RESEARCH ON CHINESE MULTI-DOCUMENT HIERARCHICAL TOPIC MODELING AUTOMATIC EVALUATION METHODS

Yu Liu[1], Lei Li[1] ,Shuhong Wan[1],Zhiqiao Gao[1]

[1] Center for Intelligence Science and Technology, School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 10086 ,China
147121500@qq.com

Abstract: Hierarchical Latent Dirichlet Allocation (hLDA) has achieved good results in the supervised and unsupervised multi-document hierarchical topic modeling. However, the result is diversified. The results maintain randomness even with the same parameters. Thus, this paper proposed automatic evaluation methods for unsupervised multi-document hLDA modeling results over previous studies. This paper used 10 topics of corpus of ACL2013 multilingual multi-document summarization and found 90 topics of news as experimental corpus, then compared the different modeling results. The results showed that automatic evaluation method can provide a good reference for the optimization of the modeling results.

Keywords: Hierarchical LDA; Hierarchical Topic Modeling; Automatic Evaluation Methods

1 Introduction

Blei et al [1] proposed the hierarchical Latent Dirichlet Allocation (hLDA) topic model in 2004, and proposed a non-parametric methods based on the nested Chinese restaurant process model , this method has a strong theoretical basis, it established relationship between topics by the levels of the semantic tree , and can automatically determine the number of topics based on the data. The hierarchical Dirichlet topic model is applied to multi-document summary in English by Asli and Dilek [2] in 2009, and they achieved remarkable results. Hongyan Liu, etc. [3] used the hLDA topic model and statements compression in Chinese multi-document summary and participated TAC 2010 evaluation, and achieved initial results. Pingan Liu, etc. [4] applied hLDA in the study of Chinese clustering and multi-document summary , they have also achieved very good results. Wei Heng et al [5] study the key factors of hLDA topic modeling , then gives an effective modeling strategies and processes. Jia Yu et al [6] study the sentence scoring method based on the hLDA topic model, it provide a strong basis for extraction of multi-document summary sentences, and achieve better results in ACL2013 evaluation.

The article is organized as follows: Section II describes the topic model, the relevant evaluation method of clustering results; Section III describes the experimental corpus of content, process and methods of experimental parameter adjustment; Section IV proposed an automatic evaluation method for hierarchical topic model from the clusters` number and similarity ; Section V compared the hLDA results with other modeling results based on the automatic evaluation method, and adjusted the parameters according to the hLDA evaluation results in different themes.

2 Related work

There is no directly relevant information found of the automatic evaluation for hierarchical topic modeling results. It has mainly two types of indirectly related information, one is the evaluation methods for LDA topic divide, such as perplexity [7], the important standard for the evaluation of the quality of the language mode, also the most widely used method for assessing the performance of LDA modeling results; Boyd-Graber et al [8] using high probability vocabulary inside the topics as the output, first proposed a system of manual evaluation method for language model. The other is the automatic evaluation of clustering results, because the results of hierarchical topic modeling are equivalent to the task of automatic clustering. It has many kinds of clustering evaluation methods been proposed so far, such as Davies-Bouldin's Index [9], CS Index [10], Dunn`s Index [11] and so on. In recent years, there are many studies on the clustering results evaluation. Yunjie Zhang et al [12] proposed an evaluation method for fuzzy clustering using change measurement and separation metrics; Jeen-ShingWang et al proposed optimization method for the support vector clustering mixing parameter [13]; Xulei Yang et al proposed the use of simulated annealing method combines clustering evaluation parameters were optimized for clustering methods [14]. Although the starting point of these methods are not the same, but in general are based on the similarity within and between the clusters of the results.

This article is for hLDA Chinese multi-document unsupervised hierarchical topic modeling, it proposes an automatic evaluation method for modeling results considering the number of clusters and similarity to assist enhancing and to ensure good results of the unsupervised hierarchical topic modeling.

978-1-4799-4719-5 /14/$31.00 ©2014 IEEE

Authorized licensed use limited to: National University of Singapore. Downloaded on February 16,2025 at 02:51:37 UTC from IEEE Xplore. Restrictions apply.

3 HLDA modeling experiments` corpus and process

3.1 Experimental corpus

The experimental corpus comes from two parts, a total of 100 topics totaling 1000 document. One is the ACL MultingLing 2013 multi-language multi-document summary evaluation of Chinese data released, a total of 10 topics, each topic contains 10 News (M000-M009), e.g. the Indian Ocean tsunami (M000). The other is the collection of domestic news portal reported that a total of 90 topics, and each topic has also 10 reports (M010-M099), e.g. Iran sanctions (M019).

3.2 Experimental procedure

3.2.1 Pretreatment

First, merged the 10 document within the same topic to one document; then processed sentences segmentation and filtered out sentences that is too short; followed by a word segmentation processing for each sentence; then removed the stop words from segmentation results according to the dictionary; finally, based on that results, generated the word frequency dictionary and "word number: word frequency" for the format of the input file of hLDA modeling.

3.2.2 HLDA modeling

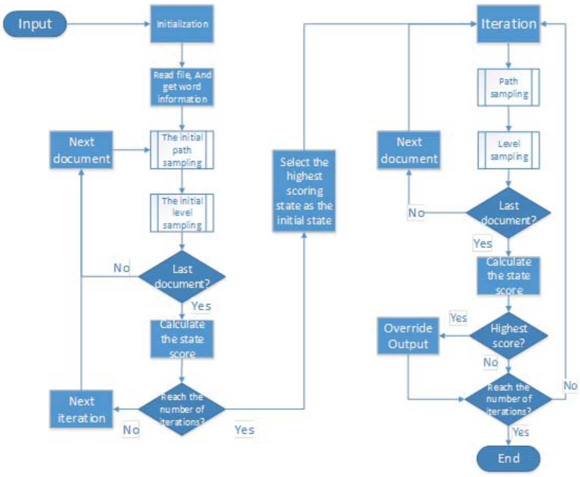
Input
Initialization
Read file, And get word information
Next document
The initial path sampling
The initial level sampling
No
Last document?
Yes
Calculate the state score
Next iteration
No
Reach the number of iterations?
Yes
Select the highest scoring state as the initial state
Iteration
Path sampling
Next document
Level sampling
No
Last document?
Yes
Calculate the state score
Override Output
Yes
Highest score?
No
Reach the number of iterations?
No
Yes
End

Figure 1 hLDA modeling flow chart

Shown in Figure 1, the two sub-critical process of hLDA modeling are document path sampling and document level sampling, through the two iterative process to get the best model of hLDA topic distribution tree. In the end, each sentence in documents is allocated in the different branches of the tree and each word is allocated in the different nodes in order to achieve the purpose of clustering.

3.2.3 Parameter adjustment

In the modeling process, setting a few of priori parameters plays a vital role, but the same set of parameters may also get totally different result with different corpus. Therefore, the formation of a set of empirical parameter adjustment method is very meaningful. This paper uses the same name of parameters as they are used in [1]. Each node`s η determines the probability of one node`s words, the greater η the probability of words that the node contains more similar, the smaller η the nodes prefer a few of words. In the document path sampling process, SCALING_SHAPE and SCALING_SCALE affect the path distribution of the model. The product of SCALING_SHAPE and SCALING_SCALE bigger, it more inclined to choose a new path when the document selected path, and thus the total number of paths become greater, otherwise smaller. When the product is the same, a larger SCALING_SCALE will make the number of clusters relatively reduced. In the course of sampling the document hierarchy, the m bigger, more word aggregation to the leaf node, which tend to be more specific topics, and vice versa. The π determines the size of the rigor of this tendency, the greater the π make words more strictly in accordance with this tendency distribution.

4 Automatic evaluation for hLDA modeling results

Admittedly, the evaluation for the hLDA hierarchical topic modeling results should be best done by human experts, because the result is reflected in a number of different topics covered in multi-document collection and the relationship between them, the result that accorded with human cognition should be the best perception. However, with the increasing amount of data, only completing the evaluation of all the results from the human is very difficult, and the speed cannot meet the needs. To this end, there is an urgent need for automatic evaluation methods. Although the accuracy of the automatic evaluation cannot be compared with the human being, it can evaluate the modeling results to a certain extent, thus to assist optimizing the results. Based on the analysis of manual evaluation process of the modeling results, this paper chose from clustering number and similarity of these two aspects for automatic assessment of hLDA's modeling results.

4.1 Evaluation of cluster number

Because the cluster number in hLDA is automatically generated based on the data, the numbers are the most direct expression of the modeling results. For the cluster number of 10 documents concerned one news topic, usually too few (3 or less) or too much (20 or more), are not in line with the results of human cognition. Therefore, the number of clusters is a preliminary assessment of the results.

4.1.1 Frequency proportion estimate cluster number

The word frequency dictionary in each topic without stop words could be got based on the results after pretreatment. The high frequency words in the document often represent the theme in one direction, so this paper, based on the number of high-frequency words in the document, proposed a method to estimate the ideal

Authorized licensed use limited to: National University of Singapore. Downloaded on February 16,2025 at 02:51:37 UTC from IEEE Xplore. Restrictions apply.

number of clusters in modeling results.

Difficulty with this method is how to determine the high-frequency words. The method used in this paper is if the word frequency proportion of each word in the total word frequency is greater than a certain threshold within the topic, the term will be considered high-frequency words. Shown in Figure 2, a more appropriate word frequency threshold ratio is between 0.008 and 0.010, thus, each topic can be taken 6 to 13 or more high-frequency words. Figure 3 to Figure 5 compared 0.008, 0.009, 0.010 these three thresholds and found the ideal cluster number was a bit more at the threshold 0.008, and less slightly at the threshold 0.010, it was between 6 and 13 at the threshold 0.009, mostly closed with number of manual sampling evaluation in 4.13, which is more in line with people's perception, and therefore, we believe that the word with frequency ratio 0.009 or more is high-frequency words.

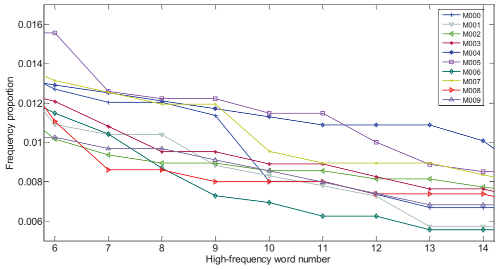

Figure 2 ideal cluster numbers under several different word frequency proportional thresholds

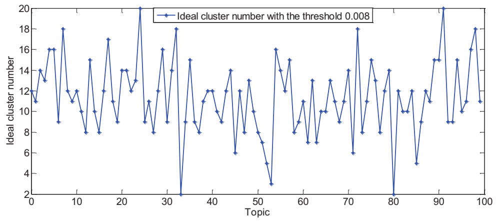

Figure 3 ideal cluster number under the threshold 0.008

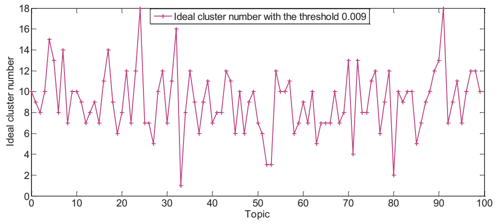

Figure 4 Ideal cluster number under the threshold 0.009

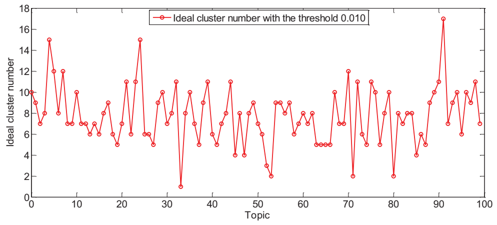

Figure 5 Ideal cluster number under the threshold 0.010

4.1.2 LSA cluster number evaluation

LSA (Latent Semantic Analysis) is using a matrix to represent words and documents, and then using the singular value decomposition on the matrix dimensionality reduction, thus the original vector space is mapped to underlying semantic space, removing irrelevant information in the document, showing its semantic structure. This paper selected the potential dimension r is based on the conclusion [15] that the selected dimensions of the singular value decomposition is not less than half of the maximum singular value of the configuration of all the singular values.

4.1.3 Evaluation methods of cluster number

High-frequency words number and potential dimensions these two reference values of document cluster number can be obtained from word frequency ratio and LSA, when the two values are similar and consistent with the document topic number, there is considerable reference value. When the cluster number of modeling results is more similar to two values, the evaluation of the modeling results will be higher. As shown in Table I, M000, M001, M006, M008, M009, M011, M012 these topic`s two reference values of cluster number are very close, it is believed that these values are approximate to the ideal cluster number of the topic. When adaptive, we should try to make the cluster number closed to these reference values. As for the big gap between the two reference values in other topics such as the M002, it is still necessary to manually analysis the corpus`s characteristics to determine the ideal cluster number.

Table I The reference values of ideal cluster number given by LSA method and high frequency words method

Topic	LSA	Frequency
M000	11	10
M001	7	9
M002	3	8
M003	7	10
M004	5	15
M006	8	8
M008	7	7
M009	8	10
M010	4	10
M011	10	9
M012	8	7

This article sampled 20 topics to score manually, on the reference values of cluster number in each topic given by LSA, the number estimated by word frequency ratio and the cluster number of hLDA actual modeling results. Four different raters carefully read these 20 topics corpus, and got a certain understanding that how many parts each topic contents covers, then gave an objective score to the three cluster number. The given number of topic clusters exactly the same as the parts the topic contents have, which is the number of sub-topics, will get 5 points, the one far from the number get 1 point, between the conformity from low to high will get 2-4 points. Score results shown in Table II, the number of

Authorized licensed use limited to: National University of Singapore. Downloaded on February 16,2025 at 02:51:37 UTC from IEEE Xplore. Restrictions apply.

the sub-topic of the article are substantially covered between 6 and 13, which is consistent with the previously mentioned figures. The closer the reference value given by LSA and the cluster number estimated by term frequency are, they get a higher reference value, the closer the actual cluster number of modeling results and the reference value are, the score is higher. When the two reference values are far away, the reference value between 6 and 13 has more reference significance.

Table II Cluster number and manual scoring (values in the table are: the number of ideal or actual clustered / manual score)

Topic	LSA	Frequency	hLDA
M004	5/4	15/1	4/4
M012	8/4	7/4	10/3
M020	6/4	8/4	9/3
M029	5/2	12/4	10/4
M034	7/3	8/3	10/2
M039	8/3	11/2	6/5
M042	8/4	8/4	10/3
M050	9/3	7/4	12/1
M051	7/4	6/4	7/4
M055	7/2	10/1	7/2
M058	4/3	6/4	9/2
M075	5/4	11/2	11/2
M085	13/1	5/4	10/2
M098	7/3	12/1	9/2

4.2 Similarity Evaluation

The evaluation of cluster number is still relatively surface, and we often get a variety of results with cluster number that meets the requirements. In order to further evaluate the clustering effect, we also used the similarity evaluation methods that commonly used to evaluate clustering results.

4.2.1 Similarity calculation method

This paper used cosine similarity. The cosine cosθ of the angle between two sentence vector represented by the TF-IDF (term frequency - inverse document frequency) value of each word in the sentence like $a(x_1,x_2,\ldots,x_n)$, $b(y_1,y_2,\ldots,y_n)$, was considered as the similarity between the two sentences.

(4-1)

4.2.2 Similarity Evaluation Content

Similarity evaluation content mainly includes five areas: intra-cluster similarity, intra-cluster variance, inter-cluster similarity, inter-cluster variance, the quotient between intra-cluster similarity and inter-cluster similarity. Calculate the average similarity between each sentence within each cluster as an intra-cluster similarity of the cluster, and weight these averages in accordance with the proportion between the number of sentences in the cluster and the total number of sentences in all clusters of each topic as the intra-cluster similarity of the topic. According to the intra-cluster similarity of the topic and each intra-cluster similarity of the clusters in the topic, we can calculate the topic intra-cluster variance. Compute the similarity between all two sentences in one cluster(e.g., sentence a, b in the cluster A, and sentence c, d in the cluster B, the similarity is calculated between ac, ad, bc and bd), and the average of these similarity is the inter-cluster similarity between the two clusters. Thus we get the average of the inter-cluster similarity between all two clusters in one topic as the inter-cluster similarity of the topic. According to the inter-cluster similarity of the topic and each inter-cluster similarity of the clusters in the topic, we can calculate the topic inter-cluster variance.

In evaluating the modeling results, the main observation is that the similarity between the cluster and the cluster similarity, while the variance between the cluster and the cluster variance small fluctuations can be. We hope that the similarity in the cluster as large as possible, and the similarity between the clusters as small as possible, so that clustering results to ensure consistency within the cluster of sentences and to ensure a distinction between the degree of clustering. To be able to considering the similarity between the cluster and the cluster similarity to cluster similarity between the results of the comparative cluster similarity with the cluster simultaneously changed when we added the inter-cluster similarity and the similarity of the cluster than as modeling results are an improvement final considerations.

5 Experiment

This part used the above-mentioned automatic evaluation methods to compare the results of different modeling results through several experiments.

5.1 hLDA comparison with other modeling results

This paper compared three modeling results which are hLDA, LDA with k-means clustering method and direct sentences TF-IDF values with k-means clustering method. From figure 6 to 9, it can be clearly seen that the hLDA modeling results got a much higher intra-cluster similarity, and the inter-cluster similarity is much lower compared with the other two. We scored the three modeling results on those topics at the same time when we did the cluster number manual scoring process as above-mentioned. Theme within the cluster is not clear, and little difference between the clusters got the worst 1 point, a high degree of aggregation within a cluster, clear theme, a lot of discrimination between clusters got the best 5 points, the modeling results between the given from low to high got 2 to 4 points. The results are shown in Table III, it is obvious that hLDA modeling results was much better than the other two modeling results which was very consistent with the automatic evaluation results, which demonstrate the effectiveness of automatic evaluation method proposed in this paper. It can also be seen hLDA modeling has superiority in terms of text clustering.

Authorized licensed use limited to: National University of Singapore. Downloaded on February 16,2025 at 02:51:37 UTC from IEEE Xplore. Restrictions apply.

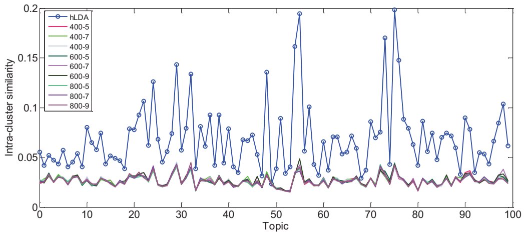

Figure 6 LDA and hLDA modeling results intra-cluster similarity comparison (400-5 said LDA topic number 400, k-means clustering number 5)

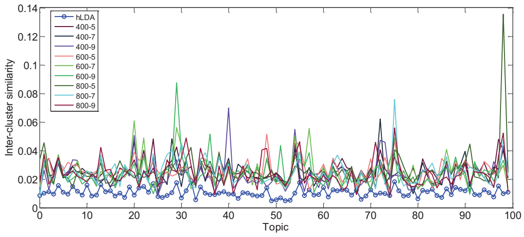

Figure 7 LDA and hLDA modeling results inter-cluster similarity comparison (400-5 said LDA topic number 400, k-means clustering number 5)

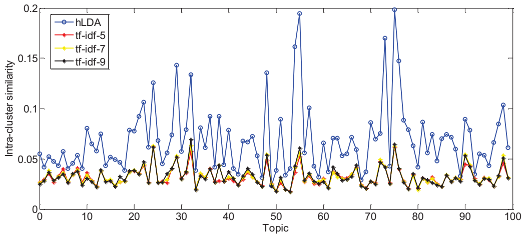

Figure 8 HLDA and k-means directly using TF-IDF values clustering results intra-cluster similarity comparison (tf-idf-5 represents the number of k-means clustering is 5)

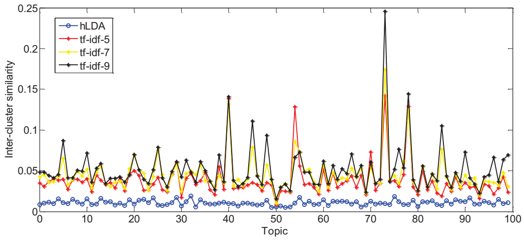

Figure 9 HLDA and k-means directly using TF-IDF values clustering results inter-cluster similarity comparison (tf-idf-5 represents the number of k-means clustering is 5)

5.2 Parameter adjustment according to different topics

Mentioned before, the same set of a priori parameter applying to all subjects in general is not possible, so when the actual and the ideal number of clusters has a big difference, it is necessary to adjust the parameters adaptively. Actually the parameter η determines the words` distribution of the probability distribution in the node. In the text clustering area, the probability of the words contained in the nodes in different topics is different, but the distribution of the distribution is stable, so that η is an area, the application-level parameters. It once got a better value from training then there was no need for a large adjustment according to different topics. And m and π determine the level and distribution of words, they don`t have directly affect for the path allocation of the document. So the parameters adaptive adjustment here focused on SCALING_SHAPE and SCALING_SCALE.

Table III Artificial score of three kinds of modeling results

Topic	hLDA	LDA	TF-IDF
M004	4	2	1
M012	4	2	1
M020	3	1	1
M029	5	2	2
M034	4	2	3
M039	4	2	2
M042	4	1	1
M050	3	1	1
M051	4	1	1
M055	3	1	1
M058	3	2	2
M075	4	2	1
M085	4	1	2
M098	3	2	2

Table IV Comparison of different parameters modeling clustered number and the ideal number of clustered

Topic	Ideal clusters		SCALING_SHAPE/ SCALING_SCALE				
M	LSA	Frequency	1/0.5	0.3/1	0.1/1	0.025/2	0.001/2
020	6	8	9	8	7	7	5
034	7	8	10	10	9	7	5
039	8	11	6	5	6	5	4
050	9	7	12	11	9	9	6
051	7	6	16	15	13	12	11

Table V The quotient between intra-cluster similarity and inter-cluster similarity of modeling results under different parameters

Topic	SCALING_SHAPE/SCALING_SCALE				
	1/0.5	0.3/1	0.1/1	0.025/2	0.001/2
M020	8.75	6.25	5.64	6.04	4.11
M034	5.57	5.31	5.47	4.6	4.38
M039	3.94	4.11	4.34	4.1	3.52
M050	6.91	6.92	6.49	6.06	4.6
M051	13.24	7.23	5.68	5.69	5.64

For the topic with a few more clusters, we can increase

Authorized licensed use limited to: National University of Singapore. Downloaded on February 16,2025 at 02:51:37 UTC from IEEE Xplore. Restrictions apply.

the product of SCALING_SHAPE and SCALING_SCALE to reduce the cluster number of the modeling results. Here adjusted the parameters which belonged to the poor hLDA modeling results, and the results shown in Table IV and Table V. Topic M039 still clustered less than the ideal number, it is needed to continue to increase the product of SCALING_SHAPE and SCALING_SCALE, and to observe the change of the quotient between intra-cluster similarity and inter-cluster similarity. If it increased or changed little it is better to select parameter values with the clustering results having a similar cluster number with the reference number. If the quotient of similarity decline much, it is necessary to consider whether the cluster number estimated is not ideal. The cluster number of Topic M051 is still much more than the ideal number in these parameters. When the cluster number is too large, the similarity artificially high, the value of reference declined, so it is necessary to continue to reduce the product of SCALING_SHAPE and SCALING_SCALE. And the cluster numbers of other topics were reached the number which is closer to the ideal value at different parameters. With the reduction in the number of cluster, the number of sentences in the cluster increased, the average similarity in cluster should decrease a little in general. So with the reduction in the number of cluster, a slightly lower similarity was acceptable, we should adjust the parameters to get a closer cluster number with the estimated value of ideal number on condition that the quotient between intra-cluster similarity and inter-cluster similarity increased or just had a little change.

6 Conclusions

In this paper, considering the stochastic hLDA modeling results and the complexity of solely manual evaluation on the results of hLDA modeling, we proposed a method for automatic evaluation of Chinese multi-document-level topic modeling on the basis of previous research. This automatic method can not only reduce the difficulty of manual evaluation, but also assist to enhance the effect of unsupervised Chinese multi-document hierarchical modeling. On this basis, the paper compared the hLDA modeling results with other modeling results and validated hLDA having superiority in Chinese multi-document topic modeling, and then adjusted hLDA parameter settings according to the results of the automatic evaluation. The evaluation method in this paper is only a preliminary exploration, but also requires a lot of in-depth research work in the future, which mainly includes two aspects, one is how to evaluate tree level division of the modeling results, and the second is to expand the study of more and better semantic features and parameters.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 61202247, 71231002, 61202248 and 61472046; EU FP7 IRSES MobileCloud Project (Grant No. 612212); the 111 Project of China under Grant B08004; Engineering Research Center of Information Networks, Ministry of Education; the Fundamental Research Funds for the Central Universities under Grant2013RC0304; the project of Beijing Institute of Science and Technology Information.

References

[1] Blei D M, Griffiths T L, Jordan M I, Tenenbaum J B. Hierarchical topic models and the nested Chinese restaurant process[M]. Advances in Neural Information Processing Systems 2004(16):106-114.

[2] Asli C and Dilek H. A hybrid hierarchical model for multi-document summarization[C]//Jan Hajic, Sandra Carberry, Stephen Clark (Eds.). Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden: The Association for Computer Linguistics 2010 ISBN 978-1-932432-66-4, 2010(7):815-824.

[3] Hongyan Liu. Multi-document Summarization Based on HLDA Hierarchical Topic Model [D].Beijing University of Posts and Telecommunications 2012.

[4] Pingan Liu. Chinese Multi-document Summarization Based on HLDA Hierarchical Topic Model [D]. Beijing University of Posts and Telecommunications 2012.

[5] Wei Heng, Jia Yu, Lei Li et al. Research on Key Factors in Multi-document Topic Modeling Application with HLDA [J]. Journal of Chinese Information Processing, 2013, 27(6): 117-127.

[6] Jia Yu, Lei Li. HLDA Based Sentence Scoring for Multi-document Summary [J]. Journal of Henan Science and Technology, 2013(22):273-274.

[7] Brown, Peter F. et al. An Estimate of an Upper Bound for the Entropy of English[J]. Computational Linguistics, 1992,18(1): 31-40.

[8] Boyd-Graber J, Chang J, Gerrish S, et al. Reading tea leaves: How humans interpret topic models[C]//Bengio Y, Schuurmans D, et al. Proceedings of the 23rd Annual Conference on Neural Information Processing Systems. British Columbia, Canada: Curran Associates, Inc. 2009(12), 288-196.

[9] Davies, DL, Bouldin, D.W. A cluster separation measure[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979, Vol. 1, No2: 224-227.

[10] Chou C.H, Su M.C, Lai. E. A new cluster validity measure and its application to image compression. Pattern Analysis and Applications, 2004, vol. 7, no. 2: 205-220.

[11] J. C. Dunn. Well Separated Clusters and Optimal Fuzzy Partitions[J]. Journal of Cybernetica, 1974, Vol. 4: 95-104.

[12] Yunjie Zhang, Weina Wang, Xiaona Zhang, Yi Li. A cluster validity index for fuzzy clustering[J]. Information Sciences, 2008,178: 1205-1218.

[13] Jeen-ShingWang, Jen-Chieh Chiang. A cluster validity measure with a hybrid parameter search method for the support vector clustering algorithm[J]. Pattern Recognition, 2008, 41: 506-520.

[14] Xulei Yang, Qing Song and Aize Cao. A new cluster validity for data clustering[J]. Neural Processing Letters, June 2006, Volume 23, Number 3: 325-344.

[15] Josef Steinberge, Karel Ježek. Using Latent Semantic Analysis in Text Summarization and Summary Evaluation[C]// Miroslav Beneš. Proceedings of the 7th International Conference ISIM. Rožnov pod Radhoštěm: Beneš, Miroslav, zemř. 2005.

Authorized licensed use limited to: National University of Singapore. Downloaded on February 16,2025 at 02:51:37 UTC from IEEE Xplore. Restrictions apply.