Extended LDA Based Topic Analysis for Big Sentiment Data

Shengzhen Xie¹, Ling Sui¹, Jiaxin Li¹, Kaiqi Kou¹ and Zhiqiong Wang²,*

¹School of Computer Science and Engineering, Northeastern University, Shenyang, China
²College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China

*Corresponding author email: wangzq@bmie.neu.edu.cn

Abstract: Owing to the massive emergence of multi-modal data related to big sentiment events, the topic analysis of the text cannot accurately reflect the topic distribution of the whole data set. There are some problems such as semantic deficiency and difficulty in integration. Therefore, the extended LDA (eLDA) based topic analysis for big sentiment data is posed. Firstly, the semantic analysis of text and image data is carried out by probability generation method respectively. The semantic correlation between multi-modal data is used for visual topic learning, a visual topic model is established, and the mapping between visual data and text topics is realized. Furthermore, the fusion of text, audio, video and image data is realized, and multi-modal theme analysis is carried out accordingly. The experimental results show that the multi-modal topic analysis method can effectively obtain the subject words related to the semantics of big sentiment data. The effect of the subject words tracked by the multi-modal data analysis method is better than that of the single text subject words extraction, which can provide basis for emotional analysis of big sentiment events.

Keywords: Theme analysis; Network sentiment; Multi-modal data; LDA.

1. Introduction

With the increasing popularity of the Internet, public opinion analysis has accumulated a large amount of heterogeneous and multi-modal big sentiment data. It is extremely urgent to study the topic analysis technology of multi-modal data. LSA¹ was originally used to mine semantic associations hidden between documents and words by decomposing singular values. On the basis of LSA, Hofmann proposed a probabilistic topic model PLSA² based on statistics, and effectively solved the problem of synonym and polysemy mixing by introducing the idea of probabilistic statistics model. However, PLSA has a series of deficiencies such as high complexity of algorithm. Aiming at these defects, Blei et al.³ proposed LDA topic model, which expanded PLSA by introducing a prior distribution of Dirichlet. Now it has become a widely used probabilistic topic model in many fields. Existing LDA topic models are all text-oriented, but images, videos and audio are also important components of network big sentiment data. It is necessary to adopt a topic analysis method based on eLDA to extract multi-modal data features from massive data sources to fully reflect the topics of current public opinion events. eLDA is proposed to extract the underlying and deep features of multi-modal data. A unified multi-modal topic description is obtained by combining image, audio and video data. The topic word extraction is realized based on the topic model of eLDA. The main research contributions are:

- From the angle of multi-modal data feature extraction, the features of different data types in the video are extracted, and the unified representation of multiple modal features is obtained, which is the basis for realizing video theme analysis.
• According to the image, audio, video and text media data existing in big sentiment data, the underlying features are extracted for topic analysis respectively to obtain the topic description of multimodal data information.
• Considering multi-modal data information, the extended LDA based topic analysis for big sentiment data is posed to realize topic analysis on the basis of traditional LDA.

Next, some methods of topic analysis combined with multi-modal data are introduced. Then, the basic idea and implementation method of eLDA are described. The proposed multi-modal data topic analysis method is experimented. Ultimately, the full text is prospected.

2. Related Work
The LDA topic model proposed by Blei et al.\[3\] discovers the hidden topic information in the text through unsupervised learning methods. Because its topic analysis effect is better than the traditional LSI and LSA models, it has been widely used. Based on the application of traditional LDA, there have been many improved methods. C. Moody introduced word vector method on the basis of LDA and proposed Lda2vec model\[6\]. Compared with traditional LDA, this method theoretically provides richer potential semantic information. However, according to the current research, compared with the traditional LDA, Lda2vec model has not significantly improved the effect of topic model in many applications. Many studies have put forward the paradox of Lda2vec model, Lda2vec model should achieve better results than classical LDA in theory, but its results are difficult to achieve expectations in practical application\[7, 8\]. Bao et al.\[4\] posed a partially supervised LDA model for cross-domain learning. Hu et al.\[10\] posed a model that can achieve topic modeling and event segmentation of events/tweets in a unified framework. However, these algorithms just premeditate the text info of the event and do not combine other modal information. In order to solve the problems, Putthividhy D\[5\] designed an improved multi-modal LDA using topic regression technology, which is used for the correlation between visual and text features of image and video annotation tasks.

3. Multi-modal Topic Learning
According to eLDA, a multi-modal data topic analysis framework is proposed indicating in Figure 1.

3.1. Visual Theme Learning Method

3.1.1 Semantic description of image data. In order to realize multi-modal topic analysis, feature extraction of image data are required. The image set \(I\) is preprocessed and SIFT features of each image \(I_i\) are extracted. It is assumed that N SIFT features are extracted from K images. Then K-means algorithm is used to cluster the extracted N SIFT features. There are k clustering centers and the length of the codebook is k. The distance between each SIFT feature of each image and the k visual words is calculated and mapped to the nearest visual word, and each image becomes a word frequency vector corresponding to the visual word sequence. Each cluster is regarded as a visual word \(\omega_i\), and the set of visual words forms a visual dictionary \(V = \{\omega_1, \omega_2, ..., \omega_k\}\), which can be used to describe the semantics of each image.
3.1.2 Construction of Visual Theme Model. The core of multi-modal topic analysis is to realize the mapping between different modal data topics\(^9\). In order to establish the topic correlation between text and image, a topic mapping framework based on visual topic learning is proposed, as indicating in Figure 2.

![Figure 2. Visual Topic Learning Framework.](image)

Assuming that there is a news set \( M \), in which text data is represented by \( d \), the number of hidden topics is represented by \( m \), image data is represented by \( I \), and the number of visual words formed by semantic description is \( k \), then this news set can be represented as \( M = \{I, d\} \). LDA is used to analyze the topic of the text \( d \) in the news set \( M = \{I, d\} \), and the topic probability distribution of the text is obtained, which contains \( m \) hidden topics: \( \Theta = \{\theta_1, \theta_2, ..., \theta_m\} \), so each news can be expressed as the probability distribution of \( m \) topics: \( d^t = \{p(\theta_1 | d), p(\theta_2 | d), ..., p(\theta_m | d)\} \). Feature extraction is carried out on the image \( I \), and \( I \) is described as the probability distribution of \( k \) visual words: \( I = \{p(o_1 | I), p(o_2 | I), ..., p(o_k | I)\} \). The topic-visual vocabulary relationship matrix \( Q \) is established through visual topic learning, and the text topic probability distribution \( d^t = \{p(\theta_1 | d), p(\theta_2 | d), ..., p(\theta_m | d)\} \) of the news \( d \) to which it belongs is set. Then the topic-visual word distribution generated by the \( t+1 \) iteration when calculating the topic-visual word distribution matrix is calculated by Equation 1.

\[
Q^{t+1}_{v,i} = Q^{t}_{v,i} + p(\theta_i | d) p(o_v | I, \theta_i)
\]

Repeating the training process for all images will obtain the visual word distribution matrix \( Q \), which shows the correlation probability of each visual word with each topic: \( Q_{v,i} = p(\theta_i | o_v) \). Assuming that the visual word distribution of the test image is: \( I = \{p(o_1 | I), p(o_2 | I), ..., p(o_k | I)\} \), the correlation probability of the image with each topic is calculated by describing the test image as a probability distribution \( I \) of \( k \) visual words and multiplying it with the topic-visual word distribution matrix \( Q \) to obtain a topic distribution vector of the test image, which is calculated by Equation 2.

\[
p(\theta_i | I) = \sum_{v} p(o_v | I) p(\theta_i | o_v)
\]

3.2. Audio topic learn method. In order to realize the topic analysis of audio data, speech recognition is proposed for audio. The audio is divided into lengths suitable for speech recognition. Each segment of audio after segmentation is kept for 1 minute to obtain an audio string \( L = \{l_1, l_2, ..., l_r\} \). The text representation \( d_v \) of the audio data is obtained by speech recognition of the audio string. Input it into the trained LDA, the topic probability distribution \( d_v = \{p(\theta_1 | d_v), p(\theta_2 | d_v), ..., p(\theta_m | d_v)\} \) belonging to the audio segment can be obtained.

3.3. Video Topic Learning Method

3.3.1 Semantic description of video data. Different from single-modal image data, video is complex multi-modal data. In eLDA, video semantics is represented by extracting key frames and audio streams to convert them into image data and audio data, and video data is processed. Assuming that there is a video \( V \), the frame with the local maximum of the average inter-frame difference strength is selected as the key frame \( V_k \) of the video based on the inter-frame difference method, and the video is converted into image data: \( V_k = \{V_k, V_k, ..., V_k\} \). The audio contained in the video is extracted and split into lengths
suitable for speech recognition: $v_i = \{v_{i_1}, v_{i_2}, ..., v_{i_s}\}$, so the video is represented as image data and audio data set: $\mathcal{V} = \{v_{i}, V_{i}\}$.

3.3.2 Video data topic analysis. In order to realize the topic analysis of video data, theme analysis is carried out on the image data and audio data set extracted from the video. Assuming that the video image set $V_i$ contains the number of key frames $v$ and the audio segment contains the number of audio strings $s$. The visual word distribution of the image is expressed as $\mathcal{V}_i = \{p(\theta_{i_1} | v_{i_1}), p(\theta_{i_2} | v_{i_2}), ..., p(\theta_{i_v} | v_{i_v})\}$, then the correlation probability between the image and each topic is calculated by describing the test image as a probability distribution of $k$ visual words. Multiplying the trained topic-visual word distribution matrix $Q$, the topic distribution vector of the test image is expressed as $p(\theta | v_{i}) = \sum_{\theta} p(\theta | v_{i})$. Then the subject probability distribution of the image set of the video is $p(\theta | V_{i}) = \sum_{\theta} p(\theta | V_{i})$. The audio string after the video processing is $L = \{l_{i_1}, l_{i_2}, ..., l_{i_s}\}$, and the text representation $d_{i_v}$ of the audio data is obtained by voice recognition. Input this text into the trained LDA theme model to obtain the theme probability distribution of the audio segment $v_{i}$ of the video: $d_{i_v} = \{p(\theta_{i_1} | d_{i_1}), p(\theta_{i_2} | d_{i_2}), ..., p(\theta_{i_v} | d_{i_v})\}$, so the theme distribution to which the video belongs is $d_{i_v} + p(\theta | v_{i})$.

4. Experimental Results and Analysis

According to the topic analysis algorithm eLDA, relevant report information of big sentiment data from October 2019 to June 2020 was collected. The multi-modal training data set involved 4999 news and the test data set involved 1200 news. After processing this test data set, some experimental results obtained from three representative news terms (News 1, News 2 and News 3) are selected for display.

4.1. Experimental Results of Visual Topic Model

LDA method is used to analyze the topic of the training data set, and the topic probability distribution of the text is obtained. By extracting SIFT features of image data, visual words are generated and visual word bags are established. Combining with text topics generated by trained LDA topic models, a topic-visual word matrix is generated. In this experiment, the number of generated visual words is 500 and the number of hidden topics in text is 10.

The test image is calculated according to Equation (1) and the generated topic-visual word matrix to obtain the probability distribution of the image on each topic. As can be indicated from Table 1 that eLDA can better obtain the theme of the image.

4.2. Experimental Results of Audio Data Topic Analysis

The audio data set of the multi-modal data set is processed by speech recognition, and the audio is converted into text for topic analysis. The analysis results of audio topics are indicating in Table 2.
Table 2. Part Test Audio and Topic Matching Degree.

| Audio | Theme Distribution | Subject Word |
|-------|--------------------|--------------|
|       | Topic (Distribution Value) | |
| News 2 Audio | Topic 2(0.327) | Examinations, candidates, CET-4 and CET-6 |
|          | Topic 1(0.267) | Functions, Pixels, Shooting |
|          | Topic 4(0.193) | Funds, companies, investments |
| News 3 Audio | Topic 2(0.852) | Examinations, candidates, CET-4 and CET-6 |
|          | Topic 8(0.065) | Game, team, heat |
|          | Topic 6(0.033) | Fashion, Match, Group Picture |

4.3. Analysis of Experimental Results

In order to measure the experimental effect of eLDA, precision P, recall R and F-measure F are used as the evaluation criteria for the extraction effect of subject words. This experiment adopts multiple groups of manual crossover methods to mark each news data to extract keywords so as to drop the deviation of personal subjective factors on the extraction results. The calculation formulas of P, R, F are:

\[
P = \frac{|W_1 \cap W_2|}{|W_1|} \times 100\%, \quad R = \frac{|W_1 \cap W_2|}{|W_2|} \times 100\%, \quad F = \frac{2PR}{P+R},
\]

where \(W_1\) represents the correctly extracted keyword set, \(W\) represents the extracted keyword set, \(W_2\) represents the manually labeled keyword set. Firstly, a comparative experiment is carried out by combining the text and different modal data in turn and the average results of P, R and F obtained after calculation are indicating in Table 3.

Table 3. Part Test Audio and Topic Matching Degree.

|            | Precision | Recall | F-measure |
|------------|-----------|--------|-----------|
| Text       | 0.827     | 0.726  | 0.773     |
| Trxt+Image | 0.835     | 0.735  | 0.782     |
| Text+Audio | 0.835     | 0.726  | 0.777     |
| Text+Video | 0.833     | 0.732  | 0.779     |
| eLDA       | 0.852     | 0.741  | 0.793     |

eLDA is compared with TF-IDF and TextRank so as to verify the feasibility of eLDA. The average results of P, R, F and running time obtained after subject analysis of the test data set are indicating in Table 4. It can be seen that eLDA is better than TF-IDF and TextRank.

Table 4. Comparative Experimental Results of Different Extraction Algorithms.

|            | Precision | Recall | F-measure | Time(min) |
|------------|-----------|--------|-----------|-----------|
| TF-IDF     | 0.771     | 0.748  | 0.759     | 58.321    |
| TextRank   | 0.748     | 0.711  | 0.729     | 28.025    |
| eLDA       | 0.852     | 0.741  | 0.793     | 27.218    |

TF-IDF and TextRank only use the text data, ignoring other modal data information, while eLDA integrates image, audio, video information into the topic analysis process. Figure 3-6 shows the comparison chart of the P, R, F and running time results under the number of articles ranging from 200 to 1200. With the increase of the number of test documents, the effect of eLDA is obviously improved in P, R, F and the running time is shortened compared with the other two algorithms.

Figure 3. Comparison of Precision.  Figure 4. Comparison of Recall.
5. Summary and Prospect

This paper proposes a topic analysis method based on eLDA. Experimental results show that eLDA can map semantically related images, video information and event text in big sentiment data to the same topic. eLDA can realize topic analysis of multi-modal data contained in big sentiment data, especially semantic analysis using visual topic modeling method, which increases the influence of image and video modal data on subject words, and improves the accuracy compared with a single text topic analysis method.

Acknowledgments

This research was partially supported by the CETC Joint Fund, the National Natural Science Foundation of China (No. 62072089), the China Postdoctoral Science Foundation (Nos. 2019T120216 and 2018M641705), the Fundamental Research Funds for the Central Universities (Nos. N2019007, N180101028, N180408019 and N2024005-2) and the fund of Acoustics Science and Technology Laboratory.

References

[1] Landauer T K, Foltz P W, Laham D. An introduction to latent semantic analysis[J]. Discourse Processes. 1998, 25(2): 259-284.
[2] T Hofmann. Probabilistic latent semantic indexing[C].Proceedings of the 22nd Annual international ACM SIGIR conference on Research and development in information retrieval 1999: 50-57.
[3] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003(3): 993-1022.
[4] Bao Y, Collier N, Datta A. A partially supervised cross-collection topic model for cross-domain text classification[C]//Proceedings of the 22nd ACM international conference on Information & Knowledge Management. ACM, 2013: 239-248.
[5] Putthividhy D, Attias H T, Nagarajan S S. Topic regression multi-modal latent dirichlet Allocation for image annotation[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010: 3408-3415.
[6] Moody C E. Mixing Dirichlet topic models and word embeddings to make lda2vec[J]. arXiv preprint arXiv:1605.02019, 2016.
[7] Chen P, Guo W, Dai L, et al. Pseudo-supervised approach for text clustering base on consensus analysis[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 6184-6188.
[8] Onan A. Topic-Enriched Word Embeddings for Sarcasm Identification[C]//Computer Science On-line Conference. Springer, Cham, 2019: 293-304.
[9] Gao B, Liu TY, Qin T. Web Image Clustering by Consistent Utilization of Visual Features and Surrounding Texts[C]. Proceedings of the 13th Annual ACM International Conference on Multimedia, 2005: 112-121.
[10] Hu Y, John A, Wang F, et al. Et-lda: Joint topic modeling for aligning events and their twitter feedback[C]//Twenty-Sixth AAAI Conference on Artificial Intelligence. 2012, pp. 59-65.