Electronic Government and Electronic Participation
E. Tambouris et al. (Eds.)

doi:10.3233/978-1-61499-570-8-72

Constructing a Methodology Toward Policy Analysts for Understanding Online Public Opinions: A Probabilistic Topic Modeling Approach

Nan ZHANG[a] and Baojun MA[b]
[a] *School of Public Policy and Management, Tsinghua University, Beijing 100084, P. R. China*
[b] *School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, P. R. China*

Abstract. Public opinion always has an important influence on the policy process. The development of social networking sites and applications has given the public more opportunities to express their views about the related policies. In cases where the coverage of the traditional hearing system challenged the policy process, how to measure accurately the public concern and attitudes regarding policies based on online public generated content using a data mining method will be very important issue in policy informatics. Our paper provides a probabilistic topic modeling approach, mainly based on Latent Dirichlet Allocation (LDA) model, to transform the complex semanteme of online public opinions into the values could be measured. A simple case could show the usefulness of the too toward policy analysts also be provided and discussed briefly.

Keywords. Online Public Opinions, Public generated content, Probabilistic Topic Modeling, Latent Dirichlet Allocation (LDA), Policy Informatics.

1. Introduction

The media, public and decision makers are considered the three primary factors in the traditional policy agenda settings [1]. The prior research believes that these three factors are independent of each other and mutually affect each other [2]. In recent years, the rapid development of Web 2.0 applications, such as Facebook, Twitter and Wikipedia, has provided citizens with more approaches to participate in public policy discussions more easily. Ordinary people have the opportunity to express their opinions and exert an unprecedented influence as long as their viewpoints are typical and comprehensive. It seems likely that ordinary people, who lack discourse power in real society, will become the opinion leaders on the new media overnight. Hence, the age in which we live now is called the 'we-media era' by scholars in the communications field [3], which means that the roles of the public and media are deeply integrated in policy agenda settings. Therefore, the capability of the professional media, such as newspapers and TV, to initiate topics for discussion and lead public opinion may be further weakened while, in contrast, public expression will play an increasingly important role in the process of making and revising policies.

These new characteristics of the age mentioned above have undoubtedly led to more challenges associated with obtaining public attitudes and feedback as was the case during the traditional policy process. The hearing system that was generally adopted before the policy promulgation is being increasingly questioned [4]. It seems that, no matter how the participants of the hearing are selected, it is impossible to satisfy everyone, which constitutes an unavoidable hindrance to the representative system in such an increasingly flat world [5]. Moreover, even online opinion surveys, that emerged with the popularity of the network, face challenges: those stakeholders directly benefiting from certain policies or the group or individuals who pay more attention to certain issues usually perform in the manner of voting several times in support of their own views, or inviting friends to vote in support of those views to exaggerate the scale of their own side which, to a certain degree, is misleading to policy makers. However, as the social network has become a crucial way for the public to express their attitudes and exchange ideas, it is possible to find a reliable method to determine public attitudes and examine the mass view in order to create policy from the fragmented expression of opinions that are to be found via various social network applications?

The direct approach that has been applied is public opinion analysis, based on keyword searches and filtering [6]. Obviously, this type of method may only describe the problems roughly, leading to a boom in related studies in the field of communication, and it seems that the method cannot really support the policy process at present. The biggest challenge originates from the definition of “data” and the expansion of the processing method, and a deeper analysis of public generated content on social media must surmount the boundary of Codd (1982)’s relational database [7]. The progressive data mining methods, including the semantic analytical methods, have made it possible to analyze these unconventional data. The related analysis contributes not only to investigating the degree of public support and practical effects made during the phase of implementing policies, but also the collecting of wider public opinion during the policy-making stage to engender more extensive public participation in the policy-making process. The decision makers need to make political responses rapidly by employing semantic analysis and other methods to learn about the public attitudes towards various policy agendas on the Internet quickly and accurately, which should also be one of the issues considered and addressed by the new research branch, policy informatics.

Under such a context, this research attempts to construct, based on Latent Dirichlet Allocation (LDA) [8], an effective approach to measuring the popularity and polarity of the discussion by means of refining related subjects from a large amount of web text. The method we present is explored by taking, as an example, public attitudes drawn from an online community regarding the car license lottery policy promulgated at the end of 2010 in Beijing, China.

2. Framework of the Methodology

2.1. Briefly Review of Topic Modeling and Latent Dirichlet allocation

Topic modeling algorithms are statistical methods that analyze the words of an original text to discover the latent themes or topics that run through it, originally proposed and applied in the field of information retrieval [9].

Without directly dealing with the topics inherent in the documents, the TF-IDF schema and vector space model (VSM) provided a rough solution to describing and modeling documents and their content or topic similarities [10, 11], with the disadvantage that documents with a similar context but different vocabulary will not be associated with each other [12]. To deal with this problem, latent semantic analysis (LSA) has been proposed to convert the high-dimensionality word-space representations of the documents into low-dimensionality vectors of the topics [13], in which topics can be obtained using singular-value decomposition (SVD). Using such a technique yielded some improvement over a TF-IDF baseline.

A closely related technique−pLSI [14] −tries to create a set of topics in a probabilistic framework [15]. The topics in pLSA are probabilistic instead of the heuristic geometric distances in LSA. pLSA has been successfully used with large collections regarding information retrieval, because its does not need to run the expensive SVD operation.

Latent Dirichlet allocation (LDA) is a popular topic modeling tool, designed to learn a set of topics (word distributions) and infer mixtures of these topics to build low-dimensionality representations of documents [9], which further refines the pLSI model within a Bayesian framework [14]. Among all of these topic models, LDA appears to be the most effective [8, 16]. A simple introduction to the LDA model will be presented in more detail in the following section.

The intuition behind the LDA model is that documents exhibit multiple topics [9]. LDA generates ''topics'' as lists of words drawn from the vocabulary used in the text corpus; the topic is based on the distributions of those words over the vocabulary. The topics are generated inductively by the model based on the likelihood of words to co-occur within documents. LDA produces the topics through a probabilistic approximation of Bayesian inference. Starting with a set of seed topics (often randomly generated), the algorithm iteratively alters these topics to best match the set of data being learned. LDA also generates proportions for each document for each topic, so that each document can be described as being proportionally composed of (or, interpretively speaking, "about'') a number of topics that are expressed by the words used in that document. For details about the algorithmic and computational aspects of LDA, see [8].

Topic modeling is a good match for a data source like public comments. Because the procedure is automated, it enables analysis of much larger text corpora than would be feasible by hand. Since topics are generated inductively by the model and not predefined by the researcher, the technique protects against implicit coding bias caused by the constraints of researcher knowledge. And because the method assumes that a single document can contain multiple topics (in contrast to some other document clustering methods; see [17]), it enables researchers to draw insights from interrelations among themes, both within documents and within the dataset as a whole.

As described, public comments are rich data because commenters express all manner of concerns in unstructured ways. Comments run the gamut from technical specifications to personal stories and from thoughtful reflection to threats and name-calling. Many comments defy easy categorization as being for or against the proposal at issue. Thus, hand-coding such documents can be a particularly difficult task; topic modeling appeals because it can uncover hidden patterns in even a varied set of documents.

Moreover, it is noted that the LDA model does not require any prior annotations or labeling of the documents and the topics emerge from the analysis of the original texts.

The LDA model enables us to organize and summarize electronic archives on a scale that would be impossible using human annotation [9].

2.2. Overall Consideration and Detailed Procedures

Given the increasing amount of content about public policy feedback being available online in the Web 2.0 era, especially via blogs, microblogs, Wikipedia and social networks, the policy makers or government do not have the time to read and study all of them to learn about the detailed contents and viewpoints of the public. In addition, due to the characteristic of being an online public communication, the topics or themes of public discussion tend to be diverse and may evolve, led by someone with time, even about a certain event or general topic. Thus, the later public discussion may deviate from the original topic, and the total content may contain many issues that are irrelevant to the policy makers or government.

To this end, we have designed a methodology framework for public policy feedback monitoring by applying a probabilistic topic modeling algorithm, specifically Latent Dirichlet Allocation (LDA) [8], which will be introduced in detail in the next subsection. In the framework, the LDA model has been utilized to discover and annotate large archives of public-generated documents with thematic information to determine how these themes are connected to each other and how they change over time. Thus, we could provide the policy makers or government with filtered, accurate content that they are actually concerned about.

We will introduce the detailed procedures of public policy feedback monitoring by applying the LDA modeling. The steps involved in probabilistic topic modeling and concerned content filtering are as follows:

- Step 1: Data collection and pre-processing.
- Step 2: Probabilistic topic modeling using the LDA model.
- Step 3: Choosing interested or concerned public policy topics.
- Step 4: Document topic assignment and concerned content filtering.
- Step 5: Concerned document hotness and relevance calculation.

Firstly, the original relevant data need to be collected from the online platform based on certain events, topics, persons or contents in which the policy makers are interested, which are normally conducted by web crawlers or programs [12, 18, 19]. Then, in order to explore the text content information further, several necessary pre-processing operations must be performed, such as word segmentation [20, 21], stemming [22, 23] and stop-word removal [18, 19]. Thereafter, each text's content is represented by a vector of words.

In Step 2, the core task lies in using the LDA to conduct probabilistic topic modeling. Specifically, LDA modeling applies training and inference to all of the text vectors and can discover any latent topics or themes inherent in these data [9], as will be discussed in the next subsection in greater detail. After the topic modeling, we obtain the following useful results related to the words, documents and topics in these documents:

- Latent topics with the most likely words in each topic;

- A Topic-document distribution matrix.

Theoretically, for any given topic, there is a corresponding distribution across all of the words in the vocabulary, and the LDA modeling process could provide the most likely words with the highest probability with respect to each topic, whereby the number of the most likely words can be given in advance. After the LDA modeling, the thematic keywords based on the given topic number become available to policy makers, and we could invite them to select topics of interest to them. Thus, this selection step is intuitive, easy to implement, and would not impose much time and cost on the policy makers. More conservatively, the topics or themes of concern to the policy makers or government could be easily predicted or learnt in advance.

In Step 4, we first utilize the topic-document distribution matrix to conduct topic assignment for each text document. In detail, the topic or topics with the highest document-topic probability will be assigned to each document. More flexibly, the topics with the top t highest document-topic probabilities could be simultaneously assigned to each document, with 1<t<K based on either the preference of the policy makers or the document number to be filtered. By combining the topics of concern selected in Step 3 and the topic assignment results together, documents about irrelevant topics could be easily filtered out.

Finally, after we obtain all of the contents of interest regarding public policy feedback, some related statistical information, such as the documents' hotness and cumulative relevance ratio per day, could be calculated and provided for the policy makers. Specifically, to show the discussion hotness of these topics each day, the cumulative document number concerned can be easily obtained. In the meantime, in order to evaluate the overall relevance of the documents concerned with respect to the topics concerned, the daily cumulative relevance and cumulative relevance ratio can be calculated.

3. Case and Discussion

3.1. Case Background and Analysis Results

The government of Beijing promulgated the policy of a car license plate lottery at the end of 2010. In our research, we selected the forum named "AutoWorld" in Shuimu Community as the targeted concern. AutoWorld forum is one of the most active forums in Shuimu Community, especially around and after the time when the lottery policy for buying vehicles in Beijing City was announced on December 23th, 2010. Since data in the recycling box can only be kept for around four or five months, in order to obtain all of the data from the AutoWorld forum about this event, the time window was chosen of around three months, which was specifically from December 15th, 2010 (i.e., around a week before the policy announcement) to April 5th, 2011. After removing the duplicate posts that appeared in both the forum and the recycling box, such as the collection post information, we finally obtained a total of 359,715 unique posts.

We conducted probabilistic topic modeling for all 359,715 post texts using the LDA model with the topic number 25, 50, 75, 100, 125, 150, 175 and 200 respectively. As we discussed above, for any given topic, there is a corresponding distribution across all of the words in the vocabulary, and the LDA modeling process could provide the most likely words with the highest probabilities with respect to each topic, in which the number of the most likely words can be given in advance. In our case, we set this number at 100. By manually looking through all 100 thematic keywords for different topic numbers, it was found that the semantic results for the thematic keywords for each topic appear best when the topic number is 50.

In this case, we assume that the policy makers and government are notably interested in public feedback on the lottery policy of car license plates in Beijing City, as introduced in the background section. Thus, public discussion of topic 10 (i.e., the public countermeasure to the lottery policy) and topic 27 (i.e., the lottery policy) were chosen as the topics of concern or interest in our case. Then, through the procedure of document topic assignment (the highest probability strategy) and concerned content filtering, we finally obtain over 23,074 related posts, with 9,543 about the lottery policy itself and 14,146 about the public countermeasure. Figure 1 represents the daily discussion hotness of the policy and public countermeasure based on our proposed LDA-based framework, showing the important policy time points.

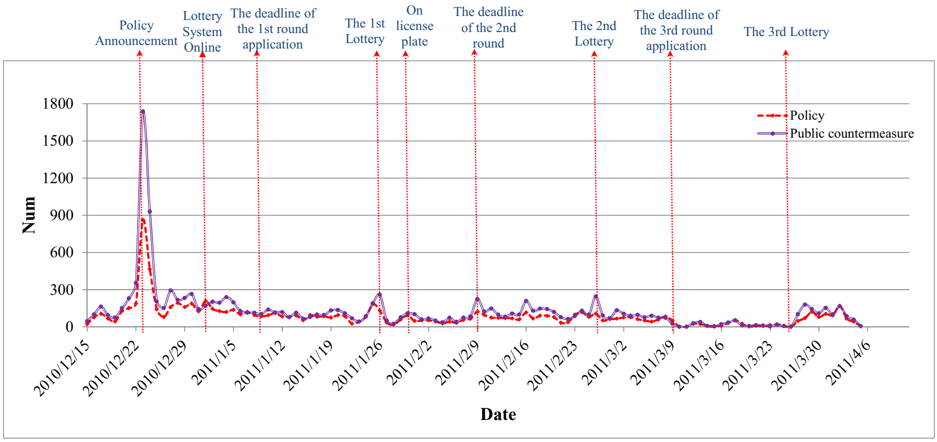

Figure 1. Comparison of daily discussion hotness of policy and public countermeasure.

Moreover, we also compared our results based on LDA modeling with those from the keywords filtering strategy, in which only containing corresponding policy terms, such as “摇号” (i.e., lottery) or “限购” (i.e., quota) in Chinese, would be retained. In detail, if any post contains at least one of the above policy terms, its original post and all of its reply posts would be regarded as potential content of interest, which resulted in 31,628 posts in total. Figure 2 shows the results of the comparison between our method and the keywords filtering method in relation to important policy time points. It is obvious that, compared with our LDA-based method, the keywords filtering strategy usually suffers from the disadvantage of containing too much irrelevant information, ignoring topic transfer as well as containing words that do not discuss the real topic, as explained above.

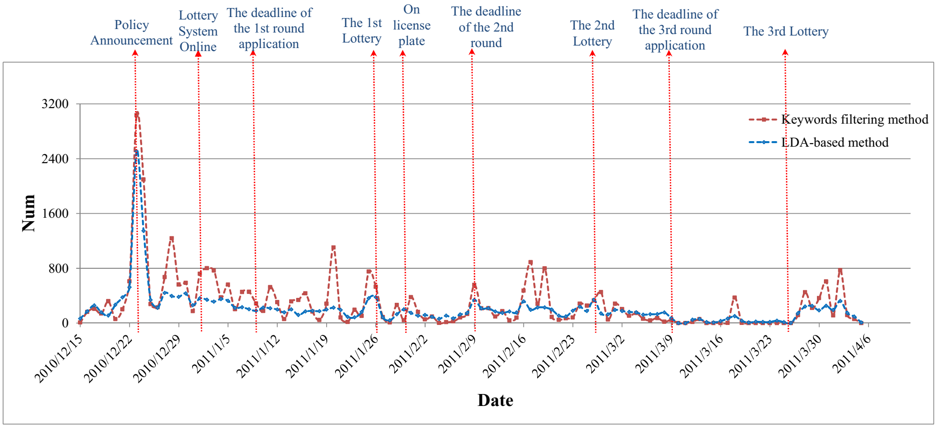

Figure 2. Comparison of daily discussion hotness by LDA-based method and keywords filtering method.

3.2. Concluding Remarks

The results of the case indicate that it is possible to obtain more accurately contents of interest from the mass of information by using the method based on LDA and be in a position to be better able to describe public concerns, and their types and polarity, related to the policy process via the related analysis. The core idea of the proposed methodology is transform the complex semanteme a relation matrix between texts and topics. After the transformation, we could construct many functions based on matrix values from different perspectives. This case only discusses one of the simplest functions and describes the trend of the hot topic. However, we could understand deeper policy related issues based on function construction.

In the age of large-scale data, a large amount of web text information that rapidly develops every day should explicitly serve as a treasure trove to be mined by policy analysis science, which will necessarily require more data mining methods and applications for policy analysis. The process undoubtedly requires that scholars of data mining and policy analysis should cooperate with each other, as in the case of the paper. It is expected that the method explored in this paper and its implications will facilitate both academia and practitioners to gain a better understanding of the opportunities and challenges existing in this rising research field.

Acknowledgements

This work was partly supported by the National Natural Science Foundation of China (71473143, 71402007, 71102010), the Beijing Social Science Foundation (15JGA008), the Tsinghua University Initiative Scientific Research Program (20131089260), and the Fundamental Research Funds for the Central Universities (2014RC0601).

References

[1] Lazarsfeld, P. F., & Merton, R. K. (1971). Mass communication, popular taste and organized social action. In Media Studies: A Reader (2nd ed., pp. 18-30). Edinburgh: Edinburgh University Press.

[2] Wanta, W., & Hu, Y.-W. (1993). The agenda-setting effects of international news coverage: An examination of differing news frames. International Journal of Public Opinion Research, 5(3), 250-264.

[3] Cottle, S. (2006). Mediatized rituals: Beyond manufacturing consent. Media, Culture & Society, 28(3), 411-432.

[4] Leighninger, M. (2012). Public Hearings≠ Public Values. Public Administration Review, 72(5), 708-709.

[5] Freidman, T. (2005). The world is flat. New York: Farrar, Straus and Giroux.

[6] Wojcieszak, M. (2011). Pulling Toward or Pulling Away: Deliberation, Disagreement, and Opinion Extremity in Political Participation*. Social Science Quarterly, 92(1), 207-225.

[7] Codd, E. F. (1982). Relational database: a practical foundation for productivity. Communications of the ACM, 25(2), 109-117.

[8] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. The Journal of Machine Learning Research, 3, 993-1022.

[9] Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.

[10] Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28(1), 11-21.

[11] Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communication of the ACM, 18(11), 613-620.

[12] Liu, B. (2007). Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data: Springer Berlin Heidelberg.

[13] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407.

[14] Hofmann, T. (1999b). Probabilistic latent semantic indexing, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 50-57). Berkeley, California, USA: ACM.

[15] Papadimitriou, C. H., Tamaki, H., Raghavan, P., & Vempala, S. (1998). Latent semantic indexing: a probabilistic analysis, Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems (pp. 159-168). Seattle, Washington, USA: ACM.

[16] Blei, D. M., & Lafferty, J. D. (2009). Topic models. In A. Srivastava & M. Sahami (Eds.), Text Mining: Classification, Clustering, and Applications (pp. 71-93). Boca Raton, FL: Taylor & Francis Group.

[17] Carterette, B., & Chandar, P. (2009). Probabilistic models of ranking novel documents for faceted topic retrieval, Proceeding of the 18th ACM conference on Information and knowledge management (pp. 1287-1296). Hong Kong, China: ACM.

[18] Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval: Addison-Wesley New York.

[19] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge, England: Cambridge University Press.

[20] Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-Driven Computations in Speech Processing. Science, 298(5593), 604-607.

[21] Yang, C. D. (2002). Knowledge and learning in natural language. New York: Oxford University Press.

[22] Hull, D. A. (1996). Stemming algorithms: A case study for detailed evaluation. Journal of the American Society for Information Science, 47(1), 70-84.

[23] Porter, M. F. (1980). An algorithm for suffix stripping. Program, 14(3), 130-137.