Clustering of conference papers using LDA based topic modelling

Mohamed Minhaj
Associate Professor - Systems
SDMIMD, Mysuru
mminhaj@sdmimd.ac.in

Abstract

With the growing number of conferences being organised across the world, researchers have access to large archives of scientic articles. While the amount of conference papers is increasing rapidly, searching papers of specific interest, organising and summarizing them has remained a key challenge. One of the ways in which the task of harnessing and harvesting of scientific data can be improved is by automatic segmentation of the huge collection of conference papers. The automatic segmentation would facilitate better organisation of the conference papers and also improve information retrieval from those collections. Topic Modeling provides a simple approach to analyse large amount of textual-data by identifying the hidden thematic structures in document collections. The Topic models help in annotating documents according to the inherent topics and aids in summarizing and searching. Topic models are currently being used for a variety of purposes like enhancing search facility, recommendation systems, classification of information assets etc. This paper has endeavoured to use LDA, a popular topic modelling algorithm to cluster the conference papers. This work may be useful in implementing

efficient information organisation and retrieval mechanism in the institutional scholarly repositories. It is envisioned that further research on this area would enable automating some of the operational aspects related to organisation of conferences like deciding on the tracks for presentation of papers and also improving the Journal Management Systems.

Keywords : *Analytics, Text Mining, Clustering, Topic Modelling*

Introduction

"You can dig out facts, but if you don't go deep enough, the facts won't tell you much" – Jeff Catlin, CEO Lexalytics, Inc.

With the constant advancements in the web based storage and information dissemination systems, the scholarly data available to the current researchers is overwhelming. The scholarly archives are exponentially growing in size as new articles are placed online and old articles are digitized and indexed. While this growth has allowed researchers to access more scientic information, it has also made it more difcult for them to find articles relevant to their specific interests. Modern researchers need new tools to search and organise these vast amounts of information. Text mining plays a pivotal role in searching, organising and understanding these vast amounts of textual data. Text mining is relatively old, yet evolving domain. Significant advancements have been made in this field, which has facilitated better organisation and understanding of texts, semi or fully automated processes in text summarization,

sentiment analysis etc. The paper endeavours to apply these established approaches in clustering scholarly papers that academic institutions deal with when they conduct conferences.

Text mining approaches and applications

Text mining, also known as text data mining or knowledge discovery from textual databases, refers generally to the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents. It can be viewed as an extension of data mining or knowledge discovery from structured databases (Tan, 1999). Text mining is an area of computer science which fosters strong connections with natural language processing, data mining, machine learning, information retrieval and knowledge management (Radovanoviæ & Ivanoviæ, 2008). While enormous advancement has been made in text mining, the domain is still evolving as it involves dealing with text data that is inherently unstructured and fuzzy. Text mining basically involves two aspects, transforming the unstructured text sources into a form which would facilitate statistical analysis of the data and studying the patterns or harvesting useful insights. As majority of the data available today is in the form of text which is predominantly unstructured, text mining is gaining prominence.

Bag- of-Words(BoW) document representation

Conventionally, most text mining processes have treated the text sources as a bag-of-words. In the bag-of-words

model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity (Sivic & Zisserman, 2009). After transforming the text into a "bag of words", we can calculate various measures to characterize the text (Wikipedia : Bag-of-words, 2016). The most common type of characteristics, or features calculated from the Bag-of-words model is term frequency, namely, the number of times a term appears in the text. The bag-of-words model is commonly used in methods of document classification where the frequency of occurrence of each word is used as a feature for training a classifier. If C is a corpus of text documents (D1, D2, D3…. Dm) and if S is set of all the words or terms (W1, W2, W3…. Wn) that are present in C, a matrix with M rows and N columns is constructed, which is referred as Term-by-Documents matrix. In the document-term matrix or term-document matrix, there are various schemes for determining the value that each entry in the matrix should take. One such scheme is Term Frequency.

	W1	**W2**	**W3**	**.**	**.**	**Wn**
D1	F(W1,D1)	F(W2,D1)	F(W3,D1)	.	.	F(Wn,D1)
D2	F(W1,D2)	F(W2,D2)	F(W3,D2)	.	.	F(Wn,D2)
D3	F(W1,D3)	F(W2,D3)	F(W3,D3)	.	.	F(Wn,D3)
.	.	.	.	.	.	.
.	.	.	.	.	.	.
Dm	F(W1,Dm)	F(W2,Dm)	F(W3,Dm)	.	.	F(Wn,Dm)

The limitation of BoW approach is that the order of the words is lost and as a consequence, the semantics or the meaning of the text is also lost. For example, “This is good” and “Is this good” have exactly the same vector representation (Quora, 2016). While BoW cannot be employed for deep text analytics due its limitations, it continues to be widely used intermediate process for many shallow text mining tasks.

Topic Models: From Bag-of-Words to Bag-of-Topics

Topic modelling is basically a text mining technique to extract themes from large unstructured text sources. In one of the prominent early works related to Topic Models, David M. Blei et.al, devised a generative probabilistic model for collections of discrete data such as text corpora called Latent Dirichlet Allocation (LDA). The goal of their wok was to find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks such as classification, novelty detection, summarization, and similarity and relevance judgments (Blei, Ng, & Jordan, 2003). Since then significant progress has been made related to this domain by both the scholarly community and the practitioners.

Some useful definitions of Topic Model :

A topic model is a type of statistical model for discovering the abstract “topics” that occur in a collection of documents. (Wikipedia, 2016).

Topic models are a suite of algorithms that uncover the hidden thematic structure in document collections. These algorithms help us develop new ways to search, browse and summarize large archives of texts. (Princeton, 2016)

Topic models provide a simple way to analyze large volumes of unlabeled text. A “topic” consists of a cluster of words that frequently occur together. Using contextual clues, topic models can connect words with similar meanings and distinguish between uses of words with multiple meanings. (Umass, 2016)

Basically Topic modelling entails a process of discovering hidden semantic structures in a text body. Given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently. For example, “Product” and “Promotion” will appear more often in documents about “Marketing”. “Selection” and “Jobs” will appear in documents about “Recruitment” and “the” and “is” will appear equally in both. A document typically concerns multiple topics in different proportions. Thus in a document that is 10% about “Recruitment” and 90% about “Marketing”, there would probably be about 9 times more “Marketing” words than “Recruitment” words. The “topics” produced by topic modelling techniques are clusters of similar words. A topic model captures this intuition in a mathematical framework, which allows examining a set of documents and discovering, based on the statistics of the words in each, what the topics might be and what each document’s balance of topics is. (Wikipedia, 2016).

Topic modelling algorithms

Topic models are based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words. A variety of probabilistic topic models have been used to analyse the content of documents and the meaning of words. All these models use the same fundamental idea – that a document is a mixture of topics – but make slightly different statistical assumptions (Steyvers & Griffiths, 2007)

Name of the Topic Model Technique	Characteristics	Limitations
Latent Semantic Analysis (LSA)	Technique is used in natural language processing and involves analysing relationships between a set of documents and the terms they contain by producing a set of concepts. LSA assumes that words that are close in meaning will occur in similar pieces of text. A matrix containing word counts per paragraph (rows represent unique words and columns represent each paragraph) is constructed from a large piece of text and a mathematical technique called Singular Value Decomposition (SVD) is used to reduce the number of	It is hard to obtain and to determine the number of topics. Also, interpreting the results in difficult.

	rows while preserving the similarity structure among columns. Words are then compared by taking the cosine of the angle between the two vectors (or the dot product between the normalizations of the two vectors) formed by any two rows. Values close to 1 represent very similar words while values close to 0 represent very dissimilar words.	
Probabilistic Latent Semantic Analysis (PLSA)	This technique which is also known as Probabilistic Latent Semantic Indexing (PLSI) is a statistical technique for the analysis of two-mode and co-occurrence data. In effect, one can derive a low-dimensional representation of the observed variables in terms of their affinity to certain hidden variables, just as in latent semantic analysis, from which PLSA evolved. Compared to standard latent semantic analysis which stems from linear algebra and downsizes the occurrence tables (usually via a Singular Value Decomposition), probabilistic latent semantic analysis is based on a mixture decomposition derived from a latent class model.	At the level of documents, PLSA cannot do probabilistic model.

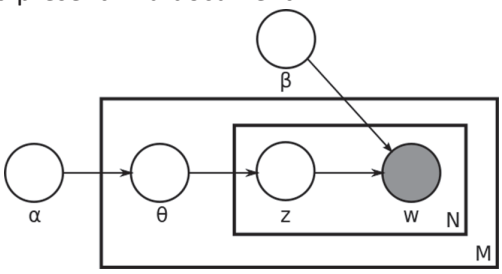
The Correlated Topic Model (CTM)	It is a hierarchical model of document collections. The CTM models the words of each document from a mixture model. The mixture components are shared by all documents in the collection. The CTM allows each document to exhibit multiple topics with different proportions. It can thus capture the heterogeneity in grouped data that exhibit multiple latent patterns.	Requires lot of calculations. Also, it includes a lot of general words inside the topics.

Prominent Topic Model Algorithms

LDA – A popular Topic Model Algorithm

Latent Dirichlet Allocation (LDA) is very popular topic model, which was devised by David Blei, Andrew Ng, and Michael I. Jordan in 2003 (Blei D. M., 2003). It is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics (Wikipedia.org, 2016).

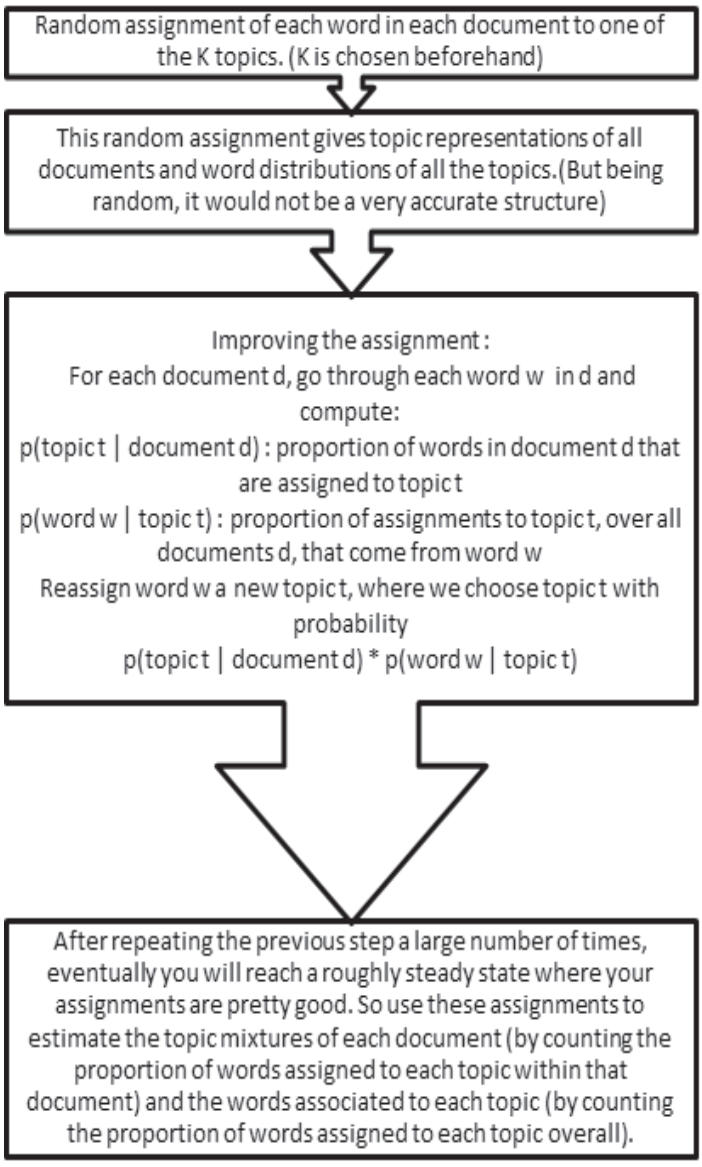
In the LDA model, each document is viewed as a mixture of topics that are present in the corpus. The model proposes that each word in the document is attributable to one of the document's topics. The LDA model discovers the different topics that the documents represent and how much of each topic is present in a document.



α is the parameter of the Dirichlet prior on the per-document topic distributions,
β is the parameter of the Dirichlet prior on the per-topic word distribution,
θ_i is the topic distribution for document *i*,
φ_k is the word distribution for topic *k*,
z_{ij} is the topic for the *j*th word in document *i*, and
w_{ij} is the specific word.

Plate notation for Latent Dirichlet Allocation (LDA)

The process involved in LDA



Existing work related to use of LDA for segmenting text sources:

For text classication, topic modelling techniques have been utilized in various ways. In the work of Zhang et al., it is used as a keyword selection mechanism by selecting the top words from topics based on their entropy. In the study conducted by Efsun Sarioglu et.al, they have removed the most frequent and infrequent words to have a manageable vocabulary size for the purpose of segregating the documents. Sarioglu et al., and Sriurai compare BoW representation to topic model representation for classication using varying and fixed number of topics respectively. In the study conducted by Banerjee, topics are used as additional features to BoW features for the purpose of classication. Further, Chen et al., developed a resampling approach based on topic modelling when the class distributions are not balanced.

Clustering of Conference Papers using LDA

With several conferences being conducted regularly across the globe, humongous research papers are produced by the scholarly community. While significant advancement has been made in the way these research papers are archived both in the institutional repositories and the public archives, fast and easy retrieval of relevant research papers from these repositories remains a key challenge. One of the ways in which harnessing and harvesting of scientific data can be improved is by automatic segmentation of the huge collection of conference papers.

The automatic segmentation would facilitate in better organisation of the conference papers and also improve information retrieval from those collections.

Topic modeling is an unsupervised technique that can automatically identify themes from a given set of documents and find topic distributions of each document. Representing documents according to their topic distributions is more compact and can be processed faster than raw text in subsequent automated processing. Topic modelling provides us with methods to organize, understand and summarize large collections of textual information. It helps in:

- Discovering hidden topical patterns that are present across the collection
- Annotating documents according to these topics
- Using these annotations to organize, search and summarize texts

Prominent Software Tools for Topic Modeling :

MALLET	It is "MAchine Learning for LanguagE Toolkit". MALLET is an integrated collection of Java code useful for statistical natural language processing, document classification, cluster analysis, information extraction, topic modeling and other machine learning applications to text.
Stanford Topic Modeling Toolkit	It is developed and supported by the Stanford Natural Language Processing Group. It manipulates text from cells in Excel and other spreadsheets. It supports several algorithms of Topic Models.
Gensim	It is an open-source vector space modeling and topic modeling toolkit, implemented in the Python programming language. It uses NumPy, SciPy and optionally Cython for performance. It is specifically intended for handling large text collections, using efficient online, incremental algorithms.

Steps involved in Topic Modeling

1. Pre-processing the conference papers : The first step as part of pre-processing, is converting all the papers into a uniform format like MS-Word, PDF, Text Files. The next and an important step is creating Term by Document Matrix from the files. To reduce the dimensionality of the matrix, stop-words (generic

words like is, a, the etc.) are removed. Optionally terms can also be filtered based on their length to discount words with very few characters and several characters.

2. Creation of Topic using any Topic Model algorithm. In case of LDA, the number of topics (K) has to be specified beforehand. The output of this step would be K topics and terms associated with each topic. For example, in case of papers pertaining to Finance conference, typical output would be as follows:

Topic 1	**Topic 2**	**……..**	**Topic K**
Index Forwards Futures Options Swaps	Amalgamation Consolidation Synergy Diversification Negotiation Valuation		

Additionally, this step would also result in all the input documents associated with probabilities of belonging to different Topics generated by the algorithm.

	Topic 1	**Topic 2**	**Topic 3**	**…….**	**Topic K**
Document 1	P (Topic 1)	P (Topic 2)	P (Topic 3)	……..	
Document 2					
Document 3					
. .					
Document M					

3. These topics can then be manually labelled, for example Topic 1 as “Derivatives”, Topic 2 as “Merger and Acquisition” etc., and based on the highest probability of documents having specific topics, the documents can be annotated with the Topic label appropriately.

Document 1	Derivatives
Document 2	Merger and Acquisition
Document 3	Derivatives
……	

Conclusion and future work

A large amount of unstructured data in the form of documents, blogs, tweets etc., is continuously being generated in the world. As a result, today the problem is not about unavailability of data, instead it is about abundant data, which is making the task of accessing specific information difficult. This issue has triggered a lot of research pertaining to tools and techniques to organise, search and understand vast quantities of information. One of the important development in this domain is Topic Modeling which provides a simple approach to analyse large amount of data by identifying the hidden thematic structures in document collections. Topic models are being used for a variety of purposes like enhancing search facility, recommendation systems, classification of information assets etc. This paper has endeavoured to use LDA, a

popular topic modeling algorithm to cluster the conference papers. This work may be useful in implementing efficient information organisation and retrieval mechanism in the institutional repositories. It is envisioned that this work can further be extended to automate some of the operational aspects related to conferences like deciding on the tracks for presentation of papers and also improving the Online Journal Management Systems.

References

(2016, 11 02). Retrieved from Quora: https://www.quora.com/What-are-the-limitations-of-the-Bag-of-Words-model

(2016, 10 27). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Topic_model

(2016, 10 27). Retrieved from Princeton: https://www.cs.princeton.edu/~blei/topicmodeling.html

(2016, 10 25). Retrieved from Umass: http://mallet.cs.umass.edu/topics.php

(2016, 10 29). Retrieved from Wikipedia.org: https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

Alghamdi, R. &. (2015). A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications (IJACSA)*.

Blei, D. M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 993-1022.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 993-1022.

Radovanoviæ, M., & Ivanoviæ, M. (2008). Text mining: Approaches and applications. *Novi Sad J. Math*, 227-234.

Sivic, J., & Zisserman, A. (2009). Efficient visual search of videos cast as text retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 591-606.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In *Handbook of latent semantic analysis* (pp. 424-440).

Tan, A. H. (1999). Text mining: The state of the art and the challenges. *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases.*

Wikipedia : Bag-of-words. (2016, 11 01). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Bag-of-words_model