SAJET

South Asian Journal of Engineering and Technology Vol.2, No.15 (2016) 184–191

ISSN No: 2454-9614

Information Retrieval and Recommendation Framework using Document Models

A. Saranya[a], V. Vennila[a*], G. Sivaselvan [a,b]

[a)] *Department Of Computer Science and Engineering, K.S.R. College of Engineering, Tiruchencode, Namakkal, Tamilnadu, India.*
[b)] *Department Of Computer Science and Engineering, K.S.R. College of Engineering, Tiruchencode, Namakkal, Tamilnadu, India.*
[c)] *Department Of Computer Science and Engineering K.S.R. College of Engineering, Tiruchencode, Namakkal, Tamilnadu, India.*

*Corresponding Author: A. Saranya

E-mail: sanjanasara333@gmail.com,

Received: 22/11/2015, Revised: 16/12/2015 and Accepted: 12/03/2016

Abstract

Text mining is the process of deriving high-quality information from text. Text categorization, text clustering, concept extraction and sentiment analysis operations are carried out in text mining. Information filtering methods are applied to remove redundant and unwanted information from the documents. Latent Dirichlet Allocation (LDA) is applied to generate statistical models to represent multiple topics in a collection of documents. Maximum matched Pattern-based Topic Model (MPBTM) is used to perform information filtering process. Statistical and taxonomic features are used to organize the topic model based patterns. Document relevance is estimated using Maximum Matched Patterns such as discriminative and representative patterns. Pattern based document model is constructed for information filtering and information retrieval tasks. Document recommendation is estimated using the user requirement information. Pattern based index scheme is adapted to perform information retrieval with ranking feature. Relevant document identification is improved with content based features.

Keywords: Data mining, Text mining, Topic modelling, Document relevance, Information filtering, latent Dirichlet allocation, supervised learning.
Reviewed by* **ICETSET'16 *organizing committee*

1. Introduction

The purpose of Text Mining is to process unstructured information, extract meaningful numeric indices from the text and make the information contained in the text accessible to the various data mining algorithms. Information can be extracted to derive summaries for the words contained in the documents or to compute summaries for the documents based on the words contained in them.

The user can analyze words, clusters of words used in documents, etc., or the user could analyze documents and determine similarities between them or how they are related to other variables of interest in the data mining

project. In the most general terms, text mining will "turn text into numbers", which can then be incorporated in other analyses such as mining projects, the application of unsupervised learning methods, etc. These methods are described and discussed in great detail in the comprehensive overview work by Manning and Schutze and for an in-depth treatment of these and related topics as well as the history of this approach to text mining.

To reiterate, text mining can be summarized as a process of "numeric zing" text. At the simplest level, all words found in the input documents will be indexed and counted in order to compute a table of documents and words, i.e., a matrix of frequencies that enumerates the number of times that each word occurs in each document. This basic process can be further refined to exclude certain common words such as "the" and "a" and to combine different grammatical forms of the same words such as "travelling," "travelled," "travel," etc. Once a table of words by documents has been derived, all standard statistical and data mining techniques can be applied to derive dimensions or clusters of words or documents, or to identify "important" words or terms that best predict another outcome variable of interest.

Once a data matrix has been computed from the input documents and words found in those documents, various well-known analytic techniques can be used for further processing those data including methods for clustering, factoring, or predictive data mining **[1].** Examples of scenarios using large numbers of small or moderate sized documents were given earlier. On the other hand, if their intent is to extract "concepts" from only a few documents that are very large, then statistical analyses are generally less powerful because the "number of cases" in this case is very small while the "number of variables" is very large.

Excluding numbers, certain characters, or sequences of characters, or words that are shorter or longer than a certain number of letters can be done before the indexing of the input documents starts. The user may also want to exclude "rare words," defined as those that only occur in a small percentage of the processed documents.

Specific list of words to be indexed can be defined; this is useful when the user want to search explicitly for particular words and classify the input documents based on the frequencies with which those words occur. Also, "stop-words," i.e., terms that are to be excluded from the indexing can be defined. Typically, a default list of English stop words includes "the", "a", "of", "since," etc, i.e., words that are used in the respective language very frequently, but communicate very little unique information about the contents of the document.

Synonyms, such as "sick" or "ill", or words that are used in particular phrases where they denote unique meaning can be combined for indexing. For example, "Microsoft Windows" might be such a phrase, which is a specific reference to the computer operating system, but has nothing to do with the common use of the term "Windows" as it might, for example, be used in descriptions of home improvement projects.

2. Related Work

“Common” and “specific” words, sparsity in topic proportions and word probabilities, as well as estimation of the number of topics have all been the subject of previous studies. Wallach et al. **[2]** introduced asymmetric

Dirichlet priors over topic distributions and word probabilities to control skewness in word and topic frequency distributions. Asymmetric priors were shown to prevent common words from dominating all topics and also help achieve sparser topic presence in documents. Similar to LDA, this approach is not parsimonious. All topics have nonzero proportion in every document and all words are modeled in a topic-specific fashion. Wang and Blei introduced a spike and slab model to control sparsity in word probabilities. Unlike our approach, does not use a shared distribution. Moreover, it does not provide the subset of relevant topics for each document. A similar approach, based on the Indian Buffet Process, was used in **[7]** to address sparsity only in topic proportions in a non-parametric topic model.

Global background models have been used in information retrieval. In these models, the probability of each word under every topic is a mixture of the background model and topic-specific word probabilities. A similar idea has been used in **[8],** with words well-modeled by the background model having small topic-specific probabilities. The mixing proportions in these models are hyper parameters that should be estimated by cross-validation. Chemudugunta and Steyvers **[9]** proposed a combination of background, general and document-specific topics to improve information retrieval. The authors argued that LDA "over generalizes" and is thus not effective for matching queries that contain both high-level semantics and keywords. Chemudugunta and Steyvers **[9]** introduced a huge set of new free parameters by adding a document- specific topic for every document. In these models, similar to LDA, each word, under every topic, has a free parameter. By contrast, in our model, the shared model is heavily used, with each topic possessing relatively few topic specific words. Also unlike these approaches, our model is sparse in topic proportions.

Zhu and Xing **[10]** presented sparse topical coding (STC), a non probabilistic topic model which gives parsimony in topic-proportions but models all words topic-specifically. Moreover, this method has three hyper parameters that must be determined by cross-validation. Non-parametric topic models have been proposed that relax the requirement of specifying the number of topics. Similar to LDA, these methods do not exhibit parsimony in their modelling.

Our approach can also be viewed from the standpoint of unsupervised feature selection. For each topic we select salient features in an unsupervised fashion, modelling the rest using a universal shared model. Unsupervised feature selection and shared feature representations have been considered in some prior works. Law et al. used a minimum message length criterion to find salient features in a mixture model. Features were tied across all components; i.e. each feature is either salient or shared in all components. A related Bayesian framework was presented in **[3].** The concept of shared feature space for mixtures was further improved proposing a component-specific feature space; i.e. a feature can be salient in some components but represented by the shared model in others. Graham and Miller used the minimum description length (MDL) **[4]** and standard mixture of unigrams for modelling documents. Boutemedjet et al. **[5]** performed unsupervised feature selection by minimizing the message length of the data, considering mixtures of generalized Dirichlet distributions.

This model was then optimized in a Bayesian framework via variational inference in **[6].** A simple example to illustrate sparsity on word probabilities and topic proportions for a corpus with two documents, six words and three topics. Contributions of this paper. Compared to previous works, our main contributions are:

1) We extend the concept of shared feature space from standard mixtures to more general topic models, allowing presence of multiple topics in documents. In doing so, we achieve sparsity in topic proportions and in topic-specific words. Prior works at best achieve sparsity in one of these two senses.

2) Unlike most works, our model allows the subset of salient words to be topic-specific. This follows the premise that some words may have common frequency of occurrence under some subset of the topics. For example, the word "component" has different meanings under "statistics" and "machine learning" than under other topics and could have higher frequencies of occurrence for these specialized topics.

3) We derive a novel BIC objective function, used for learning our model. Unlike the native form, satisfyingly, our derived objective has distinct penalty terms for the different parameter types in our model, interpretable vis a vis the effective sample size for each of the parameter types. These methods are described and discussed in great detail in the comprehensive overview work by Manning and Schutze and for an in-depth treatment of these and related topics as well as the history of this approach to text mining.

3. Document Modelling in Information Filtering

Information filtering (IF) is a system to remove redundant or unwanted information from an information or document stream based on document representations which represent users' interest. Traditional IF models were developed using a term-based approach. The advantage of the term-based approach is its efficient computational performance, as well as mature theories for term weighting, such as Rocchio, BM25, etc. But term-based document representation suffers from the problems of polysemy and synonymy. To overcome the limitations of term-based approaches, pattern mining based techniques have been used to utilize patterns to represent users' interest and have achieved some improvements in effectiveness, since patterns carry more semantic meaning than terms. Also, some data mining techniques have been developed to improve the quality of patterns for removing the redundant and noisy patterns.

All these data mining and text mining techniques hold the assumption that the user's interest is only related to a single topic. In reality this is not necessarily the case. For example, one news article talking about a "car" is possibly related to price, policy, market and so on. At any time, new topics may be introduced in the document stream, which means the user's interest can be diverse and changeable. Therefore, we propose to model users' interest in multiple topics rather than a single topic, which reflects the dynamic nature of user information needs.

Topic modelling has become one of the most popular probabilistic text modelling techniques and has been quickly accepted by machine learning and text mining communities. It can automatically classify documents in a collection by a number of topics and represents every document with multiple topics and their corresponding

distribution. Two representative approaches are Probabilistic Latent Semantic Analysis (PLSA) and LDA. There are two problems in directly applying topic models for information filtering. The first problem is that the topic distribution itself is insufficient to represent documents due to its limited number of dimensions. The second problem is that the word based topic representation is limited to distinctively represent documents which have different semantic content since many words in the topic representation are frequent general words.

In order to alleviate the ambiguity of the topic representations in LDA, we proposed a promising way to meaningfully represent topics by patterns rather than single words through combining topic models with pattern mining techniques. Specifically, the patterns are generated from the words in the word based topic representations of a traditional topic model such as the LDA model. This ensures that the patterns can well represent the topics because these patterns are comprised of the words which are extracted by LDA based on sample occurrence and co occurrence of the words in the documents. The pattern based topic model, which has been utilized in IF, can be considered as a "post-LDA" model in the sense that the patterns are generated from the topic representations of the LDA model. Because patterns can represent more specific meanings than single words, the pattern-based topic models can be used to represent the semantic content of the user's documents more accurately compared with the word-based topic models. Very often the number of patterns in some of the topics can be huge and many of the patterns are not discriminative enough to represent specific topics. In this paper, we propose to select the most representative and discriminative patterns, which are called Maximum matched Patterns, to represent topics instead of using frequent patterns. A new topic model, called MPBTM is proposed for document representation and document relevance ranking. The patterns in the MPBTM are well structured so that the maximum matched patterns can be efficiently and effectively selected and used to represent and rank documents.

The original contributions of the proposed MPBTM to the field of IF can be described as follows:

1) We propose to model users' interest with multiple topics rather than a single topic under the assumption that users' information interests can be diverse.

2) We propose to integrate data mining techniques with statistical topic modelling techniques to generate a pattern-based topic model to represent documents and document collections. The proposed model MPBTM consists of topic distributions describing topic preferences of each document or the document collection and pattern-based topic representations representing the semantic meaning of each topic.

3) We propose a structured pattern-based topic representation in which patterns are organized into groups, called equivalence classes, based on their taxonomic and statistical features. Patterns in each equivalence class have the same frequency and represent similar semantic meaning. With this structured representation, the most representative patterns can be identified which will benefit the filtering of relevant documents

4) We propose a new ranking method to determine the relevance of new documents based on the proposed model and especially, the structured pattern based topic representations. The Maximum matched patterns, which are the largest patterns in each equivalence class that exist in the incoming documents, are used to calculate the

relevance of the incoming documents to the user's interest. The maximum matched patterns are the most representative and discriminative patterns to determine the relevance of incoming documents.

4. Issues on Document Modeling Schemes

Term or pattern-based approaches are used for information filtering to generate users' information needs from a collection of documents. Document collection and user interest are categorized under multiple topics. Latent Dirichlet Allocation (LDA) is applied to generate statistical models to represent multiple topics in a collection of documents. Topic models are widely utilized in the fields of machine learning and information retrieval. Selection of the most discriminative and representative patterns from the huge amount of discovered patterns is a complex task. Maximum matched Pattern-based Topic Model (MPBTM) is used to perform information filtering process. Statistical and taxonomic features are used to organize the topic model based patterns. Document relevance is estimated using Maximum Matched Patterns such as discriminative and representative patterns. The following issues are identified from the current document modelling schemes.

- Information retrieval is not adapted by the system
  - User requirement based recommendation is not provided
  - Content based feature extraction is not supported
  - Pattern based indexing is not supported

5. Information Retrieval and Recommendation Framework

Pattern based document model is constructed for information filtering and information retrieval tasks. Document recommendation is estimated using the user requirement information. Pattern based index scheme is adapted to perform information retrieval with ranking feature. Relevant document identification is improved with content based features.

Information filtering tasks are carried out using pattern based topic models. Document search process is performed with document models and topic details. Recommendation schemes are constructed with the document and user information needs. The system is divided into six major modules. They are Document Analysis, Topic Model Construction, Statistical and Taxonomic Features, Pattern based Indexing, Information retrieval and Recommendation Process.

Document pre-process operations are carried out under the document analysis module. Topic models are constructed to organize the user information needs. Topic patterns are building with the statistical and taxonomic features. Patterns are arranged with pattern based index scheme. Information retrieval module is designed to perform document search operations. Relevant document selection is achieved using recommendation process.

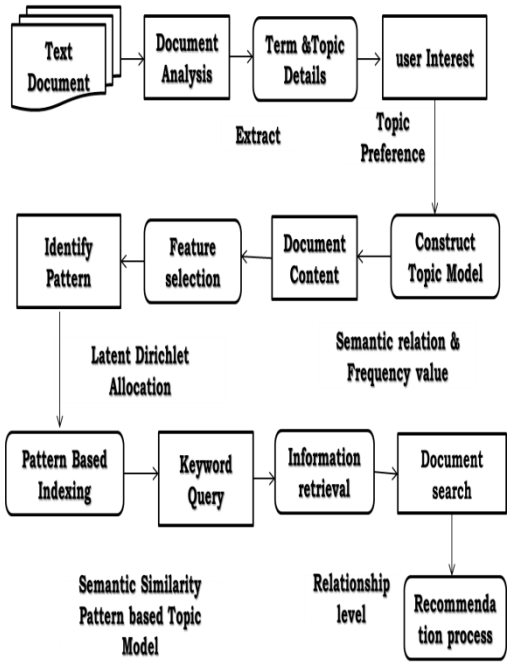

Fig. 1 Information Retrieval and Recommendation Process

5.1 Document Analysis

Document and topic information are used in the document analysis process. Document elements are extracted and updated into the database. Term collections are updated with count details. Terms and topic details are used in the pattern extraction process.

5.2 Topic Model Construction

Topic model represents the document and document collections. Topic models are generated with user information requirements. Each topic is assigned with a set of relevant term collections. Topic distributions with topic preferences for the documents are considered in the topic model construction process.

5.3 Statistical and Taxonomic Features

Document contents are processed to fetch the statistical and taxonomic features. Statistical and taxonomic features are used for structured pattern based topic representation. Equivalence class is build to organize patterns into groups. Features are updated with semantic relations and frequency values.

5.4 Pattern based Indexing

Latent Dirichlet Allocation (LDA) results are used to construct transactional set. Pattern based representation are generated with the transactional data sets. Semantic relationship based pattern index scheme is applied to arrange pattern values. Semantic Similarity Pattern based Topic Model (SSPTM) is adapted in the system.

5.5 Information Retrieval

The information retrieval module is designed to perform document search process. Keyword query is collected from the user. Query value is matched with the topic model to fetch the relevant documents. Documents are produced with ranked manner using pattern index values.

5.6 Recommendation Process

Recommendation process is designed to identify the similar documents based on the user needs. User specified document is compared with all other documents. Topic model for the document is compared with the topic models of the document corpus. Recommended documents are produced with relationship levels.

6. Conclusion

Document models are constructed with pattern based topics to perform information filtering operations. Maximum matched Pattern-based Topic Model (MPBTM) is used to organize topics with patterns. The system is enhanced to support information retrieval on document models. Recommendation features are integrated with the system to fetch relevance based document collections. The system produces Pattern representation in ranked manner. User interest based document retrieval process is adapted in the search process. Semantic relationship based recommendation mechanism is employed to fetch relevant document collections. Optimal document representation model is adapted to perform the mining operations.

References

[1] HosseinSoleimani and David J. Miller, "Parsimonious Topic Models with Salient Word Discovery", IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 3, March 2015
[2] H. M. Wallach and A. McCallum, "Rethinking LDA: Why priors matter," in Proc. Adv. Neural Inf. Process. Syst., 2009.
[3] C. Constantinopoulos, M. K. Titsias and A. Likas, "Bayesian feature and model selection for Gaussian mixture models."IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 6, pp. 1013–1018, Jun. 2006.
[4] P. D. Gr€unwald, The Minimum Description Length Principle. Cambridge, MA, USA: MIT Press, 2007.
[5] S. Boutemedjet, N. Bouguila and D. Ziou, "A hybrid feature extraction selection approach for high-dimensional non-Gaussian data clustering," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 8, pp. 1429–1443, Aug. 2009.
[6] W. Fan, N. Bouguila and D. Ziou, "Unsupervised hybrid feature extraction selection for high-dimensional non-Gaussian data clustering with variational inference," IEEE Trans. Knowl. Data Eng., vol. 25, no. 7, pp. 1670–1685, Jul. 2013.
[7] S. Williamson K. Heller and D. M. Blei, "The IBP compound Dirichlet process and its application to focused topic modeling," in Proc. 27th Int. Conf. Mach. Learn., 2010.
[8] S. Lacoste-Julien, F. Sha and M. Jordan, "Disclda: Discriminative learning for dimensionality reduction and classification," in Proc. Adv. Neural Inf. Process. Syst., 2009, pp. 897–904.
[9] D. Mimno and A. McCallum, "Topic models conditioned on arbitrary features with Dirichlet-multinomial regression," in Proc. 24th Annu. Conf. UncertainityArtif.Intell., 2008, pp. 411–418.
[10] J. Zhu and E. P. Xing, "Sparse topical coding," in Proc. 27th Conf. Uncertainty Artif.Intell., 2011, pp. 831–838.