ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/259226131

Clustering and Recommendation of Scientific Documentation Based on the Topic Model

Article *in* Lecture Notes in Electrical Engineering · November 2013

DOI: 10.1007/978-3-642-34522-7_67

CITATION
1

READS
125

1 author:

Bin Liao
North China Electric Power University
31 PUBLICATIONS **266** CITATIONS

SEE PROFILE

All content following this page was uploaded by Bin Liao on 29 July 2015.

The user has requested enhancement of the downloaded file.

Chapter 67
Clustering and Recommendation of Scientific Documentation Based on the Topic Model

Bin Liao, Weihua Wang and Chunmei Jia

Abstract In this paper, we propose a novel and efficient method for topic modeling and clustering of scientific documentation. It is a technology of content-based filtering and aims to find the same topics. Incorporating topic features will enhance the accuracy of document clustering methods. Based on the clustering results, we use the method of calculating similarity of scientific documentation to get the related documentation consistent with the content. The ranking of recommendation is according to the value of similarity of documentation. Clustering results are evaluated by F-measure. Empirical study on real-world datasets shows that the LDA's performance is better than PLSA's in the document clustering. Meantime, we find the proper number of topics in document representation.

Keywords Scientific documentation · Recommendation · LDA · Cluster

67.1 Introduction

Searching scientific documentation is an effective way for us to acquire knowledge and skills. Users get feedback webpages, when searching papers on search tools such as CNKI or Google Scholar. The result not only includes abstract, information of author and published information but also recommends a number of related papers on the terms of retrieving paper. The related papers of recommendation are from the three main sources: sometimes the related papers share the

B. Liao · W. Wang (✉) · C. Jia
Electrical and Electronic Engineering School, North China Electric Power University, No 2 Beinong Road, Beijing, Changping District, China
e-mail: wangweihua@ncic.ac.cn

W. Lu et al. (eds.), *Proceedings of the 2012 International Conference on Information Technology and Software Engineering*, Lecture Notes in Electrical Engineering 211,
DOI: 10.1007/978-3-642-34522-7_67, © Springer-Verlag Berlin Heidelberg 2013

same words in their titles, sometimes they hold the same keywords, or are from references. Those results are inseparable from the retrieving paper which is convenient for users to acquire related knowledge and expand our horizons in the field. Although the three above-mentioned ways can mostly cover scope of related papers, the result is not based on the content of the paper which is the most important thing. The content is the main part of paper certainly, so ignoring the content makes the recommended information incomplete. In many cases users focus on information of the top five which are considered to be the most relevant, the recommended sequence in the paper is also important. Users select information difficultly only relaying solely on appearance of keywords to determine accurately the recommended order. In reality some inconsistencies papers on the titles or keywords, but very close to their research and discussion the contents of these recommended information to be omitted, on the recommendation of the integrity of information left a regret.

Clustering of document is a classic problem in text mining, which overcomes the defect of tradition information retrieval [1]. That a good clustering method can quickly and efficiently with the assistance of the computer division of the document types is convenient browsing and navigation for users. By clustering the text document, the documents sharing the same topic are grouped together. Different from classification, clustering is a way of unsupervised leaning which is less human intervention and higher level of automation. An outstanding algorithm of clustering not only easily indicates the topic but also the differences of each others. A perfect document representation and an outstanding algorithm of clustering are two significant things which decide the quality of clustering result [2].

In this paper, we propose a novel and efficient method for topic modelling and clustering of scientific documents. It is a technology of content-based filtering and aims to find the same topics. Using technology of clustering greatly reduce computation of similarity among papers. Based on the clustering results, we use the similarity method of calculation of scientific documents to get documents consistent with the content. The ranking of recommendation is according to value of similarity of documents. Our framework contains three layers to complete the design goal. The first layer is for document representation of topic model which is based on latent dirichlet allocation (LDA) model. The second layer is for document clustering which is on the basis of K-means cluster algorithm. The third layer is to calculate the values of similarity of documents based on LDA model.

The following section is involving about the related works and Sect. 67.3 about proposed works based on the three layers. Section 67.4 gives the result of experiment. Section 67.5 concludes the paper and discusses about the future work.

67.2 Related Works

Proper and excellent document representation is good beginning of text mining. Vector Space Model (VSM) as a typical model of bag of words is a commonly and widely used in document representation for clustering and classification. VSM is

seen as a matrix of document collection and a feature vector represents single document. Weight of vector hold a variety of reorientation such as term frequency (TF) and term frequency-inverse document frequency (TF-IDF). TF-IDF fully take the value into account together in a single document and multi-document sets. In large set of documents vector holds thousands of dimensions, which makes matrix high sparsity. Probabilistic Latent Semantic Analysis (PLSA) [3] and LDA [4] which are two main probabilistic topic models complete the target of mapping the term-document representation to a lower-dimensional latent semantic space and treat the document as a mixture of topics. The two algorithms aim to analyze the words of original texts to discover the topics that run through them, how those topics are connected to each other. Also they do not require any prior annotations or labeling of documents.

PLSA and LDA are both deformation of VSM in low-dimension and the value of dimension could be determined by people. In PLSA an LDA, the conditional probability between terms and documents is modeled as a latent variable. Sometimes LDA which assumes that a document's topic distribution has a Dirichlet prior and that this simplifying assumption improves the Bayesian inference process shows better performance than PLSA. Another advantage of LDA is that it allows us to interpret each topic by looking at the words with a high probability of selection in that topic [5].

In [6], the dimensionality reduction via PLSA and LDA results in document clusters of almost the same quality as those obtained by using original feature vectors. And the result suggests that no difference between them for dimensionality reduction in clustering. In [7], experimental results have shown the effectiveness of the framework in clustering documents according to their mixtures of topics, and have highlighted the advantages offered by employing state-of-the-art topic model and their combination with the Bhattacharya distance.

In [8], Google Inc. generates recommendations using three approaches: collaborative filtering using MinHash clustering, PLSA and covisitation counts. Their approach is content agnostic and consequently domain independent, making it easily adaptable for other application and languages with minimal effort. In [9], after documents clustering based on LDA, a modified iterative PageRank algorithm ranked the research papers in a topic which assign an authoritative score to each paper based on the citation network.

67.3 Our Approach

In this section, we describe the proposed method in detail. As mentioned in introduction, the method is divided into three layers in Fig. 67.1, which is also treated as a linear model combing with many different algorithms for recommendations.

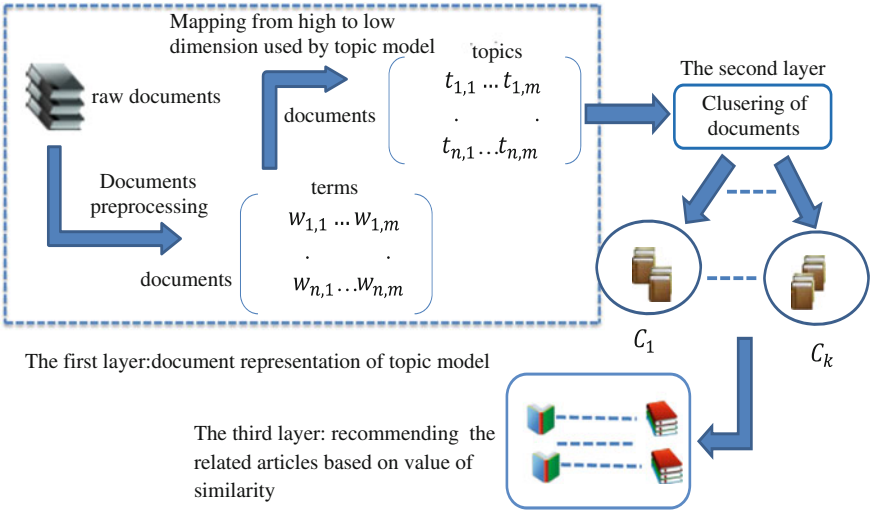


Fig. 67.1 The framework of method in three layers

67.3.1 Document Representation

The format of most scientific documents is PDF which is structured and hard to be processed by computer. *Pdf2txt*[1] is a good tool for converting pdf to txt. In the document preprocessing phrase, the paragraphs and sentences are segmented to isolated words. Through porter stemming[2] algorithm and be filtered by stopwords vocabulary, the document is represented as the format of term and TF. Let $D = \{d_1, d_2, \ldots, d_m\}$ is a set of documents. Let $W = \{w_1, w_2, \ldots, w_n\}$ is the vocabulary of terms in the set of documents, and the size of vocabulary is n.

When constructing the matrix of document collection, the weight of vector is the value of TF-IDF. Although porter stemming and stopwords filtering effectively decrease the scale of vocabulary size, the dimensionality is still high. Further using value of TF-IDF removes the words which are popular in document collection. We select a proper threshold of TF-IDF to accomplish the target of dimensionality continually. Let θ is the threshold and denotes that the percent of terms with TF-IDF. Within the scope of θ, the terms is dismissed.

$$TF - IDF(d_m, w_n) = TF(d_m, w_n) \bullet \log\left(\frac{|D|}{df(w_n)}\right) \tag{67.1}$$

We get the VSM of documents collection and the terms are isolated and not related with each others. But those words are closely linked and the combination of

[1] http://www.foolabs.com/xpdf/download.html

[2] http://tartarus.org/martin/PorterStemmer/

words constitutes the semantic information. Next we apply the LDA to mine the potential semantic information among words. After associated words coming together to form the topic, the vector of document is denoted by the fixed topics. Let $d = \{t_1, t_2, \ldots, t_k\}$ is representation of topics of document. Based on LDA, we draw the probability distribution among the topics and each document. We could clearly determine the topic of document belongs by the feature value.

The weights of document vector based on LDA are clearly show the topics which the document belongs to. Using technique of clustering automatically divides the corpus into sub-section which holds the same topics. Then, a clustering method, K-means, is performed on the vector space of document collection.

67.3.2 Document Clustering

In the phase of document representation, the topics are constituted by the collection of words. For example the *computer* topic has words about software, hardware, network with high probability. The center points of each cluster are shown by topics as same as the documents.

Algorithm 1. Documents with multiple topics clustering

Input: document collection representation based on LDA: $D = \{d_1, d_2, \ldots, d_m\}$, specified clustering value of k.

Output: clustering collection of document $C = \{c_1, c_2, \ldots, c_m\}$

1. Calculate Euclidean distance $d(d_i, d_j)$ in the document collection D, find the two vectors which hold the farthest distance in collection, make the two as initial cluster-points p_1, p_2 and delete them from collection D.

2. Find the vector which is the farthest distance with p_1 and p_2 in D, make it as p_3, and delete it from D.

3.Repeat the step 2 and find out the rest initial cluster-points $P = \{p_1, p_2, \ldots, p_k\}$

4. Calculate the Euclidean distance of every document point with cluster-points and each document is assigned to the nearest cluster.

5. Calculate mean value of cluster as new cluster-point

6. Repeat step 3 and step 5 until each cluster centroid is no longer change or reach a specified number of iterations.

67.3.3 Similarity Measure

Clustering finds out the documents which hold similar contents. We could not exhibit the all result of clustering because we recommend targeted information for every document. In sub-sections we calculate the similarity of each document with the other documents. In this proposed work, cosine similarity measure is the similarity used.

$$sim(d_1, d_2) = \cos\theta = \frac{d_1 \bullet d_2}{|d_1| \bullet |d_2|} = \frac{\sum (t_{1,i} \bullet t_{2,i})}{\sqrt{\sum t_{1,i}^2 \bullet \sum t_{2,i}^2}} \tag{67.2}$$

67.4 Experiment and Results

67.4.1 Datasets and Method of Evaluation

The recommended information is based on accurate result of clustering, and to show the accuracy of our algorithm and evaluate the experiment we prepare two test datasets. The first set is from Sougou laboratory[3] and the second set is from 20_newsgroups.[4] The set of SogouCS contains 18 channels such as Olympic games, sports, IT, the domestic and the international from Sohu News during January 2008 to June. The 20_newsgroups dataset is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. In our experiment, we select four sorts to verify the accuracy of clustering.

We use F-measure to evaluate the quality of document clustering based on topic model. It combines the precision and recall of test datasets to compute the accuracy. The F-measure can be denoted as a weighted average of the precision and recall where F-measure is range from 1 to 0. Furthermore 1 is the best value and 0 reaches the worst. The recall and precision of the cluster for each given class are calculated.

$$F - measure(clustering) = \frac{2}{k} \bullet \sum_{i=1}^{k} \frac{precision(i) \bullet recall(i)}{precision(i) + recall(i)} \tag{67.3}$$

The number of topics determines the size of document dimension based on topic model and will also impact the accuracy of clustering. We also adopt two topic models involving LDA and PLSA to calculated F-measure in those datasets. Figure 67.2 shows that LDA performs better than PLSA and with the number of topic increasing, the accuracy of clustering have an improvement. In comprehensive consideration of the number of clustering and the class of topic model, we take topic's value of 30 and LDA in subsequent experiments. In table, the value of θ mentioned in Sect. 67.3.2 is to filter the usual words in dataset and the result shows that θ range from 5 to 10 % effectively increase the F-measure (Table 67.1).

[3] http://www.sogou.com/labs/resources.html

[4] http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html

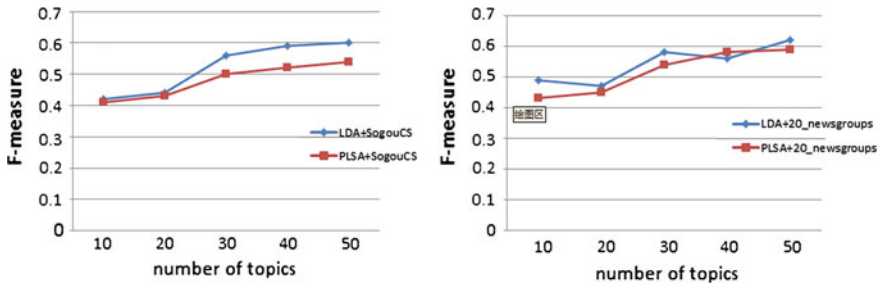
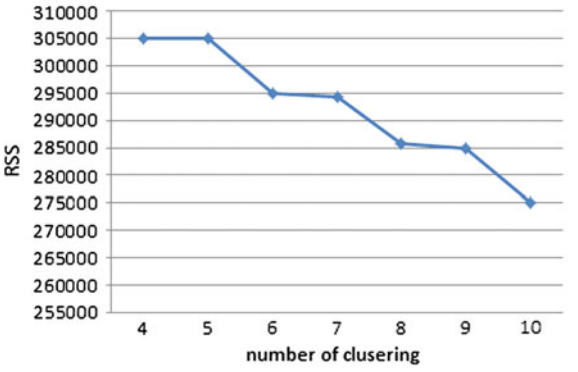


Fig. 67.2 Clustering performance by varying the clustering size on SogouCS and 20_newspapers

Table 67.1 F-measure on different value of θ

Dataset	θ	F-measure
SogouCS	0	0.56
	5 %	0.6
	10 %	0.61
20_newsgroups	0	0.58
	5 %	0.62
	10 %	0.62

Fig. 67.3 The value of RSS by varying the clustering size



67.4.2 Result

To fulfill the final target of finding out the related documents for every scientific document, the previous steps which make each cluster corresponding to a topic have laid a good foundation. We have crawled 20,000 scientific documents' abstracts from ACM. Different from those two datasets, the scientific document collection don't have fixed number of class. So we want to guess the reasonable value and get the optimal number of clustering. We use value of RSS (residual sum of squares) to achive. Figure 67.3 shows that ranging from 4 to 10 the number of

Table 67.2 Top three the most related documents for each scientific document

Topic	Scientific document	Top 3 related documents
Computer applications	Multi-processor system design with ESPAM	A framework for rapid system-level exploration, synthesis, and programming of multimedia MP-SoCs
		Automated modeling and emulation of interconnect designs for many-core chip multiprocessors
		EWD: A metamodeling driven customizable multi-MoC system modeling framework
Hardware	A model for distributed systems based on graph rewriting	Verifying parameterized networks
		Accelerating multi-party scheduling for transaction-level modeling
		Caisson: a hardware description language for secure information flow
Theory of computation	Compositional verification of concurrent systems using Petri-net-based condensation rules	Modeling time in computing: A taxonomy and a comparative survey
		Software model checking
		Using formal specifications to support testing

clustering is 6 which is considered as the inflection point. Upon implementing the algorithm of clustering based on LDA, we get six sub-sections. For each cluster, we have calculated the similarity of each scientific document with the others by cosine similarity. This can be very useful to researchers studying a given documents. Table 67.2 shows the list of the top three documents for retrieving documents in different topic.

67.5 Conclusions

In this paper, to accomplish the purpose of finding out the related documents for every scientific document, we propose a method which derives topics and clusters those documents. Through experiment, we find out the most appropriate document representation and number of topic to get optimal result. In practical application, we extend the traditional method and the users could get more information.

Our future enhancement is the clustering algorithm to bring more precise divisions. The documents which are in the cluster boundary find out the related not only in the divided cluster but also in neighboring clusters. A better similarity measure is needed to improve the accuracy.

References

1. Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval, Cambridge University Press, Cambridge
2. Rong X, Liang K, Yan Z, Min W (2011) CDW: a text clustering model for diverse versions discovery. In: FSKD 2011, pp 1113–1117
3. Hofmann T (1999) Probabilistic latent sematic analysis. Uncertainty in artificial intellogence
4. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
5. Blei DM Introduction to probablilistic topic models, http://www.cs.princeton.edu/∼blei/topicmodeling.html
6. Masada T, Kiyasu S, Miyahara S (2008) Comparing LDA with pLSI as a dimensionality reduction method in document clustering. In: LKR 2008, pp 13–26
7. Ponti G, Tagarelli A (2009) Topic-based hard clustering of documents using generative models. In: ISMIS 2009, pp 231–240
8. Das A, Datar M, Garg A, Rajaram S (2007) Google news personalization: scalable online collaborative filtering. In: WWW 2007, pp 271–280
9. Shubankar K, Singh A, Pudi V (2011) A frequent keyword-set based algorithm for topic modeling and clustering of research papers. In: DMO 2011, pp 96–102

View publication stats