A Detailed Survey on Topic Modeling for Document and Short Text Data

International Journal of Computer Applications
Foundation of Computer Science (FCS), NY, USA

Volume 178
Number 39

Year of Publication: 2019

Authors:
S. Likhitha, B. S. Harish, H. M. Keerthi Kumar

Abstract

Text mining is one of the most significant fields in the digital era due to the rapid growth of textual information. Topic models are gaining popularity in the last few years. A topic comprises of a group of words that are often take place together. Topic models are better performing techniques to extract semantic knowledge presented in the data. The various methods used for topic models are, LSA (Latent Semantic Analysis), PLSA (Probabilistic Latent Semantic Analysis), LDA (Latent Dirichlet Allocation). These methods gained popularity in extracting hidden themes from the document (corpus). Various topic modeling algorithms are developed to inquiry, summarize and extract hidden semantic structures of large corpus. In this paper, we present a detailed survey covering the various topic modeling techniques proposed in last decade. Additionally, we focus on different strategies of extracting the topics in social media text, where the goal is to find and aggregate the topic within short texts. Further, we summarize the various applications and quantitative evaluation of the various methods, with statistical and mathematical knowledge to predict the convergence of results.
References

allocation and the Dirichlet multinomial mixture model on short text. In Pattern Recognition
Association of South Africa and Robotics and Mechatronics International Conference
(PRASA-RobMech), 2016 (pp. 1-6). IEEE.
Proceedings of the 17th ACM conference on Information and knowledge management (pp.
1431-1432). ACM.
on streaming document collections. In Proceedings of the 15th ACM SIGKDD international
conference on Knowledge discovery and data mining (pp. 937-946). ACM.
Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. Multimedia Tools
and Applications, pp. 1-43.
incorporating word embeddings. In Pacific-Asia Conference on Knowledge Discovery and Data
Mining (pp. 363-374). Springer.
Proceedings of the first workshop on social media analytics (pp. 80-88). ACM.
In Proceedings of the 2018 SIAM International Conference on Data Mining (pp. 369-377).
on Intuitionistic Fuzzy Entropy to Categorize Text Documents. International Journal of
Interactive Multimedia & Artificial Intelligence, 5(3).
Representation over Short Text. In 2018 IEEE Third International Conference on Data Science
in Cyberspace (DSC) (pp. 848-853). IEEE.
text. Intelligent Data Analysis, 21(S1), pp.S55-S70.
28. Xun, G., Gopalakrishnan, V., Ma, F., Li, Y., Gao, J. and Zhang, A., 2016. Topic
discovery for short texts using word embeddings. In Data Mining (ICDM), 2016 IEEE 16th
International Conference on (pp. 1299-1304). IEEE.
text topic mining between LDA and NMF based Schemes. Knowledge-Based Systems, 163,
30. Harish, B.S. and Revanasiddappa, M.B., 2017. A comprehensive survey on various
feature selection methods to categorize text documents. International Journal of Computer
Applications, 164(8), pp.1-7.
Semantic Similarity Biterms for Short-Text Topic Discovery. In 2018 IEEE 30th International
Conference on Tools with Artificial Intelligence (ICTAI) (pp. 975-982). IEEE.
Discovering Latent Topics via Word Co-Occurrence Network Structure.
document retrieval.
51. Zhao, H., Du, L. and Buntine, W., 2017, November. A word embeddings informed
focused topic model. In Asian Conference on Machine Learning (pp. 423-438).


Index Terms

Computer Science
Information Sciences

Keywords
Text mining, Topic Modeling, Short Text, Semantic Analysis.