The application of LDA model on user profile

Wei Yao [a, *], Jia Suling, Wang Qiang, Yu Hao
School of Economics and Management, Beihang University, Beijing 100191, China
[a] weiyao1024@126.com

Abstract. This paper puts forward the method of user profile technology based on LDA model. Abstracting the behavior and transaction information of the research object into the document, the LDA model can be trained and the research object can be described into the property tags, thereby to realize the user profile. The experimental results show that describing the research object as multidimensional labels based on LDA model can effectively characterize the research object and form their user profile.

Keywords: e-commerce, user profile technology, LDA model, Gibbs sampling algorithm, big data, data mining

1 Introduction

With the explosive growth of the Internet, the problem of in- formation overload has become more and more critical. For enterprises and organizations, finding consumers of interest from a large population of users is quite a challenge as a result of the rapid growth of the user base and their diverse activities on the Web[7,8,9].

The earliest text data mining method is based on the vector space model which referred to VSM. Subsequently, Deerwesteret al. proposed the latent semantic analysis (LSA) [10]. LSA uses the singular value decomposition (SVD) method in linear algebra to reduce the dimension of the word-document matrix to map the word-document into a low-dimensional latent semantic space [6]. In 1999, Hofmann et al proposed the probabilistic latent semantic analysis model (pLSA)[11], and Blei et al. proposed the Latent Dirichlet Allocation model (LDA) in 2003[13]. The LDA model not only integrates the advantages of PLSA, but also overcomes the theoretical flaws of PLSA, which is widely used in many fields. LDA model has become the most famous topic model for the past decade[1].

User profile technology is a special user modeling technology generated under the big data environment. Through sorting of the massive user information collected, the basic information and behavior information of each user are analyzed from multiple dimensions. The results of the multidimensional analysis get together to form the property tags. These tags constitute a complete user model and interest model, which can be more close to the actual description of a real user [4].

Based on these, this paper puts forward the method of user profile technology based on LDA model. Abstracting the behavior and transaction information of the research object into the document, the LDA model can be trained and the research object can be described into the property tags, thereby to realize the user profile.

2 Related work

2.1 LDA model

The traditional way to determine the similarity of two documents is to compare the number of words commonly included in the two documents, such as TF-IDF (term frequency inverse document frequency) method. The LDA model assumes that a document is made up of a certain percentage of the topics in the topic set, and each topic is a mixture of words in the word list in a certain proportion. Through the machine learning method, the topic of the document can be got, so that the similarity of two documents can be got. LDA model is clearly stratified by the document layer, the theme layer and the word layer, in which the document is associated with the topic and the topic is associated with the word[3]. You can dig out all the potential topic information by learning the words in the document set

and use it to dig out the topic distribution of documents other than the document set.Latent Dirichlet Allocation (LDA) is a type of generative probabilistic model [2].

As described above, the basic concepts involved in LDA mode are as follows:

1) word: Word is the basic unit of discrete data, all the words in the corpus are from a certain set of words, which size is V.

2) Document: Just like the pLsa,LDA is based on the assumptions of bag of words model, that is, a document is composited the words. Only the number of words in the document is considered, regardless of their order and constraint relationship in the document. So a document can be represented by a word vector. Assuming that there are N words in the document, then the document can be expressed as w = (wl, w2, ..., wN).

3) corpus: Also named the document set, composed that it is composed of M documents, so it can be expressed as d = (d1, d2, ..., dM).

4) topic : Suppose the number of implicit object is T ,so that the topic set z can be expressed as z = {z1, z2, ..., zT}. In general, the number of topics(T) is known.

For a document in the corpus, LDA defines the following generation process:

select a few topics from the topic distribution for each document;

Select a word from the word distribution corresponding to the selected topic in step 1;

Repeat the steps 1 and step 2 until all the words in the document are generated;

Each of the documents in the corpus corresponds to the distributions consisting of a predetermined T-topic mix, and the polynomial is denoted as θ. Each topic corresponds to the distribution of the V words in the corpus, and the distribution is denoted as Φ. Θ and Φ have a Dirichre prior distribution with the hyperparameter α and β. For a document d,t a topic z is extracted from the polynomial relation θ corresponding to the document, and then a word w is extracted from the polynomial corresponding to the topic z. Nd times repeated, where Nd represents the number of words in the document, the document d is produced

The diagram of LDA model structure is shown below.

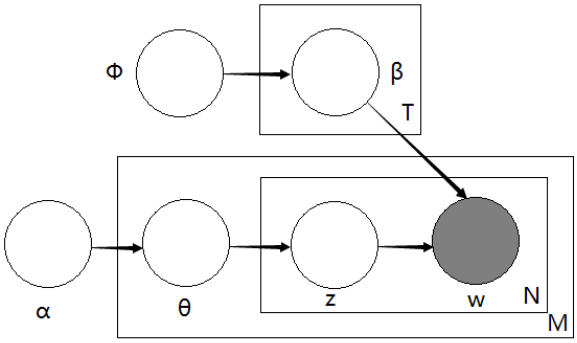

Fig. 1 LDA model structure

2.2 User profile

Tim Berners-Lee invented the World Wide Web in 1991. Until 2011, the Internet exactly went into the "big data age", which is a new milestone. After nearly two years of boiling, people gradually calm down to focus on how to use the big data to mine the potential value of business as well as how can these enterprises applying the big data technology in an effective way[5]. Compared with the traditional management of offline members, questionnaires, shopping basket analysis, big data allows enterprises to easily access to users' extensive feedback through the Internet. And it also provides a sufficient data base to analysis the users' behavior habits, consumption habits and other important business information in a more accurate and rapid way. The concept of user profile which can perfectly abstract a user's information is born. User profile technology also can be seen as the basis for enterprises to apply the big data into reality life.

User profile technology is a special user modeling technology generated under the big data environment. By sorting the collected mass information of users', the basic information and behavior information of each user are analyzed from multiple dimensions. The results of these analysis get

together to form property tags in order to constitute a complete user model and interest model, which can be more close to the actual description of a real users[4].

3 User profile technology based on LDA model

3.1Keyword extraction

At first before training the model, pretreatment of data is needed. The original data is produced by SQL in order to acquire the effective data. Each record of the filter data is seen as a document. The corpus is made up of these record, also these “documents”. The size of the corpus is denoted as M. In order to get the words list, the documents in corpus should be cut into words. For this operation, Jieba tool for words cut of Chinese is used, which is run basis on python.

After cutting the corpus into divided words, frequency of these words should be count. Spark Mllib can simply achieve this under python or scala , which is a machine learning library to process the big data. All of the words can be sorted by its frequency from high to low. From the sorted words, few words should be chosen artificially based on their frequency as well as semantics. After this step, the words set can be got, whose size is V.

3.2 Conversion of document - word matrix

In order to convert the documents of the corpus into document-word matrix, each document should be processed individually into w = (wl, w2, ..., wN). Based on the V-size word list, each of the documents are divided into wi = (wl, w2, ..., wV).Matrix J = (w1,w2…wM)T.

In the step of matrix translation, spark mllib based on python is used. Cut each of the document into divided words first, jieba tool based on python can be user to process this step. In order to get the vector of document-keywords for each document, the frequency of key words of words list in each document must be count. From the key-words count of each document, the list of documents’ vector whose size can be got. This step can be operated by spark mllib who has the special structure of RDD. Under RDD, operation of data can be exactly brief. There are also many other tools can be used to define the data structure. After combine the vector list into a matrix and saved the matrix into input-data file, the matrix of document-word which is the input data for model training can be got.

3.3 Building of LDA model

With the input data of document-word matrix, training of the LDA model can be run which is developed based on Spark Mllib. Parameter of K which is the number of clustering centers, also the number of topics can be set for an estimate number decide of the size of word list. The number of K can be adjust by P(w|T) or perplexity. In this paper, In order to get the P(w|T), the hyperparameter α and β should be estimated with the Gibbs sampling algorithm.

After the estimate of parameter, Spark Mllib can be used to train the model. The trained console includes the word-topic distribution and the document-topic distribution. With the console, things can be done as below:

(1) predict the topic distribution of a document.
(2) get the "document - topics" distribution.
(3) choose the topic of the largest probability.
(4) query the keywords of a topic its corresponding probability in LDA moel.

3.4 Building of user profile

With the distribution of word-topics, different topics can be divided into a group of labels according to their corresponding words. In this way, each of an object whose behavior was abstracted to a document can be divided into different labels of a dimension. Processing the object in the same way from different information of documents can the object be characterized into a profile with different dimensions of labels. With the user profiles, we can better know what kind of preference does the customer like and provide them more accurate service.

4 Experiments and analysis

4.1 Introduction of the experiment

In our experiment, the develop tool we chose is Intellij IDEA who was installed with the Spark machine learning framework. The language we used were Python as well as Scala. Based on these environments, the tool we used for the word segment is the Jieba Chinese word segmentation tool. The basic java development environment was JDK 1.8.

In this experiment, the transaction data of insurance policies from intermediary of insurance were used. Based on the insurance company, we extracted all of the insurance policies each company underwritten and combined their insurance type into a document for one company. In this way,193 of the companies were abstracted as 193 documents.

4.2 Console of the experiment

Based on the frequency and semantics of words, key words were chosen as below， the key words were translated into English from Chinese：

Table 1 Chosen key words

ID	Key words	Frequency	ID	Key words	Frequency
1	motor vehicle	19057	8	pollution	1195
2	business	10276	9	construction	1192
3	integrated	9244	10	cargo transportation	569
4	group	1440	11	belongings	291
5	co-insurance	1494	12	engineering	172
6	accident	1434	13	machine	147
7	civil liability	1212	14	installation	146

Five topics were set as the K number and the LDA model output the word-topic distribution as below:

Table2 word-topic distribution

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
motor vehicle	0.144436631	105.333223	0.765269757	0.103172936	1.65389763
business	0.064337531	25.653931	0.526025103	0.131318082	5.624388282
integrated	0.080895589	230.6048191	399.8259849	122.5435235	174.9447769
group	387.8464992	0.035611685	0.04545343	0.0288121	0.043623631
co-insurance	0.076915285	358.2860028	131.9055264	0.282424013	197.4491314
accident	242.8855525	0.026026537	0.035343467	0.020733539	0.03234399
civil liability	0.095738357	0.04555567	68.63054805	0.168610999	0.059546928
pollution	0.140306046	0.033898851	22.66648408	0.116440861	0.042870159
construction	0.023811619	0.043021231	0.046458533	0.043376908	116.8433317
cargo transportation	0.197446906	0.058995648	27.54386651	0.129558169	0.070132765
belongings	0.020635877	0.038484174	0.040719125	0.04141793	106.8587429
engineering	0.027159601	0.051686492	0.25732635	468.589171	0.074656585
machine	378.8553162	0.033552571	0.04291882	0.02732602	0.040886387
installation	0.040259509	0.06303015	8.506087937	284.2992379	0.0913845

From the word-topic distribution, we got the weights of each word for each topic, according to the weights, each topic with its corresponding key words are list as below:

Table3 topic's corresponding words

topic	Key words
Topic 0	group,accident,machine
Topic 1	motor vehicle,business,integrated,co-insurance
Topic 2	integrated,co-insurance,civil liability,pollution cargo transportation
Topic 3	Integrated,engineering,installation
Topic 4	Integrated, co-insurance, construction, belongings

The model also output the document-topic distribution. Part of the console for 10 documents(totally 193 documents)were list as below，the companies' name and labels were translated into English from Chinese:

Table4 document-topic distribution

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
Junlong Life Insurance Company Limited	0.999706 46	6.90E-0 5	8.30E-0 5	5.45E-0 5	8.71E-0 5
China Union Property Insurance Company Limited	1.42E-04	0.07113 43	0.15381 028	0.71319 14	0.06172 23
China United Property Insurance Company Limited Shanghai Branch	5.37E-04	0.01342 51	0.55642 494	0.41778 33	0.01182 92
China United Property Insurance Company Limited Shanghai Branch	5.63E-05	0.03912 33	0.09057 259	0.83691 94	0.03332 84
China United Property Insurance Company Limited Shanxi Branch	0.002334 448	0.86437 79	0.01827 388	0.00334 28	0.11167 1
China United Property Insurance Company Limited Hebi Branch	0.001633 247	0.95940 55	0.01290 825	0.00152 18	0.02453 11
China Life Insurance Company Limited Xiamen Branch	0.999706 46	6.90E-0 5	8.30E-0 5	5.45E-0 5	8.71E-0 5
China Life Insurance Company Limited Tianjin Branch	0.999567 386	1.02E-0 4	1.22E-0 4	8.00E-0 5	1.29E-0 4
China Life Insurance Company Limited	3.64E-04	5.83E-0 4	6.69E-0 4	4.89E-0 4	0.99789 42
China Life Insurance Company Limited Beijing Branch	2.20E-04	0.04202 72	0.02066 271	0.23313 15	0.70395 9

According the topics，3 groups of labels were list as bellow.

Table5 labels of topics

Label l	large-scale	small-scale
topic	2,3,4	0,1
Label 2	high quota	low quota
topic	0，2,3，4	1
Label 3	exclusive insurance	co-insurance
topic	0,3	1，2，4

Finally, companies are labeled as below (10 companies are shown from totally 193)

Table6 labels of documents

	label 1	label 2	label 3
Junlong Life Insurance Company Limited	small-scale	high quota	exclusive insurance
China Union Property Insurance Company Limited	large-scale	high quota	exclusive insurance
China United Property Insurance Company Limited Shanghai Branch	large-scale	high quota	co-insurance
China United Property Insurance Company Limited Shanghai Branch	large-scale	high quota	exclusive insurance
China United Property Insurance Company Limited Shanxi Branch	small-scale	low quota	co-insurance
China United Property Insurance Company Limited Hebi Branch	small-scale	low quota	co-insurance
China Life Insurance Company Limited Xiamen Branch	small-scale	high quota	exclusive insurance
China Life Insurance Company Limited Tianjin Branch	small-scale	high quota	exclusive insurance
China Life Insurance Company Limited	large-scale	high quota	co-insurance
China Life Insurance Company Limited Beijing Branch	large-scale	high quota	co-insurance

4.3 Analysis of console

From the final console, each companies were described into different dimensions of labels. It can be seen that the console intuitively reflected features of each company. But there are also some

aspects to improve such as the chosen of key word of “integrated” might not be that effective for training the model.

When parting the labels, there are also topics of blurred boundaries to corresponds the label， for example, topic 0 might be blurred between “high quota” and “low quota”. In this way, how to expand the label and how many kinds of labels should be set in one dimension needs to be better measured.

5 Conclusions

This paper started from LDA model, extracting information of each research object as a document to form the corpus. Training the LDA model from the corpus and divided the topics into different dimensions of labels, these research objects were labeled into user profile. With the use of user profile based on the LDA model, objects are intuitively described into groups of labels.

There are also directions for future research such as the evaluation system of the effectiveness of the user profile.

From the experiment, we thought that the user profile technology based on LDA model is effective to describe research object into dimensions of labels. It’s an effective way to process the mass record of transaction data such as the insurance policies from intermediary of insurance.

References

[1] Yezheng Liu,Jiajia Wang,Yuanchun Jiang. PT-LDA: A latent variable model to predict personality traits of social network users [J]. Neurocomputing. 2016. 210: 155-163

[2] Alexander Gross,Dhiraj Murthy. Modeling virtual organizations with Latent Dirichlet Allocation: A case for natural language processing [J]. Neural Networks. 2014. 58: 38-49

[3]Shan Bin,Li FangA.Survey of Topic Evolution Based on LDA[J]. JO URNA L OF CHINESE INFORMA TION PROCESSING 2010(06): 43-49 (in Chinese)

[4]Fang Yaoyao. Research and Implementation of Information Push System based on Mobile Internet [D]. North China University of Technology, 2016

[5]Yu Mengjie.Data Modeling of User Portrayal in the Product Development: From Concrete to Abstract [J].Design Research, 2014, 4(6):60-64

[6] Xiuze Zhou,Shunxiang Wu. Rating LDA model for collaborative filtering [J]. KNOWLEDGE-BASED SYSTEMS. 2016. 110: 135-143

[7]Peng Zhang,Hansu Gu,Mike Gartrellet al. Group-based Latent Dirichlet Allocation (Group-LDA): Effective audience detection for books in online social media[J]. Knowledge-Based Systems. 2016. 105: 134-146

[8] E. Ferrara , P. De Meo,G. Fiumara , R. Baumgartner , Web data extraction, applications and techniques: a survey, Knowl.-Based Syst. 70 (2014) 301–323 .

[9] H. Gu, H. Hang, Q. Lv, D. Grunwald , Fusing text and frienships for location inference in online social networks, in: Proceedings of the 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT), 1, IEEE, 2012, pp. 158–165 .

[10] S. Deerwester, S. Dumais, T. Landauer, etc. Indexing By Latent Semantic Analysis. Journal of the American Society for Information Science.1990, 41(6):391–407

[11] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1999, ACM, New York, NY, USA, 1999, pp. 50–57.

[12] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.