Osong Public Health and Research Perspectives

Journal homepage: <http://www.kcdcpkhp.org>

Editorial

Topic Modeling

Hae-Wol Cho ^{a,b,*}

^a Osong Public Health and Research Perspectives, Korea Centers for Disease Control and Prevention, Cheongju, Korea

^b College of Medicine, Eulji University, Daejeon, Korea

<https://doi.org/10.24171/j.phrp.2019.10.3.01>
pISSN 2210-9099 eISSN 2233-6052

A topic model is a type of statistical model to determine abstract “topics” that occur in a collection of documents. Topic modeling is a frequently used text-mining tool employed in the discovery of hidden semantic structure in a body of text. Intuitively, given that a document is about a particular topic, it would be expected that certain words would appear in the document more or less frequently. A document typically is concerned with multiple topics in different proportions thus, in a document that is 10% about “A” and 90% about “B”, there would probably be about 9 times more “B” words than “A” words. The “topics” returned by topic modeling techniques are clusters of similar words. A topic model captures this notion in a mathematical framework, which allows the examination of a set of documents, based on the statistics of the words used in each document, what the topics might be, and what each document’s balance of topics is.

Topic models are also described as probabilistic topic models, which refers to statistical algorithms used for the discovery of latent semantic structure in an extensive body of text. In this age of information, understanding large collections of unstructured bodies of text is simply beyond an individual’s processing capacity due to the amount of written material encountered each day. Topic models can help to organize written material and offer insight. Originally developed as a text-mining tool, topic models have been used to detect “instructive” structure in data such as genetic information, images, and networks. They also have applications in other fields such as bioinformatics [1].

Topic modeling is a type of big data analysis methodology for discovering abstract topics that repeatedly occur in a collection of documents. When writing an article, an author has a specific

keyword in mind, this keyword is repeated throughout the article. The collection of keywords is modeled as a finite mixture over an underlying set of topic probabilities, and then a latent topic in a specific document is returned [2].

The web scraping method has recently been developed where data mining is employed to automatically collect and analyze large-scale web documents. There are many open-source libraries provided for data mining, thus making it possible to analyze big data with only a small amount of coding. Text mining is not only a good method for identifying the structure of a text and extracting concepts, but is also useful for visualization [3]. Text mining is being used to determine trends in journals, social network services such as twitter and blogs, customer types (through online reviews), and the discourse of big data in news outlets [4-8].

In the current issue of Osong Public Health and Research Perspectives, Cho et al examined the flow of topics in the field of women’s health, by analyzing news articles on healthcare using topic modeling [9]. The data obtained by preprocessing were analyzed using the Latent Dirichlet Allocation algorithm and major topics from the news articles were extracted.

Chronological analysis demonstrated that most of the articles were focused on “Healthcare,” while the other 9 topics represented only a relatively small proportion of the information from news articles between the years 1993 to 2000. During 2001 to 2005, most of the articles were focused on “Medical Service” and “Dietary supplement,” whilst some prominent topics were magnified by public interest, compared to those focused on Healthcare in the 1990s. It was observed that the difference in the proportion of each topic was small after 2011.

*Corresponding author: Hae-Wol Cho
College of Medicine, Eulji University, Daejeon, Korea
E-mail: hwcho@eulji.ac.kr

The authors concluded that changes in topics related to disease were not clearly distinguished in the 1990s. Since the 2010s, it was possible to verify that the articles related to “women disease” were more prevalent and articles had various topics of disease.

References

- [1] Blei D. Probabilistic topic models. *Commun ACM* 2012;55(4):77-84.
- [2] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *J Mach Learn Res* 2003;3(Jan):993-1022.
- [3] Paranyushkin D. Visualization of text's polysingularity using network analysis. *Prototype Letters* 2011;2(3):256-78.
- [4] Cho Y, Fu P, Wu C. Popular research topics in marketing journals, 1995–2014. *J Interact Mark* 2017;40:52–72.
- [5] Scanfeld D, Scanfeld V, Larson EL. Dissemination of health information through social networks: Twitter and antibiotics. *Am J Infect Control* 2010;38(3):182-8.
- [6] Michelson M, Macskassy SA. Discovering users' topics of interest on twitter: A first look. *Proceedings of the fourth workshop on analytics for noisy unstructured text data*. p.73-80.
- [7] Chen R, Xu W. The determinants of online customer ratings: A combined domain ontology and topic text analytics approach. *Electron. Commer Res* 2017;17(1):31-50.
- [8] Qiao Z, Zhang X, Zhou M, et al. A domain oriented LDA model for mining product defects from online customer reviews. *Proceedings of the 50th Hawaii International Conference on System Sciences*; 2017; Hawaii. 2017. p.1821-30.
- [9] Cho KW, Kim SY, Woo YW. Analysis of women's health online news articles using topic modeling. *Osong Public Health Res Perspect* 2019;10(3):158-69.