Probabilistic Topic Models for Text Data Retrieval and Analysis

ChengXiang Zhai
University of Illinois at Urbana-Champaign
USA
czhai@illinois.edu

ABSTRACT

Text data include all kinds of natural language text such as web pages, news articles, scientific literature, emails, enterprise documents, and social media posts. As text data continues to grow quickly, it is increasingly important to develop intelligent systems to help people manage and make use of vast amounts of text data ("big text datafi). As a new family of effective general approaches to text data retrieval and analysis, probabilistic topic models, notably Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocations (LDA), and many extensions of them, have been studied actively in the past decade with widespread applications. These topic models are powerful tools for extracting and analyzing latent topics contained in text data; they also provide a general and robust latent semantic representation of text data, thus improving many applications in information retrieval and text mining. Since they are general and robust, they can be applied to text data in any natural language and about any topics. This tutorial systematically reviews the major research progress in probabilistic topic models and discuss their applications in text retrieval and text mining. The tutorial provides (1) an in-depth explanation of the basic concepts, underlying principles, and the two basic topic models (i.e., PLSA and LDA) that have widespread applications, (2) a broad overview of all the major representative topic models (that are usually extensions of PLSA or LDA), and (3) a discussion of major challenges and future research directions.

1 MOTIVATION

Text data include all kinds of natural language text such as web pages, news articles, scientific literature, emails, enterprise documents, and social media posts. In contrast to non-textual data which are usually generated by physical devices, text data are generated by humans and meant to be consumed by humans. Due to the rapid growth of text data, we can no longer digest all the relevant information in a timely manner. Thus there is a pressing need for developing intelligent software tools to help people manage and make use of vast amounts of text data ("big text datafi) for various tasks, especially those involving complex decision-making. Logically, to harness big text data, we would need to first identify the relevant text data to a particular application problem (i.e., perform text data retrieval) and then analyze the identified relevant text data in more depth to extract any needed knowledge for a task (i.e. text data analysis). Text data retrieval is essential for filtering out non-relevant text data to a particular problem, thus dramatically improving the efficiency since any further processing can be focused on the much smaller subset of relevant text data and avoid touching all the raw text data. As a subsequent step, text data analysis can further help users digest all the relevant text content by revealing the most useful knowledge or interesting patterns in the text data, thus providing users with more direct support for their tasks.

Due to the difficulty in natural language understanding by computers, the approaches that work well for text retrieval and text analysis tend to be statistical approaches. In the past decade, a special kind of statistical approaches, called probabilistic topic models, represented by Probabilistic Latent Semantic Analysis (PLSA) [2] and Latent Dirichlet Allocations (LDA) [1] and many of their extensions, have been studied actively with widespread applications. These topic models provide a general and robust latent semantic representation of text data, thus improving many applications in information retrieval and text mining. Since they are general and robust, they can be applied to text data in any natural language and about any topics.

Specifically, probabilistic topic models have many applications in information retrieval, including particularly improvement of text representation by representing text documents with low-dimensional latent topic vectors, summarization of search results, modeling user information needs, and topic-based feedback. Probabilistic topic models have also been applied to analyze text data to discover topics and reveal their variations over context variables such as time, location, and sources of text data, generating useful topic patterns to facilitate digestion of text content and knowledge discovery from text data. As a result, recent years have seen increasing uses of topic models in the papers published in ACM SIGIR conferences. However, there has not yet been a tutorial at SIGIR on this topic. This tutorial is meant to fill in this gap.

2 OBJECTIVES

The goal of the tutorial is to systematically review the major research progress in probabilistic topic models and discuss their applications in text retrieval and text mining. The tutorial will provide an in-depth explanation of the basic concepts, underlying principles, and the two basic topic models (i.e., PLSA and LDA) that have widespread applications, a broad overview of all the major representative topic models that are usually extensions of PLSA or LDA), and a discussion of major challenges and future research directions.

The tutorial would be appealing to anyone who would like to learn about topic models, how and why they work, their widespread

SIGIR'17, August 7–11, 2017, Shinjuku, Tokyo, Japan

978-1-4503-5022-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3077136.3080823

applications, and the remaining challenges to be solved in terms of research, including especially graduate students and researchers in both academia and industry who want to do research in this area to develop new topic models. The tutorial would also be appealing to industry practitioners who want to apply topic models to solve many application problems.

3 FORMAT AND CONTENT

The tutorial is a half-day tutorial with the following outline.

3.1 Background

This part will provide any necessary background to the audience to prepare them for understanding the main content which starts from the next part.

3.1.1 Text Data Retrieval and Mining. This part will provide an overview of text retrieval and text mining and position them in a unified framework where text retrieval serves for the purpose of converting the very large raw text collection into a much smaller more relevant set of documents that would be actually needed for a particular application (thus avoiding processing of a lot of non-relevant text data), and text mining intends to further help users digest the found relevant text data and finish their tasks. Of course, the boundary between text retrieval and text mining will not be drawn rigorously, and the separation of the two is mostly to facilitate understanding of somewhat different flavors of applications of topic models (they mostly differ in the importance of "query"). This part will set the context for understanding applications of topic models in text retrieval and text mining.

3.1.2 Statistical Language Models. This part will give a brief introduction to the general topic of statistical language model (topic models are a special kind of language models). Some basic concepts such as likelihood function, statistical estimation, maximum likelihood, Bayes rule and Bayesian estimation would be introduced to facilitate understanding of probabilistic topic models.

3.2 Basic Topic Models

This part will explain the two basic topic models (i.e., PLSA and LDA) in sufficient detail to ensure the audience to understand them well since they are the foundation of most other topic models.

- Probabilistic Latent Semantic Analysis (PLSA)
  This part will start from the simplest mixture model with just one topic and a background language model to gradually introduce more general mixture model with a detailed and thorough explanation of the EM algorithm and why it converges. Maximum a Posteriori (MAP) estimation would also be introduced for extending PLSA with an informative prior.
- Latent Dirichlet Allocation (LDA)
  This part will discuss deficiencies of PLSA and how LDA would address the deficiencies. The likelihood function and generative process of LDA will be introduced. Notation of the plate representation of graphic models will be introduced to facilitate understanding of more advanced models later. Inference algorithms for LDA will be explained with a focus on explaining the collapsed Gibbs sampling algorithm (since it is efficient and applicable to all cases with a conjugate prior).

3.3 Applications of Topic Models in Text Retrieval

This part will review applications of basic topic models (PLSA and LDA) in text retrieval .

3.3.1 Dimension reduction and latent semantic representation. This part will cover how PLSA/LDA can be used to improve text representation. They can be used to perform dimension reduction so as to obtain a low-dimentional latent semantic representation of text documents, which can then be used for many applications that rely on vector space representation of text data, especially in combination with the keyword-based vector space representation.

3.3.2 Topic models for ad hoc retrieval. This part will cover how topic models have been used to improve ad hoc retrieval, mostly invovling integration of topic models with language modeling approaches to ad hoc retrieval.

3.3.3 Topic models for other retrieval tasks. This part will cover other applications of topic models to retrieval tasks, including e.g., subtopic retrieval, summarization, and cross-lingual retrieval etc.

3.4 Applications of Topic Models to Text Mining

This part will review applications of basic topic models (PLSA and LDA) in text mining.

3.4.1 Topic discovery and analysis . This part will cover how the basic output from topic models, including word distributions characterizing topics and topic coverage distributions representing documents , can be used for applications that require discovery of latent topics from text data and analyze their patterns.

3.4.2 Topic labeling and interpretation. This part will cover methods for automatically generating labels for interpreting topics that are represented by word distributions.

3.5 Advanced Topic Models

This part will review a variety of more advanced topic models that are usually extensions of the basic topic models in interesting ways.

3.5.1 Capturing Topic Structures. This part will cover topic models that introduce topic structures (e.g., hierarchical structures or correlated topics). These models can discover not only topics but also their latent structures.

3.5.2 Contextualized Topic Models. This part will cover topic models extended to model both text data and the companion non-text data such as all kinds of meta-data (e.g., time, location, authors, sources) as well as complicated context such as social networks.

3.5.3 Supervised Topic Models. This part will cover topic models that can model text with companion labeled data (e.g., sentiment ratings or topic category labels). Since the companion labels or ratings can provide additional supervision for topic modeling, these

models are often very powerful for discovering very sophisticated latent patterns related to topics or sentiment.

3.6 Summary

This part will summarize the major points of the proposa with the main take-away messages, recommendations for applications of topic models, and a discussion of remaining challenges and future research directions.

4 PRESENTER BIOGRAPHY

ChengXiang Zhai is a Professor of Computer Science and Willett Faculty Scholar at the University of Illinois at Urbana-Champaign (UIUC), where he also holds a joint appointment at Carl R. Woese Institute for Genomic Biology, Statistics, and School of Information Sciences. His research interests include information retrieval, text mining, natural language processing, machine learning, biomedical and health informatics, and intelligent education systems. He has published over 200 papers in these areas with high citations. He served as an Associate Editor of ACM Transactions on Information Systems, and Information Processing and Management, and Program Co-Chair of NAACL HLT 2007, ACM SIGIR 2009, and WWW 2015. He is an ACM Distinguished Scientist, and received a number of awards, including ACM SIGIR Test of Time Award (three times), the 2004 Presidential Early Career Award for Scientists and Engineers (PECASE), an Alfred P. Sloan Research Fellowship, IBM Faculty Award, HP Innovation Research Award, Microsoft Beyond Search Research Award, UIUC Rose Award for Teaching Excellence, and UIUC Campus Award for Excellence in Graduate Student Mentoring. He has two MOOCs on Coursera on Text Retrieval and Text Mining, respectively. He has given many tutorials, including a tutorial on Statistical Language Models for Information Retrieval at HLT-NAACL 2004, SIGIR 2005, SIGIR 2006, HLT-NAACL 2007, and 2011 CCF Advanced Disciplines Lectures, a tutorial on Axiomatic Analysis and Optimization of Information Retrieval Models at ICTIR 2013 and SIGIR 2014, a tutorial on Statistical Language Models for Text Data Mining at 2012 CCF Advanced Disciplines Lectures, and Statistical Methods for Mining Big Text Data at 2014 PhD School, Queensland University, Australia. More information about him and his work can be found at http://czhai.cs.illinois.edu/.

REFERENCES

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993–1022. http://dl.acm.org/citation.cfm?id=944919.944937

[2] Thomas Hofmann. 1999. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*. ACM, New York, NY, USA, 50–57. DOI: http://dx.doi.org/10.1145/312624.312649