International Journal On Engineering Technology and Sciences – IJETS™
ISSN(P): 2349-3968, ISSN (O): 2349-3976
Volume III, Issue X, October- 2016

Document Hierarchical Model Construction and Indexing Approach for Textual Data Mining

Ms. J. Sharmila Ananthi, MPhil., Research Scholar,
Mrs. R. Sasiregha, MSc., MPhil, Assistant Professor,
PG & Research, Department of Computer Science,
SSM College of Arts and Science, Komarapalayam, Tamilnadu, India

Abstract

The textual data mining methods are applied to extract quality information from the text data values. Text mining tasks are carried out to perform classification, clustering, concept identification and sentiment discovery operations. Conflict and irrelevant information are removed in the information filtering process. Information filtering is achieved with term and pattern based techniques. User's information needs from document corpus are discovered using the term and pattern based methods. Topic categorization is carried out on the document corpus and user interest details.

Multiple topic based statistical model is constructed with Latent Dirichlet Allocation (LDA) technique. Machine learning and information search operations are carried out using the topic models. Discriminative and representative patterns are selected from the discovered patterns. Information filtering is achieved with Maximum matched Pattern-Based Topic Model (MPBTM) mechanism. The topic model based patterns are organized with Statistical and taxonomic features. Discriminative and representative features are used to estimate document relevance.

Document hierarchical model is build to perform information filtering and search operations. User requirement based relevant documents are identified for the recommendation process. The information search and relevant document identification operations are performed in ranked manner with pattern based index scheme. Content relationship levels are analyzed to discover the relevant documents. The concept relationships are used in the document search and recommendation process.

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. 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.

2. Related Work

Several previous studies have included author or user information when modeling documents using topic models. For example, in the author-topic model [1], each author is associated with a mixture over topics and each word in a document is associated with a mixture of the authors' topics mixtures. Built upon the AT model, the author-recipienttopic (ART) model [2] conditions per message topic distribution from emails jointly on both the author and individual recipients. The role-author-recipient-topic (RART) model [3] extends from ART that it assumes an author can have multiple roles and a role is a persona represented by a topic distribution.

While our work is closely related to these studies, it is different from them in the following aspects. First, the aforementioned models assume either a single user specific interest [1] or a shared set of personas or roles among all the users [2]. On the contrary, the role-based topical interests in our models are both role- and user-specific, which are opposed to the role-based topic distribution shared among all the users as in RART. Second, "roles" defined in previous models are intrinsically different from the social roles defined in our models which essentially correspond to different social activities. Third, ART explicitly characterizes the two roles in a specific relation and is not easy to generalize to multi-role modeling. RART cannot model user-level interactions but only role-level interactions. Last, the models built upon sender-receiver relations assume that there are at least one "sender" and one "receiver" for any document, which

is clearly not the case in our datasets. For example, there are many tweets which are never retweeted on Twitter.

In addition to various author/user topic models mentioned above, there has been some work incorporating underlying network structures into topic modeling using regularization methods [3]. Our work is partly inspired by NetPLSA [3], but our focus is to model role-specific interests of users and we propose a principled approach which incorporates SRT into the generative process of topic models. Recently, there have also been some studies analyzing users' retweeting behaviors when performing topical analysis in Twitter [4]. However, they only focus on finding topical authorities and do not model both users and topics.

3. Information Filtering with Document Models

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. The system models 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, the system uses 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. The system selects the most representative and discriminative patterns, which are called Maximum matched Patterns, to represent topics instead of using frequent patterns. A topic model, called MPBTM is applied 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 MPBTM to the field of IF can be described as follows:

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

2) Data mining techniques are integrated with statistical topic modeling techniques to generate a pattern-based topic model to represent documents and document collections. The 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) A structured pattern-based topic representation is constructed with 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) A ranking method is designed to determine the relevance of new documents based on the 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. Problem Statement

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 drawbacks are identified from the existing system.

- 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. Semantic Supported Pattern based Topic Model (SSPTM)

Text documents are unstructured data values with raw collection of terms. Document modeling approaches are adapted to normalize the text document contents. The preprocessing methods are applied to identify the terms in the document collections. The documents are maintained with labels. The term collections are arranged with labels or topic information. User interest and topic overlapping factors are also considered in the document modeling process.

Document models are used for the mining and analysis operations. 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.

The text documents are build with raw contents. The information filtering methods are applied to refine the text document contents. Term and pattern based document models are constructed for the information filtering process. All the text mining operations are carried out on the document models. The Pattern Based Topic Model (PTM) is build with terms and taxonomical information. The terms are extracted from the documents with its frequency values. The terms and associated topic values are analyzed with Latent Dirchlet Analysis (ODA) Method. The similar topics and associated terms are refined into topic models. The index process is initiated to rank the topic patterns. All the document retrieval and recommendation operations are carried out using the constructed document models.

The Pattern based Topic Model (PTM) scheme constructs the document model term information only. Term relationship details are not considered in the PTM based document model construction process. The term and concept relationships are maintained under the Ontology. The Ontology is a repository that maintains the concepts and associated terms with its relationship details. Three types of relationships are maintained in the Ontology. They are Synonym, Meronym and Hypernym relationships. The synonym relationship indicates the terms with similar meaning of the concepts. The meronym relationship shows the logical part of the concept relationship details. The hypernym indicates the typical relationship of the concepts with the terms. Semantic weight or concept weight is estimated with reference to the relationship levels. The Semantic Supported Pattern based Topic Model (SSPTM) is

constructed with term and concept relationships. The topics are extracted fom the associated concept details maintained under the Ontology repository. The document models are building with SSPTM mechanism.

The document search and recommendation operations are performed using the SSPTM based document models. The document search operation is carried out with the user input query value. The system fetches the documents that are related to the given query text value. The documents are ranked with the semantic relationship weight values. The document recommendation process is carried out with the given input document. The input document is constructed as document model. The document model is compared with all document models constructed for the entire corpus. The similar documents are produced as recommended documents for the input document.

6. Performance Analysis

The document search and relevance discovery operations are performed using the Semantic Similarity Pattern based Topic Model (SSPTM) construction scheme. The Pattern based Topic Model (PTM) and the (SSPTM) schemes are adapted in the system analysis. The system uses Ontology for data mining domain. Document search is tested with different query levels.

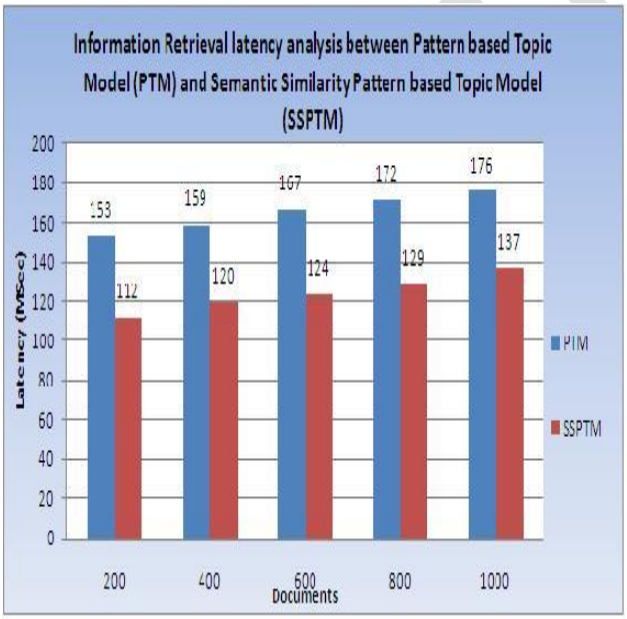

Figure No. 6.1. Information Retrieval latency analysis between Pattern based Topic Model (PTM) and Semantic Similarity Pattern based Topic Model (SSPTM)

The document retrieval and recommendation system is tested with three performance measures. They are information retrieval latency, information retrieval accuracy and topic relevancy analysis. The system is tested with 1000 text documents collected from the IEEE web site. The retrieval latency analysis is carried out to compare the document retrieval period. Figure 6.1. shows the retrieval rate analysis between the Pattern based Topic Model (PTM) and Semantic Similarity Pattern based Topic Model (SSPTM) techniques. The analysis result shows that the Semantic Similarity Pattern based Topic Model (SSPTM) minimizes the retrieval latency 10% than the Pattern based Topic Model (PTM) technique. The information retrieval accuracy level analysis is carried out to compare the document retrieval accuracy levels. Figure 6.2 shows the information retrieval accuracy between the Pattern based Topic Model (PTM) and Semantic Similarity Pattern based Topic Model (SSPTM) techniques. The analysis result shows that the Semantic Similarity Pattern based Topic Model (SSPTM) increase the information retrieval accuracy 15% than the Pattern based Topic Model (PTM) technique.

The topic relevancy analysis is performed to measure the topics relationship level under different pattern analysis mechanism. Figure 6.3. and shows the topic relevancy analysis between the Pattern based Topic Model (PTM) and Semantic Similarity Pattern based Topic Model (SSPTM) techniques. The analysis result shows that the Semantic Similarity Pattern based Topic Model (SSPTM) increase the topic relevancy 10% than the Pattern based Topic Model (PTM) technique.

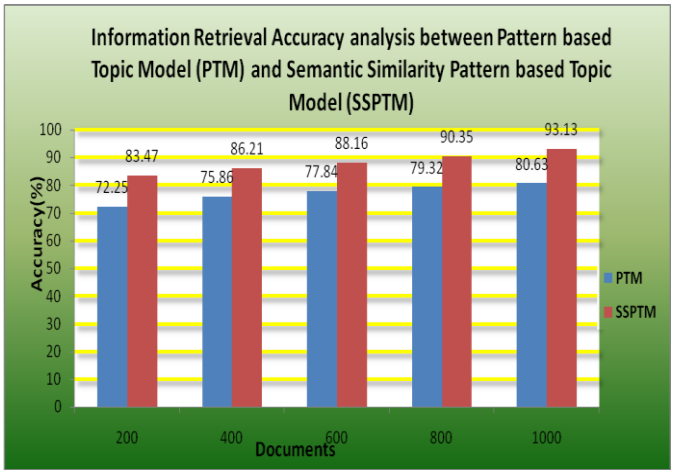

Figure No.6.2. Information Retrieval Accuracy analysis between Pattern based Topic Model (PTM) and Semantic Similarity Pattern based Topic Model (SSPTM)

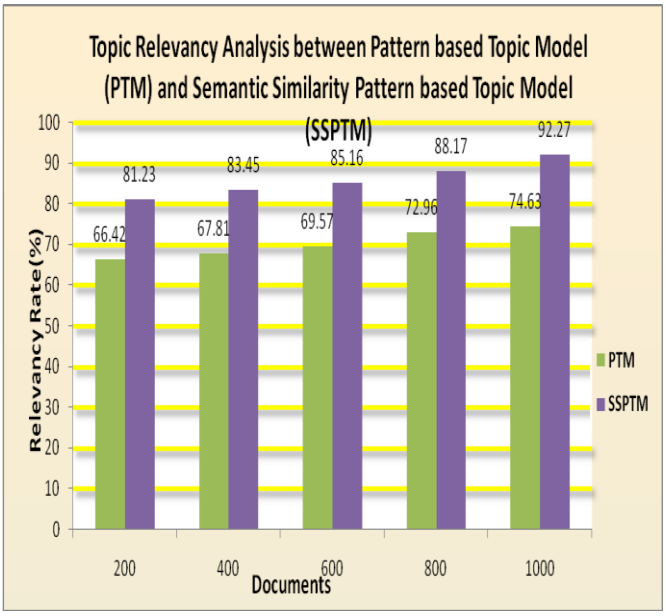

Figure No. 6.3. Topic Relevancy Analysis between Pattern based Topic Model (PTM) and Semantic Similarity Pattern based Topic Model (SSPTM)

7. Conclusion And Future Enhancement

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. The system can be enhanced with the following futures.

- The information retrieval and recommendation system can be enhanced to perform document model based search operations under web document based environment
- Clustering techniques can be integrated with the document model construction scheme to improve the document model construction process

References

1. M. Steyvers, P. Smyth, M. Rosen-Zvi and T. L. Griffiths, "Probabilistic author-topic models for information discovery," in Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2004, pp. 306–315.
2. Q. Mei, D. Cai, D. Zhang and C. Zhai, "Topic modeling with network regularization," in Proc. 17th Int. Conf. World Wide Web, 2008.
3. A. McCallum, A. Corrada-Emmanuel and X. Wang, "Topic and role discovery in social networks," in Proc. 19th Int. Joint Conf. Artif. Intell., 2005, pp. 786–791.
4. D. M. Romero, W. Galuba, S. Asur and B. A. Huberman, "Influence and passivity in social media," in Proc. 20th Int. Conf. Companion World Wide Web, 2011, pp. 113–114.
5. Bo Tang, Haibo He, Paul M. Baggenstoss, and Steven Kay, "A Bayesian Classification Approach Using Class-Specific Features for Text Categorization", IEEE Transactions on Knowledge and Data Engineering, June2016
6. Haoji Hu, Kai Zheng, Xiaoling Wang and Aoying Zhou,"GFilter: A General Gram Filter for String Similarity Search", IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 4, April 2015.
7. H. Hu, G. Li, Z. Bao and J. Feng "Top-k Spatio-Textual Similarity Join", IEEE Transactions on Knowledge and Data Engineering, Volume:28, Issue: 2 , January 2016.
8. Hossein Soleimani and David J. Miller, "Parsimonious Topic Models with Salient Word Discovery", IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 3, March 2015.
9. T. Lou and J. Tang, "Mining structural hole spanners through information diffusion in social networks," in Proc. 22nd Int. Conf. World Wide Web, 2013, pp. 825–836.
10. K. Henderson, B. Gallagher, T. Eliassi-Rad, H. Tong, S. Basu, L. Akoglu, D. Koutra, C. Faloutsos and L. Li, "RoIX: Structural role extraction and mining in large graphs," in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2012, pp. 1231– 1239.