ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 148

International Conference on Information Technology and Management Engineering (ICITME 2018)

Extracting Latent Topics from User Reviews Using Online LDA

Zilin Wang

School of Humanity & Economic Management, China University of Geosciences, Beijing, Haidian District, Beijing, 100083

***Abstract*—As local business directory service sites like Dianping.com and Yelp.com are increasingly popular, user reviews are becoming more and more important in informing customers of product and service quality. The reviews can also provide meaningful insights to business owners. However, huge amounts of online user reviews are displayed in texts and are of high dimensionality. They also imply different latent topics. Therefore, it is intractable to pinpoint the demand of customers from a large amount of incremental user reviews manually. The goal of this paper is to help businesses discover user demands from enormous reviews of high dimensionality, which in turn will help improve their business. To this end, we propose using online Latent Dirichlet Allocation (LDA) as topic model to discover latent topics from user reviews. We used the open dataset from Yelp Dataset Challenge, and further cleaned and filtered the dataset to focus on the user reviews of restaurants in Phoenix, Arizona, US. By running Online LDA over the cleaned dataset, we discovered 50 latent topics. In this paper, we present the breakdown of latent topics over all reviews and the word distribution of topics. Furthermore, the method adopted by this paper could prove useful to specific business owners in discovering user demands and points of interest.**

Keywords—natural language processing; topic model; latent dirichlet allocation; yelp reviews

I. INTRODUCTION

Consumers need to know the quality and services of goods, so online reviews increasingly become the expert's opinion and social network is worked as supplements. Many consumers will take the users' evaluations which from the Dazhongdianping (www.dianping.com), Yelp.com and other websites of local life service as an important reference for making decision before they choose businesses to consume. Research shows that the user's evaluations from Yelp.com has

A significant effect on restaurant businesses, half extra star (five stars rating system) can make a predetermined ratio of a restaurant by 19 percent (up from 30 percent to 49 percent), and when the information of other channels are scarce, its influence will be more remarkable. For those that almost don't have external comments or other promotional channels, the half extra star will make a predetermined ratio by 27 percent [1]. So, it has practical significance to discover the hidden topics from online reviews. However, huge amounts of online user reviews are displayed in texts and are of high dimensionality. They also imply different latent topics. Therefore, it's difficult to pinpoint the needs from a large number of online user reviews. The key to the problem is to find topics from multi-dimensionality user reviews, which means when dealing with high dimensional data related questions, how to extract relevant topics from large amounts of data. However, this data usually has relatively simple implicit structures, such as topics of document, user preference, topics of discussion, and so on. We can roughly estimate these effects with dimension reduction model. The topic model is used to find a statistical model implicit in the theme of the collection of documents. The topic model can simulate the process of document generation, and then obtains each topic through parameter estimation. By using the topic model to decompose the user's online evaluation into implicit sub themes, we can see the user's demand much more simply and intuitively. To automatically extract hidden topics from online reviews of business users, this paper proposes a method to discover hidden topics in online user reviews using online hidden Dirichlet Lickley distribution (online LDA) as topic model. Based on the open Yelp Dataset Challenge (http://www.yelp.com/dataset_challenge/) data sets which were collected in Arizona Phoenix Restaurant, more than 140 thousand user review data are used as the experimental data, and the use of online LDA model in the data set as the topic model text for unsupervised training can get implicit comment text theme.

II. LDA

A. Probabilistic Description of LDA

LDA[2] is a generative probability model for discrete data sets such as text sets that can be used as a topic model to discover topics that are implicit in text documents. It is a three-tiered Bayesian model of "document-topic-word" that models each of the data sets, such as each text, as a mixture of sets of unknown topics. Each topic is modelled as some sort of mixed probability distribution. LDA and other thematic models are part of the probabilistic model, and the probabilistic model is a much larger category. In general, LDA is based on the posterior distribution framework. The computational problem of inferring the underlying subject structure from the document is to calculate the posterior distribution, which is the conditional distribution of implicit variables for a given document [3].

We formally define the subject as a distribution on a fixed vocabulary. The theme model can be expressed in two ways: the generation process and the graph model. The simplest way

to express LDA is its generation process, that is, the stochastic process that the model assumes. The LDA generates words for each document in the collection, which is shown in Figure 1.

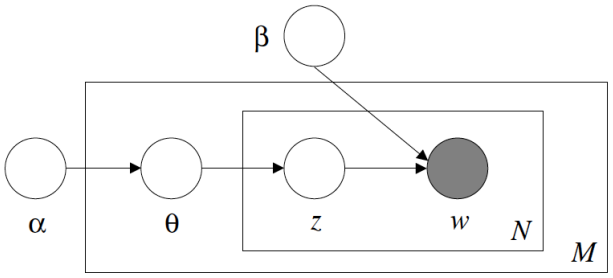

FIGURE I. GRAPH MODEL REPRESENTATION OF LDA

In this figure, the box represents the content of which is repeated, the lower right corner N is the number of repetitions; the grey node w represents the observed value, and the remaining open nodes represent the implied random variables or parameters. The arrow indicates the dependency. θ is the subject distribution vector, z is one of the topics, α is the parameter of the prior distribution Dirichlet distribution of the subject distribution, β is the parameter of the prior distribution Dirichlet distribution of the subject word distribution.

B. Posterior Calculation of LDA

The principle and representation of LDA are introduced above, now we go to the problem of calculation, that is, given the observation document and calculating the conditional distribution of the implied topic structure, which is also called posterior distribution. For the subject model and other Bayesian models, posterior distributions are difficult to calculate, so researchers have to estimate posterior inferences. A central research problem in modern probability models is an efficient way to find approximate posterior. Thematic model algorithms are usually an adapted version of the general approach to approximate posterior distribution [3]. The approximate posterior distribution algorithms commonly used in thematic models are generally divided into two types: sampling approaches and optimization approaches. The sampling method is usually based on the Markov Chain Monte Carlo[10] (MCMC) method. The optimization method is usually based on variational inference. In the Bayesian hierarchy model, this method is called variational Bayes. The variational approach opens the door to innovation in optimization methods that has a real impact on the probability model. Although the posterior distribution of choice approximation is biased, the VB method is empirically analyzed to prove much faster with the same accuracy as the MCMC method. Therefore, the VB method is a good choice for applying the Bayesian model to a large dataset [4].

C. Online Variational Bayesian Learning Algorithm for LDA

As mentioned earlier, the VB method is a good choice for applying large Bayesian models to large datasets. However, large-scale data analysis using VB is computationally very difficult [5]. This problem is even more prominent in the application of the topic model. To solve this problem, Hoffman etc. [6] proposed an online variational Bayes learning algorithm for LDA to fit the variational posterior parameters of the subject distribution. Online LDA is almost as easy as a batch VB algorithm, but much faster on big datasets.

The online LDA algorithms converge into the local optimum of the variational Bayesian objective function. This online VB method is a practical method for estimating the posterior distribution in a complex Bayesian hierarchy model. Online LDA is based on online stochastic optimization and has been studied to confirm that good parameter estimation results on large datasets are much faster than batch algorithms, so online LDA can easily be used for large-scale document datasets.

This paper selects the LDA model as the theme model, at the same time, chooses online LDA this online learning algorithm. Online LDA algorithms converge faster, user comments are batched, and subject models incrementally and continually update after each batch is processed. In this way, constant memory can be used in constant time to complete the calculation, but also to satisfy the continuous growth of online restaurant users' comments. Although the experimental data in this paper is small, the method proposed in this paper is more scalable and can be easily extended to larger datasets. In this paper, online LDA method is chosen.

III. RESEARCH DESIGN

A. Research Objective

User reviews on local life service sites such as Yelp.com are of great importance to local businesses and have a strong positive correlation between user ratings and profitability of merchants. Therefore, it is very important for the merchants to find out the user's needs in the user's comments so as to improve it in a targeted manner, thereby improving the user's rating and increasing the positive rating. This article translates the problem of finding user needs from online reviews into the problem of finding hidden topic structures in corpora. The goal of the study is to automate the process of extracting implicit topics from merchant users 'online reviews, which will help merchants find users' needs and points of interest from reviews, thereby improving product and service quality.

B. Research Route Chart

Here is the route chart.

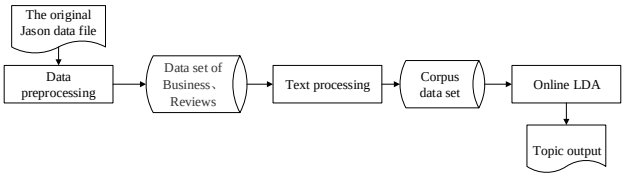

FIGURE II. THE ROUTE CHART

C. Tools

Text data pre-processing and online LDA model in this paper are implemented using the open source Python scripting language. The natural language processing related Python libraries used in this article are mainly NTLK[9]

(http://www.nltk.org/) and Gensim (https://radimrehurek.com/gensim/). NLTK is a Python library dedicated to natural language processing, which provides many corpora and lexical resources such as WordNet. At the same time, NLTK also provides a library of text processing, including text classification, POS tagging, parsing, semantic analysis and so on [7]. Gensim is the Python library that implements the thematic model for large corpora.

In addition, this paper also uses MongoDB to access the data. MongoDB is a document-oriented database, not a relational database that can get more convenient scalability. Document is the core concept of MongoDB, putting together in order by multiple keys and their associated values. Collections can be thought of as tables without patterns. Each database has its own collection and permission[8]. MongoDB implements the object-oriented idea well that every record in MongoDB is a ddocument object. All MongoDB data operations without the need to manually write SQL statements can easily be called (Create), retrieve (retrieve), update (Update) and delete operations (CRUD[11] operations). At the same time, the interaction between MongoDB and Python is also very simple and can be easily implemented by using the Pymongo library.

IV. EXPERIMENTAL PROCESS AND RESULTS

A. *Experimental Process*

1) Data pre-processing

Because of lots of redundant data in the dataset, we should organize and clean the data, filter and import it and then for subsequent processing. Firstly, we filter categories for restaurants and businesses in the city of Phoenix from a json file named "yelp_academic_dataset_business" in the Yelp dataset and store the 'business_id' of the relevant merchant into the data collection called "Business". Then we extract the 'business_id' rating that matches 'business_id' in Business from a json file named 'yelp_academic_dataset_review' and import it into a MongoDB collection named 'Reviews', so that only retain the relevant business and rating data.

2) Importing filtered and cleared data into MongoDB to get the result

3) Text processing

Text processing mainly uses NLTK. This article for text corpus processing methods are based on the background of this article. After the initial sorting and screening, this article further processes the initial Reviews dataset, mainly for corpus processing. First, loop through the following actions for each comment in the initial database: step 1, tokenization; step 2, stopping words; step 3, filter punctuation; step 4, calculating the number of words in the corpus; step 5, POS tagging; step 6, simplifying the inflection.

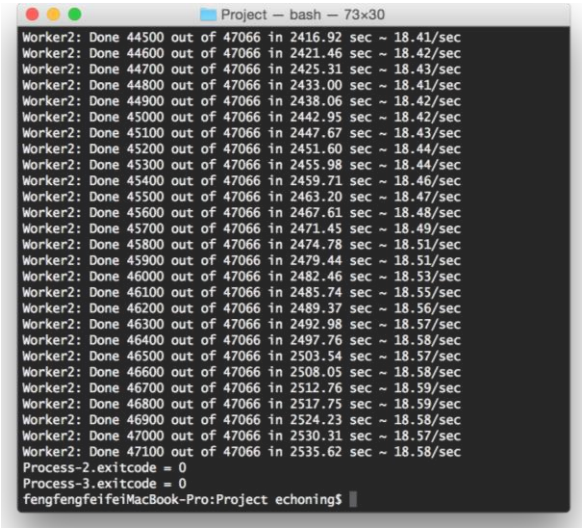

FIGURE III. DO TEXT PROCESSING TO THE TEXT IN THE REVIEW DATABASE

All the comments of the data set in MongoDB after the above steps are processed in a loop in the following way: First, filter out all the words that are not nouns, and then use WordNetLemmatizer to find the vocabulary of each nouns. Finally, each comment including review ID, the name of the restaurant business, the comment text, and the glossary of terms in the commentary, are stored in a new MongoDB collection and named Corpus. That is, until this step, a corpus for model training has been formed. Processed databases and data collections are shown in afterwards.

4) Corpus processing MongoDB database and data collection situation

5) Implementation of online LDA

In the corpus theme model which is used in this paper, the word dictionary size is found by the calculation of the previous text processing, that is, the topic. After the stop word was removed, only the top 10,000 words left were considered. The experiment found that at that time, the experimental results on the restaurant user review dataset were quite satisfactory. If the number of topics is too small, terms originally belonging to different topics will be grouped under the same topic; conversely, if values are too large, the terms that should belong to the same topic will be assigned to different topics. Therefore, this article through a number of comparison of test results, finalized. The specific algorithm to achieve the main use of the open source Python library Gensim, which is shown in figure 6.

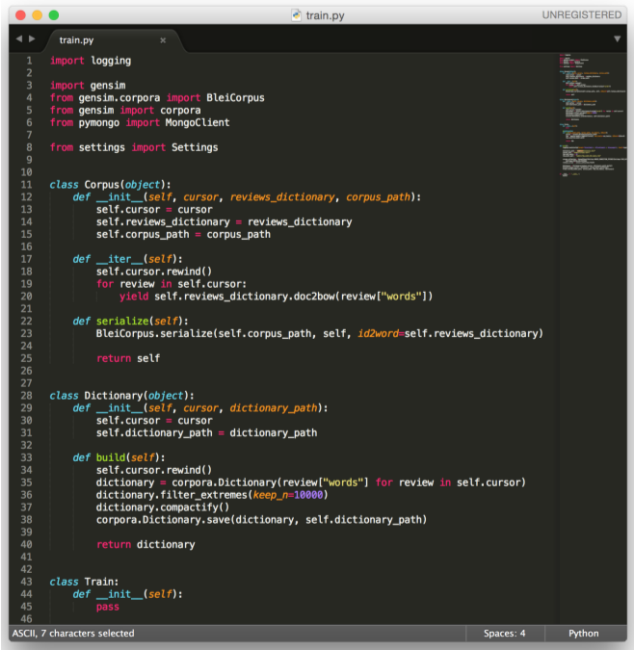

FIGURE IV. THE SCREENSHOT OF ONLINE LDA ALGORITHM CODE

6) Using LDA algorithm code to process the data

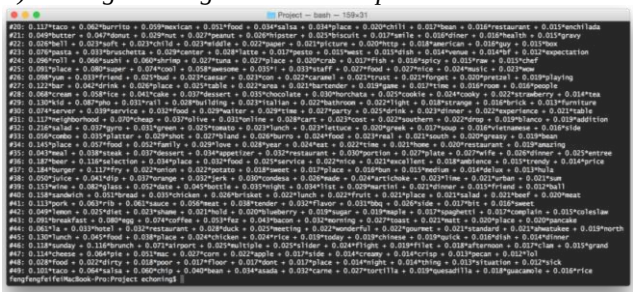

FIGURE V. THE SCREENSHOT OF ONLINE LDA ALGORITHM CODE

B. Experimental Results

1) Thematic distribution of topics

By running an online LDA of 50 topics on a lexical dictionary of more than 10,000 terms, we conclude with 50 implicit topics and lexical distributions of the corresponding topics in this paper. As an example, the specific topic distribution and frequency of some topics are shown below in Table 1. Each term belongs to at least one of these topics and may have overlapping parts.

TABLE I. PART OF THE THEME OF THE WORD DISTRIBUTION

Service	Italian food	Thai food	Mexican food
6.8% service	31.9% pizza	5.8% Thai	11.7% taco
4.9% order	2.9% slice	5.4% curry	6.2% burrito
4.4% server	2.4% cibo	3.7% spicy	5.9% Mexican
4.2% time	2.3% topping	3.4% dish	3.5% salsa
3.4% wait	2.2%	2.9%	3.4% chili
	cheese	tea	
Parking	**Location**	**Flavour**	**Decor**
5.7% parking	25.5% location	4.6% flavour	7.0% patio
4.8% lot	3.6% downtown	3.4% dish	5.5% seating
3.1% car	2.4% Scottsdale	2.8% texture	3.6% inside
2.7% valet	2.0% phoenix	2.1% taste	3.3% vibe
2.4% place	1.9% sister	2.1% sauce	2.4% nice

2) User reviews mainly hidden topics

From the 50 topics generated by the online LDA model, the paper selects 10 of the more significant and frequently occurring topics to examine, and the percentage of topics and corresponding topics to the implied topics of all the comments is shown in Table 2.

TABLE II. THE MAIN HIDDEN TOPICS IN THE DISTRIBUTION OF COMMENTS

theme	service	value	take out	parking	healthiness
percent	8.9%	5.86 %	3.75%	3.01%	2.71%
theme	**wait**	**decor**	**breakfast**	**dinner**	**lunch**
percent	2.62 %	1.79 %	1.64%	0.97%	0.76%

V. CONCLUSIONS

In this paper, the review data of Phoenix Restaurant is screened on the Yelp dataset, and an online LDA model is implemented on this dataset. Based on the online LDA model training, we obtained 50 implicit topics i the user comment text and the lexical distribution of each topic, which successfully solved the proposed problem. At the same time, the experiment also found that Phoenix restaurant users in the choice of restaurants most concerned about the point of interest. This article has completed the task of using online LDA as a topic model to discover implicit topics for user reviews on Yelp datasets. According to the experimental results, restaurants can quickly and easily find the needs of users, based on that, they can improve and enhance the quality of service in all aspects to improve the rating of restaurant businesses and enhance their performance. Meanwhile, experiments show that for a merchant, the experimental method in this paper can efficiently attract the customer's needs and points of interest, which is beneficial to dig customer preferences in different locations.

REFERENCES

[1] M. Anderson and J. Magruder. Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database. The Economic Journal. 2012, Volume 122, Issue 563: 957–989.

[2] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003: 993–1022.

[3] D.M. Blei. Probabilistic Topic Models. Communications of the ACM, 2012, vol. 55, no. 4: 77–84.

[4] D. Blei, M. Jordan. Variational methods for the Dirichlet process. ICML '04 Proceedings of the 21st international conference on Machine learning. 2004: 12.

[5] D. Newman, A. Asuncion, P. Smyth, and M.Welling. Distributed inference for latent Dirichlet allocation. Neural Information Processing Systems, 2007.

[6] M. Hoffman and D. Blei. Online Learning for Latent Dirichlet Allocation. Neural Information Processing Systems, 2010.

[7] S. Bird, E. Klein and E. Loper. Natural Language Processing with Python. Sebastopol, CA: O'Reilly Media, 2009.

[8] K. Chodorow, M. Dirolf. MongoDB:The Definitive Guide (Translated by Cheng Xianfeng)[M]. Beijing: Posts & Telecom Press, 2011: 5-38.

[9] Devanshi Gupta,Priyank Singh Hada,Deepankar Mitra,Niket Sharma. Identification of Lost or Deserted Written Texts Using Zipf's Law with NLTK[M].Springer International Publishing:2014-06-15.

[10] N. A. Sheehan. On the Application of Markov Chain Monte Carlo Methods to Genetic Analyses on Complex Pedigrees[J]. International Statistical Review, 2000,68(1).

[11] Daniel J. Walter,Brian K. Kendrick,Victor Petrov,Annalisa Manera,Benjamin Collins,Thomas Downar. Proof-of-principle of high-fidelity coupled CRUD deposition and cycle depletion simulation[J]. Annals of Nuclear Energy,2015,85.

[12] Jongseong Yoon,Doowon Jeong,Chul-hoon Kang,Sangjin Lee. Forensic investigation framework for the document store NoSQL DBMS: MongoDB as a case study[J]. Digital Investigation, 2016,17.