ISSN: 2278-4632
Vol-10 Issue-7 No. 11 July 2020

EVALUATING TOPIC MODELS USING LATENT DIRICHLET ALLOCATION

Mr. B. RAVITEJA

Assistant Professor, Department of CSE, Sri Mittapalli College of Engineering, Tummalapalem, NH-16, Guntur, Andhra Pradesh, India.

1) S. CHANDRA SEKHAR , 2) P. SAI KRISHNA , 3) SK. DEENUL ANSARI

B. Tech Students, Department of CSE, Sri Mittapalli College of Engineering, Tummalapalem, NH-16, Guntur, Andhra Pradesh, India.

ABSTRACT: It is observed that distinct words in a given document have either strong or weak ability in delivering facts (i.e., the objective sense) or expressing opinions (i.e., the subjective sense) depending on the topics they associate with. Motivated by the intuitive assumption that different words have varying degree of discriminative power in delivering the objective sense or the subjective sense with respect to their assigned topics, a model named as identified objective–subjective latent Dirichlet allocation (LDA) (iosLDA) is proposed in this paper. In the iosLDAmodel, the simple Pólya urn model adopted in traditional topic models is modified by incorporating it with a probabilistic generative process, in which the novel "Bag-of-Discriminative Words" (BoDW) representation for the documents is obtained; each document has two different BoDW representations with regard to objective and subjective senses, respectively, which are employed in the joint objective and each document has two different BoDW representations with regard to objective and subjective senses, respectively, which are employed in the joint objective and subjective classification instead of the traditional Bag-of-Topics representation.

I. INTRODUCTION

There is a growing demand of automatic analysis on the multimodal data (e.g., electronic documents, images, audio and video data, and so on) that can be easily found and obtained from the Internet. So far, various machine learning algorithms have been employed in accessing, retrieving, clustering, and summarizing the data. Among them, topic models are more and more popular due to their ability to efficiently discover the latent structure embedded over a group of documents and provide low-dimensional representation for large-scale data. The earliest topic model is probabilistic latent semantic analysis (pLSA) that evolves from LSA. In pLSA, documents are projected into a low-dimensional topic space by assigning each word with a latent topic, where each topic is usually represented as a multinomial distribution over a fixed vocabulary. While various extensions of pLSA have been proposed in recent years, the most famous and successful one amongthem remains to be Latent Dirichlet Allocation (LDA).

ISSN: 2278-4632
Vol-10 Issue-7 No. 11 July 2020

It is observed that distinct words in a given document have either strong or weak ability in delivering facts (i.e., the objective sense) or expressing opinions (i.e., the subjective sense) depending on the topics they associate with. Motivated by the intuitive assumption that different words have varying degree of discriminative power in delivering the objective sense or the subjective sense with respect to their assigned topics, a model named as identified objective–subjective latent Dirichlet allocation (LDA) (iosLDA) is proposed in this project. In the iosLDA model, thesimple Pólya urn model adopted in traditional topic models is modified by incorporating it with a probabilistic generative process, in whichthe novel "Bag-of-Discriminative- Words" (BoDW) representation for the documents is obtained; each document has two different BoDW representations with regard to objective and subjective senses, respectively, which are employed in the joint objective and subjective classification instead of the traditional Bag-of-Topics representation. The experiments reported on documents and images demonstrate that: 1) the BoDW representation is more predictive than the traditional ones; 2) iosLDA boosts the performance of topic modeling via the joint discovery of latent topics and the different objective and subjective power hidden in every word; and 3) iosLDA has lower computational complexity than supervised LDA, especially under an increasing number.

II. METHODOLOGY

A) Probabilistic Topic Models: As our collective knowledge continues to be digitized and stored—in the form of news, blogs, Web pages, scientific articles, books, images, sound, video, and social networks—it becomes more difficult to find and discover what we are looking for. We need new computational tools to help organize, search, and understand these vast amounts of information. Right now, we work with online information using two main tools—search and links. We type keywords into a search engine and find a set of documents related to them. We look at the documents in that set, possibly navigating to other linked documents. This is a powerful way of interacting with our online archive, but something is missing. Imagine, searching and exploring documents based on the themes that run through them. Rather than finding documents through keyword search alone, we might first find the theme that we are interested in, and then examine the documents related to that theme.

B) Using Incremental PLSI for Threshold-Resilient Online Event Analysis:

The goal of online event analysis is to detect events and track their associated documents in real time from a continuous stream of documents generated by multiple information sources. In this project, we propose a threshold resilient online algorithm, called the Incremental Probabilistic Latent Semantic Indexing (IPLSI) algorithm, which alleviates the threshold-dependency problem and simultaneously maintains the continuity of the latent semantics to better capture the story line development of events. Event analysis is a challenging research topic that has many applications such as "hot" news stories provided by Internet portals, Internet event detection, e-mail event detection and discussion board topic detection.

ISSN: 2278-4632
Vol-10 Issue-7 No. 11 July 2020

The challenge of online event analysis is to detect unknown events and track their story line development from a continuous document stream generated by uncoordinated information sources.

C) Bayesian Learning for Latent Semantic Analysis:

D) Probabilistic latent semantic analysis (PLSA) is a popular approach to text modeling where the semantics and statistics in documents can be effectively captured. In this project, a novel Bayesian PLSA framework is presented. We focus on exploiting the incremental learning algorithm for solving the updating problem of new domain articles. This algorithm is developed to improve text modeling by incrementally extracting the up-to-date latent semantic information to match the changing domains at run time.

E) Online PLSA: Batch Updating Techniques Including Out-of-Vocabulary Words:

The pLSA employs a fixed-size moving window over a document stream to incorporate new documents and at the same time to discard old ones (i.e., documents that fall outside the scope of the window). In addition, the pLSA assimilates new words that had not been previously seen (out-of-vocabulary words), and discards the words that exclusively appear in the documents to be thrown away. To handle the new words, Good-Turing estimates for the probabilities of unseen words are exploited. The experimental results demonstrate the superiority in terms of accuracy of the pLSA over well-known PLSA updating methods, such as the PLSA folding-in (PLSA fold.), the PLSA rerun from the breakpoint, the quasi-Bayes PLSA, and the Incremental PLSA.

F) Latent Dirichlet Allocation:

We describe latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities.

G) Hierarchical Bayesian Models for Applications in Information Retrieval:

We present a simple hierarchical Bayesian approach to the modeling collections of texts and other large-scale data collections. For text collections, we posit that a document is generated by choosing a random set of multinomial probabilities for a set of possible "topics," and then repeatedly generating words by sampling from the topic mixture.

H) On an Equivalence between PLSI and LDA:

LDA is a fully generative approach to language modelling which overcomes the inconsistent generative semantics of Probabilistic Latent Semantic Indexing (PLSI). PLSI is maximum posteriori estimated LDA model under a uniform Dirichlet prior, therefore the perceived shortcomings of PLSI can be resolved and elucidated within the LDA framework.

ISSN: 2278-4632
Vol-10 Issue-7 No. 11 July 2020

III. ANALYSIS

The Systems Development Life Cycle (SDLC), or Software Development Life

Cycle in systems engineering, information systems and software engineering, is the process of creating or altering systems, and the models and methodologies that people use to develop these systems. In software engineering the SDLC concept underpins many kinds of software development methodologies.

A) Existing System: The earliest topic model is probabilistic latent semantic analysis (pLSA). In pLSA, documents are projected into a low-dimensional topic space by assigning each word with a latent topic, where each topic is usually represented as a multinomial distribution over a fixed vocabulary. As an unsupervised model, the original LDA model is built based on the "Bag-of-Words" (BoW) representation, where the documents are treated as unordered collections of words, disregarding any linguistic structures embedded in them. LDA utilize the Bag-of-Topics (BoT) representation of one document for the prediction of its corresponding label, in which the proportion of topics (instead of the word proportion in BoW) in the document is considered to be the predictive feature.

Disadvantages of Existing System:

1.It is often found not to be so strongly predictive.**2.**The unsupervised manner employed in LDA unfortunately loses sight of the nature of various discriminative tasks, such as classification and regression, and thus provides no guarantee on the effectiveness of the learned representation**.**

B) Proposed System: As a result, a supervised approach named as identified objective– subjective LDA (iosLDA) is proposed in this project that extends the basic framework of multiclass sLDA in many aspects. In the iosLDA model, the simple Pólya urn (SPU) model followed by traditional topic models is modified by incorporating it with a probabilistic generative process to obtain the novel "Bag - of-Discriminative-Words" (BoDW) representation for the documents.

Each document has two BoDW representations with regard to objective and subjective senses, respectively, which are then employed in the joint objective and subjective classification.

Advantages of Proposed System:1) The BoDW representation is more predictive than the traditional ones.**2)** The model iosLDA boosts the performance of topic modeling via the joint discovery of latent topics and the different objective and subjective power hidden in every word. **3)** The model iosLDA has lower computational complexity than sLDA, especially under an increasing number of topics.**_**SDLC is nothing but Software Development Life Cycle. It is a standard which is used by software industry to develop good software.

IV. DATA FLOW & WORD REGIONS

It illustrates how data is processed by a system in terms of inputs and outputs. Data flow diagrams can be used to provide a clear representation of any business function. The technique starts with an overall picture of the business and continues by analyzing each of the functional areas of interest. A DFD can be

ISSN: 2278-4632
Vol-10 Issue-7 No. 11 July 2020

easily drawn using simple symbols. Additionally, complicated processes can be easily automated by creating DFDs using easy-to-use, free downloadable diagramming tools. A DFD is a model for constructing and analyzing information processes. DFD illustrates the flow of information in a process depending upon the inputs and outputs.

Objective Region: If the word that represents one region in the given image is identified as strongly objective, this region is detected to be a descriptor of the dominant object (e.g., "face" or "car") in the image.

Subjective Region: The regions consisting of visual words that are strongly subjective will be identified as a region that mainly describes the overall sentiment (e.g., "sad" or "surprise").

V. CONCLUSION AND FUTURE WORK

In this project, a supervised topic model named as iosLDA is proposed to discover the words that either discriminative or trivial in delivering an objective or a subjective sense with respect to their assigned topics. To achieve this goal, first, the SPU model adopted in traditional topic models is modified by incorporating it with a probabilistic generative process, making it possible to obtain the novel BoDW representation for the documents; after that, each document is defined to have two different BoDW representations with regard to objective and subjective senses, respectively, which are employed in the joint objective and subjective classification instead of the traditional BoT representation. 1)The BoDW representation is more predictive than the traditional ones; 2) iosLDA boosts the performance of topic modeling via the joint discovery of latent topics. 3) iosLDA has lower computational complexity than sLDA, especially under an increasing number of topics.

VI. REFERENCES

1. D. M. Blei, L. Carin, and D. Dunson, "Probabilistic topic models," IEEE Signal Process. Mag., vol. 27, no. 6, pp. 55–65, Nov. 2010.

2. T. Hofmann, "Probabilistic latent semantic indexing," in Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., 1999, pp. 50–57.

3. M. W. Berry, S. T. Dumais, and G. W. O'Brien, "Using linear algebra for intelligent information retrieval," SIAM Rev., vol. 37, no. 4, pp. 573–595, 1995.

4. N. Chen, J. Zhu, F. Sun, and B. Zhang, "Learning harmonium models with infinite latent features," IEEE Trans. Neural Netw. Learn. Syst., vol. 25, no. 3, pp. 520–532, Mar. 2014.

5. T.-C. Chou and M. C. Chen, "Using incremental PLSI for thresholdresilient online event analysis," IEEE Trans. Knowl. Data Eng., vol. 20, no. 3, pp. 289–299, Mar. 2008.

6. J. T. Chien and M. S. Wu, "Adaptive Bayesian latent semantic analysis," IEEE Trans. Audio, Speech, Language Process., vol. 16, no. 1, pp. 198–207, Jan. 2008.

7. N. K. Bassiou and C. L. Kotropoulos, "Online PLSA: Batch updating techniques including out-of- vocabulary words," IEEE Trans. Neural Netw. Learn. Syst., vol. 25, no. 11, pp. 1953–1966, Nov. 2014.

8. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993– 1022.