Fuzzy Approach Topic Discovery in Health and Medical Corpora

Amir Karami, Aryya Gangopadhyay, Bin Zhou, Hadi Kharrazi, Deval Shaileshkumar Mali

October 23, 2022

1 The main idea

In the research authors had described fuzzy latent semantic analysis (FLSA), a novel approach in topic modeling using fuzzy perspective. FLSA can handle health & medical corpora redundancy issue and provides a new method to estimate the number of topics. The quantitative evaluations show that FLSA produces superior performance and features to latent Dirichlet allocation (LDA), the most popular topic model. There is a growing need to analyze large collections of electronic documents. Moreover, very large-scale scientific data management and analysis is one of the data intensive challenges identified by National Science Foundation (NSF) as an area for future study.

2 The methodology

Each file in data is related to one twitter account of a news agency. For example, bbchealth.txt is related to BBC health news. Each line contains tweet id—date and time—tweet. The separator is '—'. This text data has been used to evaluate the performance of topic models on short text data. However, it can be used for other tasks such as clustering. They leveraged five available health and medical datasets in this research.

The researchers have proposed Fuzzy Latent Semantic Analysis (FLSA) model for health and medical text mining. This model shows better performance in both redundant and non-redundant document and can help topic models estimating number of topics in corpus. The remainder of this paper is organized as follows. In the related work section, we review the related research. In the methodology section, we provide more details for FLSA. An empirical study was conducted to verify the effectiveness of FLSA. Finally, we provide an illustrative example for FLSA, and present a summary and future directions in the last two sections. They described fuzzy latent semantic analysis (FLSA), for uncovering latent semantic features from text documents. FLSA treats fuzzy view as a new approach in topic modeling and will be validated through a series of experiments, conducted on health and medical text data.

The traditional reasoning has precise character that is yes-or-no (true-or-false) rather than more-or-less. Fuzzy logic added a new extension to move from the classical logic, 0 or 1, to the truth values between zero and one, [0,1]) The main goal of fuzzy models is to formulate uncertainty for applications such as decision-making [32, 28]. For example, a voter decides to select some candidates among a set of candidates in an election. The voter has different preferences in terms of economic, foreign policy, health, etc. Based on the preferences, the distance between each candidate's plans and the voter's preferences can be changed. These preferences can be formulated and measured in fuzzy clustering with µ, degree of membership.

3 The results

The vast array of health and medical text data represents a valuable resource that can be analyzed to advance state-of-the-art medicine and health. Large electronic health and medical archives such as PubMed provide an extremely useful service to the scholarly community. However, the needs of readers go beyond a simple keyword search. Topic modeling is one of the popular unsupervised methods to automatically discover a hidden thematic structure in a large collection of unstructured health and medical documents. This discovered structure facilitates browsing, searching, and summarizing the collection. Existing techniques of topic modeling are based on two main approaches: linear algebra and statistical distributions; however, this paper proposes FLSA to utilize fuzzy perspective for disclosing latent semantic features of health and medical text data. FLSA is a new competitor to the established topic models such as LDA and has the flexibility to work with a wide range of dimension reduction and fuzzy clustering techniques. FLSA also works with both discrete and continuous data, estimates the optimum number of topics, and avoids the negative effect of the redundancy issue in health and medical corpora.

4 Recommendation

They would develop dynamic and hierarchal topic models using fuzzy perspective. In addition, FLSA should be applied on social media data to track public opinions and would be used for online review and SMS spam detection.