Topic Model Allocation of Conversational Dialogue Records by

Latent Dirichlet Allocation

Jui-Feng Yeh\textsuperscript{1}, Chen-Hsien Lee\textsuperscript{1}, Yi-Shiu Tan\textsuperscript{1}, and Liang-Chih Yu\textsuperscript{2}

\textsuperscript{1}Department of Computer and Information Science, National Chia-Yi University, Taiwan (R.O.C.)
ralph@mail.nyu.edu.tw, S1020443@mail.nyu.edu.tw, S1000441@mail.nyu.edu.tw

\textsuperscript{2}Department of Information Management, Yuan-Zu University, Taiwan (R.O.C)
lcyu@saturn.yzu.edu.tw

Abstract—The topic information of conversational content is important for continuation with communication, so topic detection and tracking is one of important research. Due to there are many topic transform occurring frequently in long time communication, and the conversation maybe have many topics, so it’s important to detect different topics in conversational content. This paper detects topic information by using agglomerative clustering of utterances and Dynamic Latent Dirichlet Allocation topic model, uses proportion of verb and noun to analyze similarity between utterances and cluster all utterances in conversational content by agglomerative clustering algorithm. The topic structure of conversational content is friabilty, so we use speech act information and gets the hypernym information by E-HowNet that obtains robustness of word categories. Latent Dirichlet Allocation topic model is used to detect topic in file units, it just can detect only one topic if uses it in conversational content, because of there are many topics in conversational content frequently, and also uses speech act information and hypernym information to train the latent Dirichlet allocation models, then uses trained models to detect different topic information in conversational content. For evaluating the proposed method, support vector machine is developed for comparison. According to the experimental results, we can find the proposed method outperforms the approach based on support vector machine in topic detection and tracking in spoken dialogue.

Index Terms: Conversational dialogue, latent Dirichlet allocation, topic detection and tracking, spoken language processing.

1. Introduction

In this information suffusion age, we can get the numerous of knowledge and information on different ways, such as science document, newspaper, blog, voice messages. It’s important issues that how to arrange this information and deal with the hidden information of the unstructured documents. Document classification is a search in data mining, the purpose is, how to classify and analyse the huge and complicated document data, such news document class, science paper class, image analysis. With appropriate classification, it can analyse the data’s structure and semantic content, it also can use on the object detection and recognition. Topic classification also can call “Topic Detection and tracking,” we can classify the document when we know the topic, it can analyse the data or information which relative or similar topics. Conversational dialogue is one of the unstructured documents. Spoken language processing usually faces the ungrammatical and disfluency noise. The topic detection and tracking is more essential for understanding of spoken language especially in dialogue systems.

In recent years, there have many approaches about the classification methods for spoken documents or conversational utterances, especially in topic detection and tracking. Chen et al [3] use the web segmentation cut the web to visual blocks, and use the SVM to create the capture rule, it can capture the conference topic which on the web. Chen et al [4] propose a mix method to topic detection on interactive document. The method is mixed with semantic dependency distance and PLSA, its use semantic and syntactic information to reduce the problem that lack of semantic information. In recent years, blog is popular more than past, arrange and analysis the blog’s content were more important. Xuan et al [5] propose a tree level structure method, they use the LDA and Markov Cluster Algorithm, and finally they show the topic of content by text list. Wu et al [6] was compared tree topic model, these models is using to topic clustering on network news.

Many researches are building the news document or science document topic model. This paper will investigate topic detection and tracking with dialogues record based on latent Dirichlet allocation models. For evaluation the proposed method, support vector machine proposed by Vapnik [1] is also developed for dialogues topic detection and tracking. It is a supervised machine learning methods, and it has two advantages: (1) In the case of non-linear, SVM can mapping low dimensions on high dimensions or vector space by mapping functions; (2) SVM find the Hyperplane by vector space, it can cut up two data because the data's margin is made to maximize. LDA is proposed by David Blei [2] originally, it is an unsupervised machine learning methods. It regarded the text as data information which constitute by “Bag-Of-Word”, and the model is built by word arisen frequency.
2. The proposed method

In this paper, the topic detection and tracking is divided into two sections, one is the “Train Section” and the other is “Test section.” The dialogue is composed of utterances those are with single or multi sentences. Here, Bag-Of-Word is used to be the main attributes, word break for sentence are essential for extracting the words within the utterance. The word breaks have many different results because everyone has a different view. For the unity, we will use the CKIP system which develops by Academia Sinica. The structure chart of model train sees Figure 1. In the LDA topic model build section, using the concept of the Bag-Of-Word to show the word frequency, in order to train the LDA topic model. In the SVM topic classification model train section, the text will as feature parameters and use the word frequency table to build feature vector, the topic classification model is trained by feature vector. These models will compare their performance each other. The Figure 2. Is LDA and SVM model test performance section, the preprocess is using CKIP to do word break too, then use this model to compare each topic model performance.

2.1. SVM Topic Classification Model

The SVM classification in this study is at baseline. In SVM classification model section, we use the libsvm [9] to train multi topic classification model, in select on the feature parameters, we use words as the feature parameters to training model, and record the number of times the word appears in the document to generate a feature vector. Some word is not important at topic relationship, so we choose the word base as feature parameters on experience. We will delete these words to get the better classification model. The kernel functions we use is Radial Basis Function (1).

$$K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2), \gamma > 0$$

In classification model, we build the model for each topic category. Each topic category regards as different category, and gives each one category a label, use these methods to train multi category topic classification model. In test section, input the test data in topic classification model, and classification results is based on output. This result is a baseline topic classification result.

2.2. LDA Topic Classification Model

LDA is proposed by David Blei [2], it is a statistical probability generation model, it and generate a random text which constitute by multi words. Through LDA model we can get the text distribution probability $P(w|z)$ in the topic, and the topic distribution probability $P(z|d)$ in the text. We will explain the LDA model by figure, see Figure 3.
above process, LDA model can be expressed in the form of joint distribution
\[
P(B, z, w)P(\theta|\alpha) \prod_{n=1}^{N} P(z_n|\theta) P(w_n|z_n, \theta) \quad (2)
\]
In this experiment, it uses GibbsLDA [7][10] to train the LDA model. Gibble Sampling is a special case in Markov chain Monte Carlo, when the joint distribution is not clear, the conditions of each variable distribution is known under. It can be randomly generated sample based on currently variable, and sequentially generated sample for the other distributed. After continuous iterations until the convergence of the parameters to be estimated, it can do joint probability distribution statistics. The speed is more than original LDA model, and suitable for analysis of large-scale documents. The first is Bag-Of-Word notation, we chose the word according to analysis of experimental corpus. The LDA lexicon builds section, we delete some word which without effect the topic, it is also according to experimental experience, and show per experimental corpus by Bag-Of-Word notation. The LDA model is built by four topics. Observing the topic category of test data, the output P (z|d) can get the topic of this document, and the P (w|z) can get the probability generated by word on every topic.

3. Discussion

This search is point on topic analysis of the interactive dialogue record. The experimental data are collecting the dialogue record in real life, and translate the record to text, see Table 1. We divide all of the dialogue record to four topics according to content analysis, each topic has 50 piece of dialogue record, total have thousands of sentences, see Table 2. Chinese has not auto word break like English, so we need to reprocess the data. In preprocess, we use the CKIP system to break the word, and, it can divide each sentence to multiple word, the result after process see in Table 3.

Table 2.
\[
\begin{array}{|c|c|c|}
\hline
\text{Dialogue Topic} & \text{Number of Document} & \text{Total Dialogue Sentence} \\
\hline
\text{Travel} & 56 & 2115 \\
\text{Diet} & 52 & 1515 \\
\text{Leisure} & 43 & 1500 \\
\text{Care} & 57 & 1793 \\
\hline
\text{Total} & 208 & 6923 \\
\hline
\end{array}
\]

Table 3.
\[
\begin{array}{|c|c|}
\hline
\text{Dialogue Record Word Break Process} \\
\hline
\text{Speaker A:} & \text{今年好冷} \\
\text{Speaker B:} & \text{那要不要去泡個溫泉啊} \\
\text{Speaker A:} & \text{去哪泡啊} \\
\text{Speaker B:} & \text{看你啊你要去北投還是關子嶺啊} \\
\text{Speaker A:} & \text{關子嶺好了啦比較近} \\
\text{Speaker B:} & \text{那你不想去台北玩玩噢我覺得台北比較精彩} \\
\hline
\end{array}
\]

In baseline classification model section, the cost parameter is 25 in SVM training model, the kernel function is RBF. The training data are 60% on dialogue record, 40% are test data. The result which uses the SVM see Table 4.

Table 4.
\[
\begin{array}{|c|c|c|}
\hline
\text{Topic} & \text{Correct Rate} & \text{Avg. Correct Rate} \\
\hline
\text{Travel} & 79.16\% & 75.07\% \\
\text{Diet} & 81.81\% & \\
\text{Leisure} & 52.38\% & \\
\text{Care} & 86.95\% & \\
\hline
\end{array}
\]

The results show, the correct rate of “Leisure” is the worst in these topics, and the “Care” is the best. Speculating the reason for the leisure words less obvious features, the topic of leisure is similar the topic of the travel, but the performance still best.

In LDA train model section, the train data are 60% on dialogue record, and the 40% is test data. The parameter alpha is 0.5, the beta is 0.1, the sampling iterations is 1000, the number of topic is 4, the lexicon has 3633 words. The number of topic is hidden parameter in LDA, so we need observation and define the topic by a model which had trained. After training, we can get words for each topic. These words can define the topic category, we define “Topic 1”is a diet, “Topic 2” is leisure, “Topic 3” is caring, and “Topic 4” is travel, see Table 5. These classify will used in the test section.
In the test section, the data we use is 40% dialogue record, and classify the topic by model. According the P(z|d) matrix, we can decide the text in which topic. Table 6 is classifying result for “Leisure” text, the “休閒” word isn’t, there usually have many superfluous words (e.g. “你好(How are you?)”, “哈囉(Hellow)”, “的(De)”, “不(Not)”), these words always confuse the topic classification, how to find the identical word is very important. The LDA model is better than SVM in building model section, especially the topic is not obvious, it can use the probability statistics to fix and guide the defined topic. The future work is how to determine the sentence topic in real time, how to find the identifiable word on LDA lexicon, and consider the structure of the sentence. Finally, build a real time topic classification system.

### 5. References


### 4. Conclusions

In this paper, we use the dialogue record to discuss the topic classification problem. In the past, most of the topic classification search is focused on science document or newspaper, because these texts has related to their topic. The dialogue record usually translate the topic or deviate the topic, event has the sentence which does not associate with the topic (e.g. Greetings, yes-no answer.) These sentences are confusing the topic and influence the identification results. The superfluous word and expetutive usually interferes the completeness of sentence topic. In this paper, LDA is used Bag-Of-Word, so it is important to choose the word, it often chooses the representative vocabulary for each topic. The science document and newspaper usually have many proper nouns, using these keywords can efficiently to build the topic model. But the dialogue isn’t, there usually have many superfluous words (e.g. “你好(How are you?)”, “哈囉(Hellow)”, “的(De)”, “不(Not)”), these words always confuse the topic classification, how to find the identical word is very important. The LDA model is better than SVM in building model section, especially the topic is not obvious, it can use the probability statistics to fix and guide the defined topic. The future work is how to determine the sentence topic in real time, how to find the identifiable word on LDA lexicon, and consider the structure of the sentence. Finally, build a real time topic classification system.