International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 02 Issue: 09 | Dec-2015 www.irjet.net p-ISSN: 2395-0072

A SURVEY PAPER ON MODELING METHODS FOR INFORMATION FILTERING AND RELEVANCE RANKING OF DOCUMENTS

Ms Manjiri M. More [1], Prof. Mrs Archana S. Vaidya[2]

[1] *PG- Student, Department of Computer Engineering, Gokhale Education Society's R. H. Sapat College of Engineering, Nashik, India*

[2] PG-Coordinator, *Department of Computer Engineering, Gokhale Education Society's R. H. Sapat College of Engineering, Nashik, India*

Abstract: ***Topic modeling is useful in the area of machine learning and text mining, etc. It was proposed to create statistical models to classify many topics in a collection of documents. A fundamental supposition for these approaches is that the documents in the collection are all about one topic. Patterns are always more discriminative than single terms for describing documents. Selection of the most representative patterns from the huge amount of discovered patterns becomes crucial. To deal with the above limitations and problems, a novel information filtering model is needed. Where user information needs are created in terms of multiple topics where each topic is represented by patterns. In this paper, we present three technical categories of model includes topic modeling methods, pattern mining methods and term-based methods.***

Key Words: ***Topic Modeling, Pattern Mining, Information Filtering.***

1. INTRODUCTION

Many data mining techniques have been used for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still a research issue. Traditional Information filtering models were developed using a term-based approach [1], [2].The advantage of the term-based approach is its efficient computational performance. Term-based document representation tolerate from the problems of polysemy and synonymy. To overcome the limitations of term-based approaches, pattern mining based method have been used to utilize patterns to represent users' interest and have achieved some improvements in effectiveness [3], [4] since patterns carry more semantic meaning than terms [5], [6], [7], [8]. All these data mining and text mining techniques hold the assumption that the user's interest is only related to a single topic. However, in real world this is not necessarily the case. At any time, new topics may be introduced in the document, which means the user's interest can be diverse and changeable. Therefore, here proposes a technique to model users' interest in multiple topics rather than a single topic, which reflects the dynamic nature of user information needs.

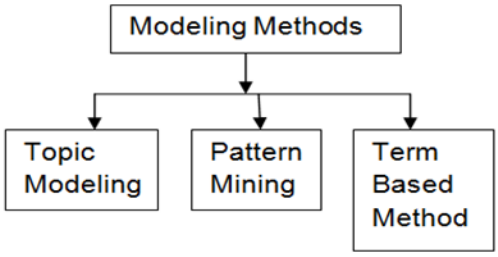

Fig -1: Three technical categories of modeling.

Topic Modelling [9], [10], [11] such as Latent Dirichlet Allocation (LDA) [11] is a probabilistic model for collections of discrete data such as text collections. 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 methods are Probabilistic Latent Semantic Analysis (PLSA) [12] and LDA [11]. However, there are two problems if we directly apply 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 repeated general words.

Here proposes a technique to choose the most representative and discriminative patterns are called Maximum matched Patterns, to represent topics instead of using frequent patterns. A new topic model, called Maximum Matched Pattern-based Topic Model (MPBTM) is proposed for document representation and document relevance ranking. The patterns in MPBTM are well

structured so that the maximum matched patterns can be efficiently and effectively selected and used to represent and rank documents

The organization of this paper is as follows. Section 2 reviews the related work. In Section 3, we present the Maximum Matched Pattern-based Topic Model (MPBTM). In Section 4, we explained algorithm. Section 5 presents conclusions and outline directions for the future work.

2. RELATED WORK

Three technical categories of baseline model include topic modeling methods, pattern mining methods and term-based methods. For each category, some methods were chosen as the baseline models. For the topic modeling category, three topic modeling methods are chosen as baseline models, PLSA word and LDA word, Pattern-based Topic Model (PBTM). For the pattern mining category, the baseline models include Frequent Closed Patterns (FCP), frequent Sequential Closed Patterns (SCP) and phrases (n-Gram). The third category includes the classical term-based method Support Vector Machine (SVM). An important distinguish between the topic modeling methods and other methods is that, the topic modeling methods consider multiple topics in each document collection and use patterns (e.g. PBTM and MPBTM) or words (e.g. LDA word) to represent the topics, whereas the pattern mining and term based methods assume that the documents within one collection are about one topic and use patterns or terms/words to represent documents directly. Literature Survey of these baseline models are given below.

2.1 Topic modeling based category.

The work proposed by X. Wei and W. B. Croft [9] shows LDA-based document models for ad-hoc retrieval. They studied how to efficiently use LDA to improve ad-hoc retrieval. They proposed LDA-based document model method within the language modeling framework.

Collaborative topic modeling for recommending scientific articles is proposed in [10]. They developed an algorithm to recommend scientific articles to users of an online community. This method combines the excellence of traditional collaborative filtering and probabilistic topic modeling. It provides an interpretable latent structure for users and items, and form recommendations about both existing and newly published articles.

Author Qiaozhu Mei, Xuehua Shen, Chengxiang Zhai [13] has studied specially Automatic Labeling of Multinomial Topic Models. They proposed a labeling method that is quite effective to generate labels that are meaningful and useful for interpreting the discovered topic models. This method is common and can be applied to labeling topics learned through every type of topic models such as PLSA, LDA, and their variations. This method is effective and robust when applied on various genres of text collections to label topics generated using various statistical topic models (e.g., PLSA and LDA).

A.Tagarelli and G. Karypis, in [14] has presented a segment-based approach to clustering multi-topic documents. They addressed the problem of multi-topic document clustering by leveraging the natural composition of documents in text segments, which bear one or multiple topics on their own.

In [15] author has studied enriching text representation with frequent pattern mining which is a useful probabilistic topic modeling. They proposed a general way to go beyond the bag-of-words presentation for topic modeling by using frequent pattern mining to find out frequent word patterns that can show semantic associations between words and using them as additional supplementary semantic units to augment the conventional bag-of-words presentation.

In [16] probabilistic topic models are proposed. Generative models for text, such as topic model, have the potential to make important contributions to the statistical analysis of multiple document collections, and the development of a deeper understanding of human language learning and processing.

The work proposed by L. Azzopardi, M. Girolami, and C. Van Rijsbergen [17] shows topic based language models for ad hoc information retrieval. They explored the possibility of using a document specific term prior based on inferred topics induced from the corpus.

2.2 Pattern-based category

In [18] effective pattern discovery for text mining is proposed. Author studied an innovative and effective pattern discovery method which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating generated patterns for finding relevant and interesting information. In this research work, an effective pattern discovery method has been proposed to minimize the low-frequency and misinterpretation problems for text mining. The proposed method uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents.

Author R. J. Bayardo Jr [5] has studied efficiently mining long patterns from databases. They presented a pattern-mining algorithm that scales approximate linearly in the number of maximal patterns embedded in a database regardless of the length of the longest pattern. They presented and evaluated the Max-Miner algorithm for mining maximal frequent item sets from large

databases. Max- Miner applied several new techniques for reducing the space of item sets considered through superset-frequency based pruning. Max-Miner is also easily made to incorporate additional constraints on the set of frequent item sets identified.

2.3 Term-based category

Frequent term based text clustering is proposed in [02]. This approach utilizes frequent term for text clustering. Such frequent sets can be efficiently find out using algorithms for association rule mining. To cluster based on frequent items; they measure the mutual overlap between frequent sets with respect to the sets of supporting documents. They presented two algorithms for Frequent Term-based text Clustering (FTC) which creates flat clustering's and Hierarchical Frequent Term-based text Clustering (HFTC) for hierarchical clustering. This algorithm obtains clustering of comparable quality significantly more efficiently than text clustering algorithms.

In [19] an author has studied a multilevel approach to intelligent information filtering. A filtering model is proposed that divides the overall task into subsystem functionalities and highlights the requirement for multiple adaptation techniques to cope with uncertainties. Proposed filtering system implemented based on the model, using traditional methods in information retrieval and artificial intelligence. These methods include document representation by a vector-space model, document categorization by unsupervised learning, and user modeling by reinforcement learning. Systems can filter information based on content and a user's specific interests. The user's interests are spontaneously learned with only finite user intervention in the form of optional relevance feedback for documents.

Table -1: Summary of Models

Model	Advantages	Disadvantages	Representative approaches
1)Topic model	1) Model can automatically categorize documents in a collection by a number of topics.	1) The topic distribution & representation is insufficient due to its limited number of dimensions. 2) The topic representation is limited to distinctively present documents which have different semantic contents.	PLSA, LDA,PBTM
2)Term based Model	1) Efficient computational performance.	1) Model suffers from the problems of polysemy and synonymy	SVM
3)Pattern based model	1) Models are used to represent the semantic content of the user's documents more accurately.	1) Number of patterns in some of the topics can be huge 2) Many times the patterns are not discriminative enough to represent specific topics.	FCP, SCP ,n-Gram

3. MAXIMUM MATCHED PATTERN-BASED TOPIC MODEL (MPBTM)

In proposed system user's interest with multiple topics are considered. The proposed model Maximum Matched Pattern-based Topic Model [20] consists of topic distributions describing topic preferences of each document or the document collection and pattern-based topic representations presenting the semantic meaning of each topic.

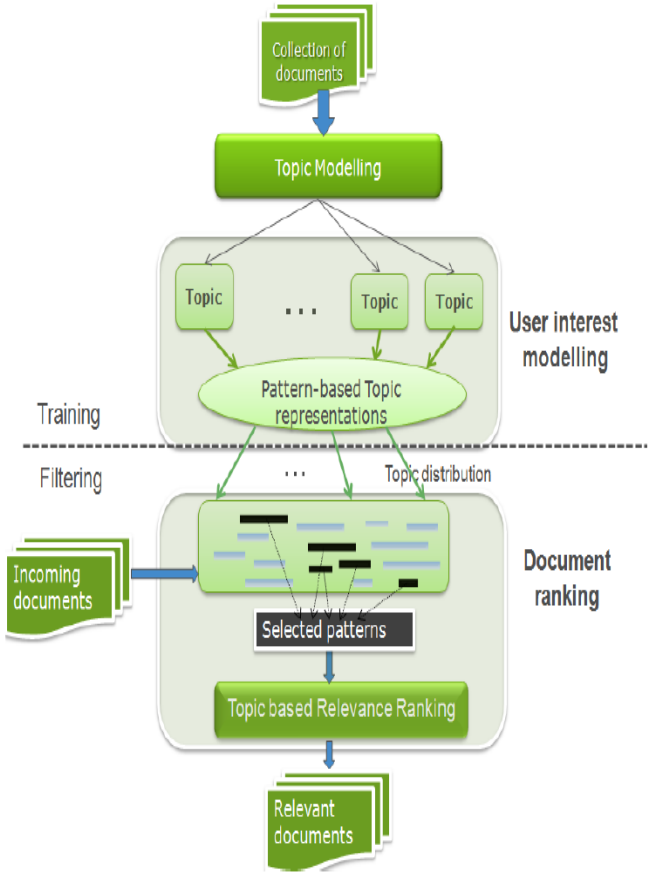

Fig -1: Maximum matched Pattern-based Topic Model (MPBTM)

Here proposed that a structured pattern-based topic representation in which patterns are organized into groups, called equivalence classes, based on their taxonomic and statistical features. With this structured representation, the most representative patterns can be identified which will benefit the filtering of relevant documents. In this system a new ranking method determines the relevance of recent documents based on the proposed model and especially the structured pattern-based topic representations. The maximum matched patterns are the largest patterns in each equivalence class that exist in the incoming documents, and used to evaluate the relevance of the incoming documents to the user's interest.

4. ALGORITHM

The proposed Information filtering model is explained in two algorithms: user profiling (generating user interest models) algorithm and document filtering (relevance ranking of incoming documents) algorithm. First algorithm generates pattern-based topic representations to represent the user's information needs. Algorithm uses functions such as construction of transactional dataset, construction of user interest model and construction of equivalence class. The latter algorithm ranks the incoming documents based on the relevance of the documents to the user's needs. Document filtering algorithm scan the documents to find maximum matched pattern and update the ranking of documents.

5. CONCLUSIONS

Modeling methods are very useful in information filtering, document ranking, content-based feature extraction and modelling tasks, such as information retrieval and recommendations. The main objective of this paper is to review various modeling methods and to study different modeling methods. Here paper presents an innovative pattern enhanced topic model for information filtering including user interest modelling and document relevance ranking. The proposed MPBTM model generates pattern enhanced topic representations to model user's interests across multiple topics. In the filtering stage, the MPBTM selects maximum matched patterns, instead of using all discovered patterns, for estimating the relevance of incoming documents.

REFERENCES

[1] S. Robertson, H. Zaragoza, and M. Taylor, "Simple BM25 extension to multiple weighted fields," in Proc. 13th ACM Int. Conf. Inform. Knowl. Mana, 2004, pp. 42–49.

[2] F. Beil, M. Ester, and X. Xu, "Frequent term-based text clustering,"in Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2002,pp. 436–442.

[3] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal, "Mining frequent patterns with counting inference," ACM SIGKDD Explorations Newslett., vol. 2, no. 2, pp. 66–75, 2000.

[4] H. Cheng, X. Yan, J. Han, and C.-W. Hsu, "Discriminative frequent pattern analysis for effective classification," in Proc. IEEE 23rd Int. Conf. Data Eng., 2007, pp. 716–725.

[5] R. J. Bayardo Jr, "Efficiently mining long patterns from databases," in Proc. ACM Sigmod Record, 1998, vol. 27, no. 2, pp. 85–93.

[6] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: Current status and future directions," Data Min. Knowl. Discov. vol. 15, no. 1, pp. 55–86, 2007.

[7] M. J. Zaki and C.-J. Hsiao, "CHARM: An efficient algorithm for closed itemset mining." in Proc. SDM, vol. 2, 2002, pp. 457–473.

[8] Y. Xu, Y. Li, and G. Shaw, "Reliable representations for association rules," Data Knowl. Eng., vol. 70, no. 6, pp. 555–575, 2011.

[9] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2006, pp. 178–185.

[10] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in Proc. 17th ACM SIGKDD Int. Conf.Knowl. Discov. Data Min., 2011, pp. 448–456.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.

[12] T. Hofmann, "Probabilistic latent semantic indexing," in Proc.22nd Annu. Int. ACM SIGIR Conf. on Res. Develop. Inform. Retrieval, 1999, pp. 50–57.

[13] Q. Mei, X. Shen, and C. Zhai, "Automatic labeling of multinomial topic models," in Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discov.Data Min., 2007, pp. 490–499.

[14] A. Tagarelli and G. Karypis, "A segment-based approach to clustering multi-topic documents," Knowl. Inform. Syst., vol. 34, no. 3,pp. 563–595, 2013.

[15] H. D. Kim, D. H. Park, Y. Lu, and C. Zhai, "Enriching text representation with frequent pattern mining for probabilistic topic modeling," in Proc. Am. Soc. Inform. Sci. Technol., 2012, vol. 49, no. 1, pp. 1–10.

[16] M. Steyvers and T. Griffiths, "Probabilistic topic models," Handboo Latent Semantic Anal., 2007, vol. 427, no. 7, pp. 424–440.

[17] L. Azzopardi, M. Girolami, and C. Van Rijsbergen, "Topic based language models for ad hoc information retrieval," in Proc. Neural Netw. IEEE Int. Joint Conf., 2004, vol. 4, pp. 3281–3286.

[18] N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for text mining," IEEE Trans. Knowl. Data Eng., vol. 24, no. 1, pp. 30–44, Jan. 2012.

[19] J. Mostafa, S. Mukhopadhyay, M. Palakal, and W. Lam, "A multilevel approach to intelligent information filtering: Model, system, and evaluation," ACM Trans. Inform. Syst., vol. 15, no. 4, pp. 368– 399, 1997.

[20] Yang Gao, Yue Xu, and Yuefeng Li, "Pattern-based Topics for Document Modelling in Information Filtering,"in Knowledge and Data Engineering,2015.