well as the words themselves into the consideration. As a result, ssLDA gains stable and superior classification performance independent of the number of topics. The main drawback is Demanding word discriminatory power estimation mechanism.

B. Parsimonious Topic Models with Salient Word Discovery

According to Hossein Soleimani and David J. Miller [2], The system constructs the topic models with topic related words collection fetched from the documents. We suggest a parsimonious topic model for text corpora. Topic specific probabilities and Latent Dirichlet Allocation (LDA) methods are used to fetch words with relevant topic models. In Latent Dirichlet Allocation (LDA), all words are modeled topic specifically, even though many words occur with related frequencies across different topics. In LDA all topics are in principle present in every document. Our model gives sparse topic representation, determining the subset of relevant topics for each document. By using an Bayesian Information Criterion (BIC), balancing model difficulty and goodness of fit. Compared to previous works, our main contributions are:

All rights reserved by *www.ijsrd.com*

- We achieve sparsity in topic proportions and in topic-specific words. Prior works at best achieve sparsity in one of these two senses.
- Our model allows the subset of salient words to be topic-specific. This follows the premise that some words may have mutual frequency of existence under some separation of the topics
- We originate a novel BIC objective function, used for learning our model.

Experiments show that our model outperforms LDA and a sparsity-based topic model with respect to several clustering performance measures, including test set log-likelihood and agreement with ground-truth class labels. The main disadvantage is Topic models are not connected with document analysis applications.

C. TASC:Topic-Adaptive Sentiment Classification on Dynamic Tweets

According to Shenghua Liu, Xueqi Cheng, Fuxin Li, and Fangtao Li [3], The system performs sentiment classification on dynamic Tweets. A topic-sensitive task have an sentiment classification. A classifier trained from one topic will achieve not as good as on another. This is particularly a problem for the tweets sentiment analysis. Semi-supervised Topic-adaptive Sentiment Classification (TASC) model performs the sentiment classification with a classifier build on many features and varied labeled data. It minimizes the pivot loss to adapt to unlabeled data and features. Text and non-text features are extracted and naturally split into two opinions for co-training. The TASC learning algorithm apprises topic-adaptive structures based on the collaborative selection of unlabeled data, which in try helps to choice more consistent tweets to boost the performance.Therefore our work focuses on cross-domain sentiment analysis on tweets, and we propose a semi-supervised topic-adaptive sentiment classification model (TASC). It transfers an initial common sentiment classifier to a specific one on an emerging topic. TASC has three key components.

- The semi-supervised multiclass SVM model is formalized. Given a small amount of mixed labelled data from topics, it selects unlabeled tweets in the target topic, and minimizes the structural risk of labeled and selected data to adjust the sentiment classifier to the unlabeled data in a transductive manner
- We set feature vector in the model into two parts: fixed common feature values and topic-adaptive feature variables
- To tackle the content sparsity of tweets, more features are extracted, and split into two views: text and non-text features.

Sentiment classifications on tweets suffer from the problems of lack of adapting to unpredictable topics and labeled data, and extremely sparse text. The main disadvantage is Topic patterns are not considered in the system.

D. Automatic Taxonomy Construction from Keywords via Scalable Bayesian Rose Trees

According to Yangqiu Song, Shixia Liu, Xueqing Liu, and Haixun Wang [4], The system constructs domain related taxonomy with keyword relationship values. We learn a challenging problem of deriving a taxonomy from a usual of keyword phrases. A solution can benefit many real-world applications. However, it is impossible to create a taxonomy out of a keyword set itself. Bayesian rose tree algorithm is used to build a hierarchical taxonomy for a agreed set of keywords. We decrease the complexity of earlier hierarchical clustering approaches from Oðn2 log nÞ to Oðn log nÞ using a nearest-neighbor-based approximation. In this paper, we consider the challenging problem of inducing a taxonomy since a set of keyword phrases instead of after a text quantity. The problem of inducing a taxonomy from a set of keywords has one major challenge. Although using a set of keywords allows us to more accurately characterize a extremely focused, even fast-altering domain, the set of keywords itself does not comprise obvious relationships from which a taxonomy can be constructed.

The assumption is that the text corpus exactly denotes the domain. Although these text-corpus based methods have attained some success, they have several disadvantages.

- It is very difficult to find a text corpus that accurately characterizes that domain.
- If we can find a corpus that accurately characterizes the domain, we may still have a data sparsity problem. High quality patterns typically have very low recall.

Our approach can be regarded as specifying the general-purpose knowledge (or world knowledge) of taxonomy to a domain specified task by three steps:

- World knowledge formulation
- Concept deduction for domain keywords
- Hierarchical cluster induction.

The main drawback is Conceptual relationships are not considered.

E. Exploring Topical Lead-Lag across Corpora

According to Shixia Liu, Yang Chen, Hao Wei, Jing Yang, Kun Zhou, and Steven M. Drucker [5], The system performs lead-lag relationship identification between the text collections. Classifying which text quantity leads in the situation of a topic presents a great experiment of significant interest to researchers. Real-world applications have a terrible need to appreciate lead-lag patterns both globally and locally. We introduce TextPioneer, an cooperating visual analytics tool for exploring lead-lag across extents from the global level to the resident level. The major contributions of this work are:

- An collaborating visual analytics tool that tightly mixes interactive visualization with lead-lag analysis to help users to improved understand lead-lag relationships crossways corpora both globally and locally.
- A two-phase analysis mechanism that impeccably accomplishes lead-lag investigation at both the globaland local levels.
- Coherent visualization appliance that encrypts the two-level analysis results.

The major article of this tool is that it allows users to visually analyze topical leadlag relationships both locally and globally. In this situation, mining can consider terms that defines the concept of the sentence which results

All rights reserved by www.ijsrd.com

creation of topic TextPioneer provides three significant benefits over previous methods. Topic influence estimation is not performed.

F. Autocratic Decision Making using Group Recommendations based on Intervals of Linguistic Terms and Likelihood-Based Comparison Relations

According to Shyi-Ming Chen, Fellow, and Bing-Han Tsai [6], Linguistic Terms interval and Likelihood based group recommendation is identified by the system. The future method builds a mutual interval semantic preference matrix and uses likelihood-based comparison relations of intervals of linguistic terms to build a mutual preference matrix for all professionals. The proposed method can overcome the drawbacks of Chen and Lee's method and Ben-Arieh and Chen's method for autocratic decision creating using cluster approvals. Autocratic decision making scheme is applied for the group recommendation process. Firstly, The proposed method collections the interval linguistic preference matrix of each proficient to build a collective interval linguistic preference matrix. Secondly, it uses likelihood-based comparison relations of intervals of semantic terms to build a preference matrix for each expert and to build a mutual preference matrix of all experts, respectively. Based on the gained collective preference matrix, it analyses the score of each alternative. Thirdly, it forms a agreement matrix for each proficient and calculates the consensus degree for each expert. Finally, it calculates the group consensus degree of all experts. The main disadvantage is Ordered weighted average (OWA) based similarity analysis is not performed.

G. GFilter: A General Gram Filter for String Similarity Search

According to haoji hu, kai zheng, xiaoling wang, and aoying zhou [7], The system performs string search on biomedical data collections. Several submissions such as data integration, protein detection, and article copy detection share a parallel essential problem. Assumed a string as the query, how to efficiently find all the similar answers from a huge scale sequence collection. Many prevailing methods accept a prefix-filter-based framework to explain this problem. Gram-based framework is used to achieve near maximum filter performance in string search process. The main notion is to carefully choose the high-quality grams as the prefix of query rendering to their assessed facility to filter entrant. we make the following major contributions in this paper:

- We develop a general gram filter. This simplification offers a chance to select optimized combination of grams from q-gram set of a query.
- We present a choose-and-extend framework to efficiently find the high feature grams in the query process. This framework have an different approaches can be prolonged

A theoretical analysis for this filter model to prove that it is NP-hard problem to obtain best gram filter. Greedy algorithm is not good when t is not large. We then devise an effective measure to sort grams based on their probable capability to reduce entrant size, which is used to select better grams efficiently and effectively. The main disadvantage is Topic relevance based string search is not adapted by the system.

H. Automatic Labeling of Multinomial Topic Models

According to Q. Mei, X. Shen, and C. Zhai [8], Multinomial distributions prepared words are frequently used to model topics in text collections. A communal, foremost experiment in applying all such topic models to some text mining problem is to tag a multinomial topic model exactly so that a user can construe the discovered topic. We propose probabilistic approaches to automatically labeling multinomial topic models in an impartial way. The technique we have to use minimizing Kullback-Leibler divergence between word distributions and also used to maximizing mutual information between a label and a topic model. Various different topic models have been planned to extract interesting topics in the form of multinomial distributions automatically from text. The future methods are assessed using two text data sets with different types (i.e., literature and news). The outcomes of experiments with handler study show that the future labeling approaches are relatively effective and can automatically produce labels that are important and useful for deducing the topic models. It can be applied to labeling a topic knowledgeable over all kinds of topic models such as PLSA, LDA, and their variations. This method is not only restricted to labeling topic models it can also be used in any text management tasks. The main drawback is Without reasonable labels, the use of topic models in physical world applications uis seriously limited.

I. A segment-based approach to clustering multi-topic documents

According to A. Tagarelli and G. Karypis [9], Document clustering has been familiar as a fundamental problem in text data management. It particularly challenging when document contents are categorised by subtopical negotiations that are not automatically relevant to each other. In Existing methods a document is an indivisible unit for text representation and similarity computation, it does not handle documents with multiple topics. We address the problem of leveraging the natural composition of documents in text segments that are coherent with respect to the underlying subtopics in multi-topic document clustering. We propose a new document clustering framework that is considered to make a document association from the identification of cohesive groups of segment-based portions of the unique documents. Evolving a document clustering approach for collections in which each document can possibly belong to multiple topics. In Existing clustering multi-topic document groups discourse the problem by making overlapping clustering solutions. The main characteristic is entire document as a single unit of information.

III. PROBLEM IDENTIFICATION

From the above survey papers this paper studied the following problems: Using the Selective Supervised Latent Dirichlet Allocation (ssLDA) is boost the prediction performance for the probabilistic topic model.We use Bernouli Distribution for each word to identify word as a strongly or weakly discriminative. We also introduce Gaussian Linear Model (GLM) it primarily concentrating on the word selection mechanism for single lable. We derive a Bayesian Information Criterian (BIC), focusing on model

All rights reserved by www.ijsrd.com

complexity and goodness of fit. We propose a Topic-Adaptive Sentiment Classification (TASC) model, to derive mixed labeled data from various topic. We develop Bayesian approach to build hierarchical taxonomy for set of keywords.

IV. Conclusion

This paper proposes a natural extension of sLDA, called selective supervised Latent Dirichlet Allocation. ssLDA is capable of obtaining the analytical representation of a document taking the latent semantics original the words as well as the words themselves into the consideration. Specially, words in ssLDA are taken as structures, whose numbers are consistent and do not change with the number of topics, and their topic assignments are used to adjust the weights of words. As a result, ssLDA gains constant and larger classification performance independent of the number of topics. The experiments shown on textual documents show that ssLDA not only performs competitively over "state-of-the-art" classification methods based on both the flat "bag-of-words" demonstration and the probabilistic "bag-of-topics" representation in terms of classification concert, but also has the ability to discover the discrimination power of the words specified in the topics. As we know, the suggestion of words' discrimination power in ssLDA is shown without any previous knowledge. In the future work, we tend to seek some methods to boost the inference of words' discrimination power.

References

[1] Yueting Zhuang,Haidong Gao, Fei Wu, Siliang Tang, Yin Zhang, and Zhongfei Zhang, "Probabilistic Word Selection via Topic Modeling," In Proc. IEEE Transactions On Knowledge And Data Engineering, vol. 27, no. 6, June 2015.
[2] Hossein Soleimani and David J. Miller, "Parsimonious Topic Models with Salient Word Discovery," In Proc. IEEE Transactions On Knowledge And Data Engineering, vol. 27, no. 3, March 2015.
[3] Shenghua Liu, Xueqi Cheng, Fuxin Li, and Fangtao Li, "TASC:Topic-Adaptive Sentiment Classification on Dynamic Tweets," In Proc. IEEE Transactions On Knowledge And Data Engineering, vol. 27, no. 6, June 2015.
[4] Yangqiu Song, Shixia Liu, Xueqing Liu, and Haixun Wang, "Automatic Taxonomy Construction from Keywords via Scalable Bayesian Rose Trees," In Proc. IEEE Transactions On Knowledge And Data Engineering, vol. 27, no. 7, July 2015.
[5] Shixia Liu, Yang Chen, Hao Wei, Jing Yang, Kun Zhou, and Steven M. Drucker, "Exploring Topical Lead-Lag Across Corpora," In Proc. IEEE Transactions On Knowledge And Data Engineering, vol. 27, no. 1, January 2015.
[6] Shyi-Ming Chen, Fellow, and Bing-Han Tsai, "Autocratic Decision Making Using Group Recommendations Based on Intervals of Linguistic Terms and Likelihood-Based Comparison Relations," In Proc. IEEE Transactions On Systems, Man, and Cybernetics: Systems, vol. 45, no. 2, February 2015.
[7] Haoji Hu, Kai Zheng, Xiaoling Wang, and Aoying Zhou, "Gfilter: a general gram filter for string similarity search," In Proc. IEEE Transactions On Knowledge and Data Engineering, vol. 27, no. 4, april 2015.
[8] Q. Mei, X. Shen, and C. Zhai, "Automatic labeling of multinomial topic models," in Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2007, pp. 490–499.
[9] A. Tagarelli and G. Karypis, "A segment-based approach to clustering multi-topic documents," Knowl. Inform. Syst., vol. 34, no. 3, pp. 563–595, 2013.

All rights reserved by www.ijsrd.com