Understanding Topic Influence Based on Module Network

Jinlong Wang[1], Congfu Xu[2], Dou Shen[3], Guojing Luo[2], and Xueyu Geng[4]

[1] School of Computer Engineering, Qingdao Technological University
Qingdao, 266033, China
WangJinlong@gmail.com
[2] Institute of Artificial Intelligence, Zhejiang University
Hangzhou, 310027, China
xucongfu@cs.zju.edu.cn
[3] Microsoft adCenter Labs, Redmond WA 98052
doushen@microsoft.com
[4] Institute of Geotechnical Engineering Research, Zhejiang University
Hangzhou, 310027, China

Abstract. Topic detection and analysis is very important to understand academic document collections. By further modeling the influence among the topics, we can understand the evolution of research topics better. This problem has attracted much attention recently. Different from the existing works, this paper proposes a solution which discovers hidden topics as well as the relative change of their intensity as a first step and then uses them to construct a module network. Through this way, we can produce a generalization module among different topics. In order to eliminate the instability of topic intensity for analyzing topic changes, we adopt the piece-wise linear representation so that we can model the topic influence accurately. Some experiments on real data sets validate the effectiveness of our proposed method.

1 Introduction

Topic analysis over academic document collections is beneficial to researchers since it can help researchers find out hot topics at a certain stage and the topic evolution patterns. It can even discover how topic changes affect the researchers' actions and vise versa.

Some recent works [1, 2] have focused on discovering the relationships among the topics. In [1], the authors work on the problem of discovering evolutionary transitions. When two themes have a smaller evolution distance, measured with the KL-Divergence, the themes are claimed to be closer to each other. However, this measurement only reflects the similarity between two themes, and cannot make multiple topics simultaneously. In [2], the authors propose a method for discovering dependency relationships among the topics in a collection of documents shared in social networks. This paper's hypothesis is that one topic evolves into another topic through the interaction between the corresponding social actors with different topics in the latent social network. Based on the hypothesis,

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 391–399, 2007.

the authors can compute the transition probability among topics and rank the authors to see who dominate the topic transition. However, this relation is only a transition between topics, it computes the quantitative realtionship of different topics through the co-authors. The Markov transition graph among topics is obtained by computing the transition between topics respectively, which is not a global graph. In [3], we use the dynamical bayesian network to model the research field development with a global graph, but it needs input to be sequential data, and it only reflects the relations among fields, not the topics.

Different from previous methods, in this paper, we attempt to generalize the topic's interactional relations reflected by topic intensity, which reflects the topic development and trend. We use the relative change of topic intensity over time to build a module network for topic interrelation generalization. As a generative model, the module network [4] can partition the variables into modules and learn the dependency structure of each module. The variables in each module share the same parents in the network and the same conditional probability distribution. By feeding a set of variables and the maximum module number, this method can generate a module network automatically. [4] applies this method to cluster stock sectors based on the stock price change, which generalizes better than the traditional bayesian network. Especially, the learned module network provides some important insights for stock sector based on the stock price. For example, the stock of high-tech service companies are in the same module, and have the same parent of manufacturer. Essentially, the topic intensity with time is similar to the stock data. Thus, we can use the module network to learn topic relation for a better understanding of the latent influence among the topics.

The rest of the paper is organized as follows. Section 2 introduces the related works. Section 3 presents our method. Section 4 describes the experimental setup and the results. The last section concludes the paper.

2 Related Works

2.1 Document Content Analysis

A variety of statistical approaches have been proposed to model a document, such as LSA (Latent Semantic Analysis) [5], pLSI (Probabilistic Latent Semantic Indexing) [6], LDA (Latent dirichlet allocation) [7] *etc.* In recent years, LDA models, as effective approach for generating topics, receives more and more attentions. The model uses the Dirichlet distribution to model the distribution of topics for each document. Each word is considered sampled from a multinomial distribution over words specific to topics. LDA models are well-defined generative models and generalize easily to new documents without overfitting. Some recent work has been concerned with temporal documents [8, 9].

2.2 Module Network

As a generative model, a module network [4] partitions the variables into modules, and studies the dependent relation among the modules. A module network

can be viewed simply as a Bayesian network [10] in which variables in the same module share parents and parameters. This provides a good representation for understanding the data. When learning module network, given a domain of random variables $X = \{X_1, \cdots, X_n\}$, we can obtain K modules $M_1, \cdots, M_K$. A module network consists of two components. The first defines a template probabilistic model for each module; all of the variables assigned to the module will share this probabilistic model. The second component is a module assignment function that assigns each variable to one of the K modules.

Fig. 1 represents an example of module network, the rectangle denotes the module, and the variables in one module share the same conditional probability table, parameter and same parent node, but may have different descendants. For example, the three variable $\{X_2, X_3, X_4\}$ in M_2 have the same parent X_1, but only the variable X_3 has descendants, which are X_6, X_7 in module M_4. Learning module network includes an iterative approach consisting of two steps: module assignment search step (assign variables to modules) and structure search step (learn the network structure).

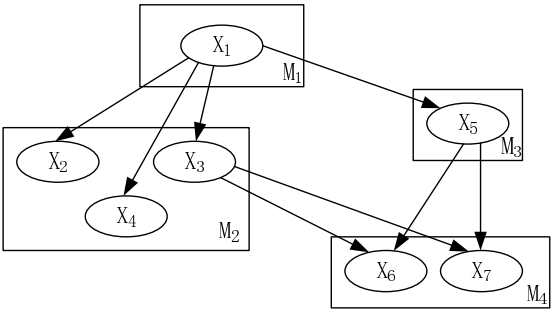


Fig. 1. The module network example

3 Methods

In our method, we take a document collection with time stamps as the input and proceed with three steps to build the module network: (1) discover the topics/themes in the collection as well as the change of their intensity; (2) process the themes strength with piece-wise linear analysis; (3) construct module networks. The module network supplies a graph reflecting topic relations. Based on the graph, we can analyze influences among topics.

3.1 Obtaining Topic Intensity Time Series

As the first step, we extract the topics from the document collection using a LDA model [7], and then obtain the topic dynamics referring to the series of topic with various strength of probability over time. The process of the method is shown in Fig. 2.

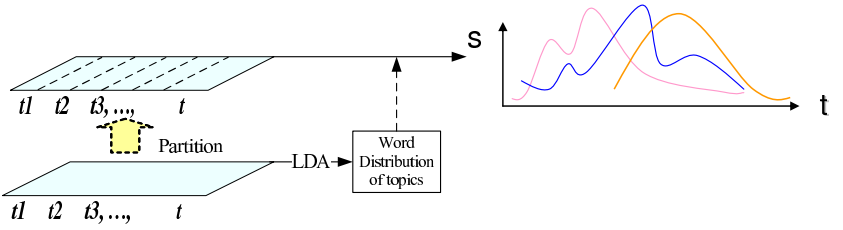


Fig. 2. Obtaining topic intensity time series

For a certain year, the strength of a topic is calculated as the normalized sum of all the probabilities of this topic inferred from all documents in that year. We partition the documents by year. For each year, all of the words are assigned to their most likely topic. The fraction of words assigned to each topic for a given year is then calculated for each of the topics and each year. These provide relative topic popularity in the document collection, and present the trend of topics over time.

3.2 Piece-Wise Linear Representation

With the topic intensity induced in Section 3.1, we can directly use the relative changes between successive years as the input of module network building. However, the induced intensity changes are not stable and may contain much noise due to many reasons such as the fluctuation of the number of documents each year and the statistical error in LDA model. These instability and noise will debase the module network's generalization performance. Actually, comparing with the concrete intensity of a topic at a certain time, the general trend is more important in our analysis. Therefore, we use the piece-wise linear segmentation [11], an effective method for trend representation of time series to eliminate the fluctuation in our problem. The piece-wise linear segmentation attempts to model the data as sequences of straight lines, which has been proved to be effective for data compression and noise filtering. In our problem, for each topic, we set its intensity every-year with the responding slope which is mapped close to the range of [-1, 1]. Fig. 3 is the topic intensity (red line) of SVM topic discovered in the above way and its piece-wise linear representation (blue line). As the input of the module network, it also ensures that the topic's change consistent in a small range of time. In this way, the module network can generate an ordered time sections which better accord with reality.

3.3 Constructing the Module Network

In step 3, we construct a module network. We treat each topic as a variable, and each instance to correspond to a year, where the value of the variable is the result processed with piece-wise linear representation, and the range is in [-1, 1]. Taking these data and the maximum module number as input, we can obtain a high scoring module network based on the scoring function in [4]. Same

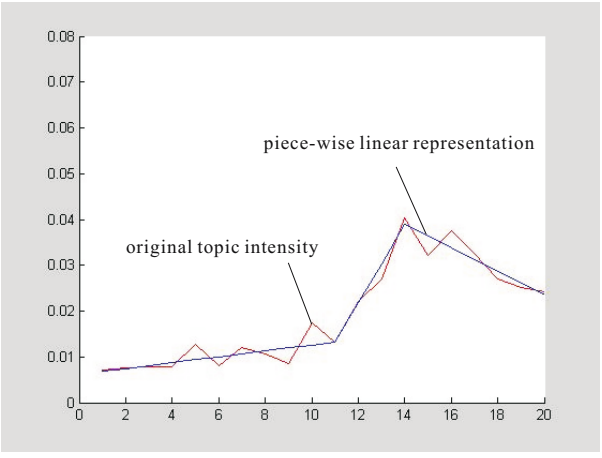


Fig. 3. The topic intensity of SVM and its piece-wise linear representation

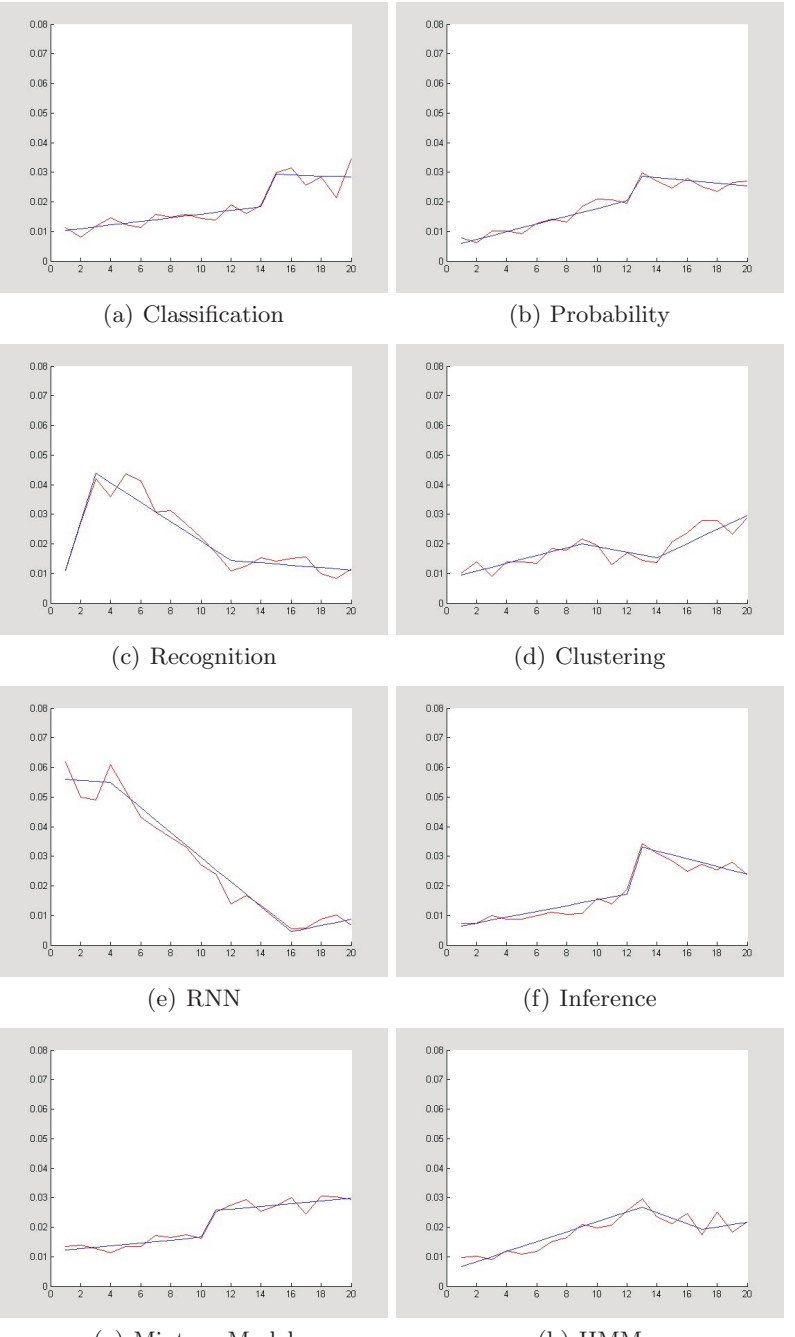
as the Bayesian network, given a scoring function over network, we can find a high scoring module network. We encourage the readers to refer to [4] for more information about module network.

4 Experiments

We use the NIPS (Neural Information Processing Systems) conference data to explain our approach. The NIPS dataset consists of the full text of the 20 years of proceedings from 1987 to 2006. In our experiments, we use the abstracts extracting from the all NIPS papers data download from the website[1]. In order to improve the accuracy of abstract extraction from pdf/ps/djvu format to text, we use different methods toward different time literature data. For the papers in 1987-1999, we use Roweis' raw data, which were corrected errors by hand on yann's djvutotxt data. For papers of 2000-2001, we use the DjVuLibre 3.5.17-1 to transform the file format of djvu to plain text, and the result is much better than using the other two formats. And for the remaining 977 papers without djvu files, we use the text save function of Acrobat7.0 for text export, which is less OCR errors then using PDF2TXT tools. For better discovering the topics, we filter some phases, such as "the", "a", *etc.* by a stopwordlist. At the same time, we deleted two letters' words except some domain words, such as "SVM", "HMM", "KNN", "RNN", "AI", "KL", *etc.* We also deleted words that appeared less than five times in the whole collection, for these words are mostly generated by OCR errors. Finally we extract 3102 abstracts and total 4517 distinct words from the collection.

With the LDA topic model, we extract 50 topics and then obtain each topic's intensity time series as shown in Section 3.1. We select 9 significant topics

[1] http://nips.cc/



(a) Classification (b) Probability

(c) Recognition (d) Clustering

(e) RNN (f) Inference

(g) Mixture Model (h) HMM

Fig. 4. The other topics' (SVM result as Fig. 3) intensity and their piece-wise representation

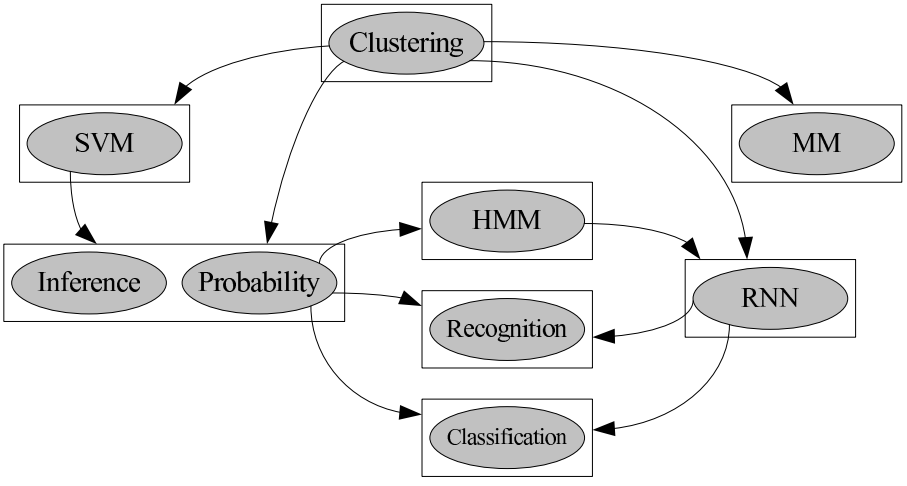


Fig. 5. The module network with 9 topics

Table 1. Conditional probabilistic table

(a) CPT of SVM node

	Clustering(Up)	Clustering(Down)
SVM(Up)	0.533	1
SVM(Down)	0.467	0

(b) CPT of Prob node

	SVM(Up) Clustering(Down)	SVM(Down) Clustering(Up)
Pro(Up)	0.8	0
Pro(Down)	0.2	1

Mixture Model (MM), Clustering, Recurrent Neural Network (RNN), SVM, HMM, Classification, Inference, Probability, Recognition, that are frequently referred to machine learning in the NIPS.

With the piece-wise linear segmentation, the topic intensity can be presented better with trend on a macroscopical view. Since we have only 20 years' data, the more the segmentation number is, the harder to see the trend. In our experiments, we vary the segment number from 2 to 3, 4 and 5. We find that the result is best for understanding with the 3 segmentation. The result is as Fig. 4.

Using the 9 topics' intensity data to construct the module network, and the maximum module number is set as 9, we obtain the network with 6 modules as shown in Fig. 5. The topic Inference and Probability are clustered together. The figure shows the topic relation/influence according to the arrowhead. For example, the topic Clustering influences the topic SVM, MM, RNN and Prob (Inference and Probability). The topic RNN influences Recognition and

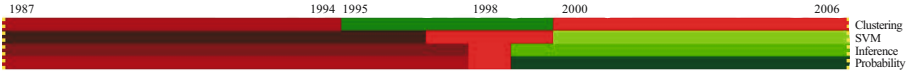


Fig. 6. The four topics intensity description (more red, more rise; more green, more fall)

Classification, and the topic Probability is related with the topic Classification, HMM and Recognition because it introduces new theory methods to these methods and applications. These results are consistent with our understanding about this field. In the following, we show the detailed relation with a part of CPT (conditional probability table) as Table 1.

We take the module SVM and Prob (Inference, Probability) as examples. As Table 1(a), when Clustering (Clu) rises (U), the influence is not obvious; when Clustering (Clu) falls (D), the influence is negative. Through the detailed description of topics (Clustering, SVM, Inference, Probability) trend in Fig. 6, we can find that, in the beginning, with the development of NIPS, the four machine learning methods increase synchronization, but in recent years, with the steady of development, the influence is obvious. Fig. 6 shows that recently SVM is positively correlated with Prob topic (Inference and Probability), but Clustering is negatively correlated with them, which is consistent with Table 1(b). And recently, when Clustering rises, SVM falls, which complements the result in Table 1(a).

5 Conclusion

This paper investigated the problem of topic influence based on the topic intensity evolution and used the topic intensity time series to construct the module network for understanding the topic influence. Future work includes improving the accuracy of topic intensity by virtue of topic tracking technology, introducing the semantic information of topic, considering the time series information in module network building.

Acknowledgements

This work was supported by the National Natural Science Foundation of P.R.China (No.60402010) and Zheiang Provincial Natural Science Foundation of P.R.China (Y105250).

References

[1] Mei, Q.Z., Zhai, C.X.: Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In: KDD 2005. Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp. 198–207. ACM Press, New York (2005)

[2] Zhou, D., Ji, X., Zha, H., Giles, C.L.: Topic evolution and social interactions: how authors effect research. In: CIKM 2006. Proceedings of the fifteenth ACM international conference on Information and knowledge management, pp. 248–257. ACM Press, New York (2006)
[3] Wang, J.L., Xu, C.F., Li, G., Dai, Z.W., Luo, G.J.: Understanding research field evolving and trend with dynamic bayesian networks. In: PAKDD 2007. Proceedings of the eleventh Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 320–331. Springer, Heidelberg (2007)
[4] Segal, E., Pe'er, D., Regev, A., Koller, D., Friedman, N.: Learning module networks. Journal of Machine Learning Research 6, 557–588 (2005)
[5] Landauer, T., Foltz, P., Laham, D.: Introduction to latent semantic analysis. Discourse Processes 25, 259–284 (1998)
[6] Hofmann, T.: Probabilistic latent semantic indexing. In: SIGIR 1999. Proceedings of the twenty-second annual international ACM SIGIR conference on Research and development in information retrieval, pp. 50–57. ACM Press, New York (1999)
[7] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet aladdress. Journal of Machine Learning Research 3, 993–1022 (2003)
[8] Wang, X.R., McCallum, A.: Topics over time: a non-markov continuous-time model of topical trends. In: KDD 2006. Proceedings of the twelfth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 424–433. ACM Press, New York (2006)
[9] Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: ICML 2006. Proceedings of the twenty-third international conference on Machine learning, pp. 113–120. ACM Press, New York (2006)
[10] Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann Publishers Inc, San Francisco (1988)
[11] Keogh, E.J., Pazzani, M.J.: An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. In: KDD 1998. Proceedings of the fourth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 239–243. AAAI Press, New York (1998)