Clustering of News Topics Integrating the Relationship among News Elements

Jiaying Hou[1,2], Zhengtao Yu[1,2,*], Xudong Hong[1,2], Feng Li[3]

[1] School of Information Engineering and Automation, Kunming University of Science and Technology, 650500 Yunnan, China
[2] Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, 650500 Yunnan, China
[3]Logistics Science Research Institute of PLA, Beijing 100166, China
ztyu@hotmail.com

Abstract. To make full use of news document structure and the relation among different news documents, a news topic clustering method is proposed of using the relation among document elements. First, the word characteristic weight was calculated by the TF-IDF method based on word frequency statistics to generate document space vector and news document similarity was calculated using text similarity measurement algorithm to obtain the initial news document similarity matrix. Then, the initial similarity matrix was modified with the relation among different news elements as semi-supervised constraint information, the clustering of news documents was realized using Affinity Propagation algorithm, and news topics were extracted from news clusters. As a result, the construction of news topic model was finished. At last, the contrast experiments were performed on manually-annotated news corpus. The results show that the Affinity Propagation clustering methods integrating the relation among document elements can achieve a better effect than those without constraint information.

Keywords: relationship among news elements; news topic; affinity propagation algorithm; semi-supervised constraints.

1 Introduction

With the development of the Internet, a large amount of data renders internet users difficult to extract news topics, and have access to the cues of and relation among events of one topic. Against this backdrop, many mainstream portal websites launch popular events and news topics within a certain period through manual editors or computer tools. For example, Baidu news updates daily hot news rankings, and Sina micro-blog updates hot topic list according to the popularity of the news. This has brought great convenience to internet users. Therefore, it is important and urgent to realize the clustering of news topics through the machine learning.

The research on finding topics both at home and abroad mainly focuses on the establishment of specific topic models, which can be roughly divided into the following categories. One is vector space models. For example, Makkonen constructed space

M. Dong et al. (Eds.): CLSW 2016, LNAI 10085, pp. 441–448, 2016
DOI: 10.1007/978-3-319-49508-8_42

vector of such models according to persons, time and places [1]. Lee converted the problem of finding topics into the clustering problem of the news text. To be specific, he realized the clustering and found topics by extracting news important words as the characteristic to calculate text similarity [2]. The second category is probabilistic topic models. Through the integration of evolutionary clustering algorithm and topic models, Ahmed et al. realized topic detection and tracking. This integrated model can achieve a better result in news topic detection and tracking [3]. D.Newman extracted figure entity in the news, and then integrated figure entity characteristics and topic models. This integrated model can also achieve a better result in topic detection [4]. The third category is Graphic Models. In 2000, Ogilvie constructed a Graphic Models with the feature words as the centroid gathered from text sets [5]. In 2004, Cathy used WordNet [6] to convert similar or related words of text sets into the lexical chain. Such a model was also constructed by Liu Ming et al in 2010 by using How net language dictionary [7].

For some of the above methods, different news elements are also used to realize news topic clustering, but the relation among them is not taken into consideration. In fact, there can be many kinds of relation among different news elements, which play a great supporting role in finding news topic.

2 Affinity Propagation algorithm

In 2007, Frey et al proposed an AP algorithm [8]. In this algorithm, the similarity among data points was used as the basis, all data points were considered as potential representative point and trust propagation theory was applied to the exchange of information among data points to find the representative point for each data point, which can make the sum of the similarity among all the data points and their nearest representative point of similarity largest. Thus clustering was completed. In figure 1 as below:

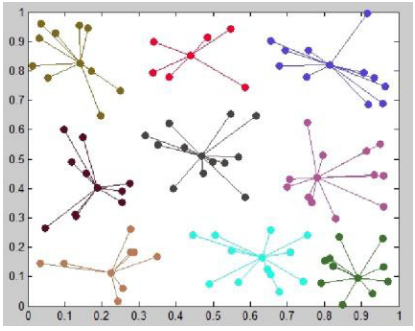

Fig. 1. Clustering results based on the algorithm of AP

AP algorithm was selected as the clustering algorithm based on similarity matrix after the adjustment. According to the literature [8-10], AP algorithm can achieve a higher accuracy and a better effect in processing large quantities of data, the accuracy and effectiveness of the algorithm is much higher than K-means [11].

3 Building the model of news topic clustering integrating relation among news elements

AP clustering algorithm can be seen as one kind of graph-based clustering. Each data point can be seen as a node in the graph, and AP is performed along the edges to find the optimal representative set of points. The most important two parameters in the algorithm are the similarity matrix S and the bias parameter P. The definition of the similarity matrix directly affects the performance of the clustering algorithm based on similarity matrix. The construction process of the news topic clustering model integrating relation among news elements is described in detail as follows.

3.1 Constructing the initial similarity matrix of the news-documents

"Topic detection and tracking is defined as the technique which can be used to automatically find topics in data flow from the newswire and broadcast news and link the piece of news with other pieces about the same topic"[12]. It can be seen that the research on topics is converted into the clustering problem of news text. Firstly, taking the new-pages collections as graph $G = (V, E, S)$, where V is the node set of the new pages, S indicates the similarity matrix between the new-page notes. E is the edge set. Assign a weight to the edge between two nodes according to the similarity between them. After some pretreatment for the retrieved new-pages, such as stripping tags, Chinese words segmentation and the stop word removal, we entirely consider the occurrence frequency of different words in all the texts and the resolution capabilities of the word to the different texts and then use the TF-IDF based on word frequency statistics to calculate the word feature weights of the two new-pages nodes. The cosine of the vector space angle of the documents is employed for defining the similarity between the two new-page nodes, based on which we can get the initial similarity matrix S.

Suppose there are two arbitrary new-page nodes $\forall x_i, x_k \in V$, TF-IDF as follows:

$$W_{t,x} = TF_{t,x} \times IDF_{t,x} \quad (1) \qquad TF_{t,x} = N / M \quad (2) \qquad IDF_{t,x} = \log(x / x_N) \quad (3)$$

Where $W_{t,x}$ is the weight of the feature item t in the document x; $TF_{t,x}$ represents the occurrence frequency of t in the document x. $IDF_{t,x}$ is known as the document frequency of features to reflect the distribution of feature item t in the whole document set and the distinction ability of this feature item to a certain extent, The following we use the angle cosine of the two document vectors to define the initial similarity S between two new-page nodes.

$$S_{ik} = Sim(x_i, x_k) = \cos\theta = \frac{\sum_{t=1}^{n} W_{t,x_1} \times W_{t,x_2}}{\sqrt{\left(\sum_{t=1}^{n} W_{t,x_1}^2\right)\left(\sum_{t=1}^{n} W_{t,x_2}^2\right)}} \qquad (4)$$

Where $W_{t,x1}, W_{t,x_2}$ is respectively devotes the weight of the feature item t in the new-page documents x_1, x_2 and further obtain the initial similarity matrix S .

3.2 Correcting the initial similarity matrix based on the relation among news elements

Selecting the structural characteristics of new documents .
The html code of news documents was analyzed to find news document structural features. News report not only gives the text describing the event, but also the report title, entity name, keywords, links, classes and other factors. The relation among these factors plays a pivotal role in the initial adjustment of the initial similarity of news pages. After the analysis, 5 kinds of news documents structural features were selected:

Table 1. Structural features of new documents

Serial number	Structural characteristics
1	title
2	entity
3	keywords
4	links
5	category

Annotation was conducted using CRF-TF tool developed in the laboratory, and the recognition model was trained using the above tool. The structured features of the test corpus were identified and extracted through the recognition model.

Analysis of relation among document elements and their characteristic value.

Table 2. Relationship among document elements and their characteristic value

Serial number	Elements-associated relationship	Feature type	Eigenvalue f_m	Feature weight α_m
1	Title association	Boolean	0,1	α_1
2	Entity co-occurrence	Boolean	0,1	α_2
3	Keywords co-occurrence or high correlation	Boolean	0,1	α_3
4	Links pointed to each other	Boolean	0,1	α_4
5	Whether belongs to the same category	Boolean	0,1	α_5

In order to avoid the frequent appearance of 0, TF-IDF to calculate feature weight was not used. Instead, the method of calculating the semantic extension of words was introduced which was based on Tongyici Cilin in semantic [14] put forward by Liu Duanyang and Wang Liangfang [13].

3.3 Page-associated feature constraints and similarity correction

Add the weights to the different relations through training the relation characteristic among the elements of the page as shown in table 1， assume two arbitrary new-page nodes , define a adjustment matrix as follows:

$$S'_{ik} = \sum_{m=1}^{4} \alpha_m f_m + \alpha_5 f_5 \tag{5}$$

Where, α_m is the elements-associated feature weight obtained by training, as the two page is belong to the same category is a important factor to determine whether two pages should be gathered to a class. Therefore set $\alpha_5 = 0.5$, $\sum_{m=1}^{4} \alpha_m = 0.5$ and f_m denotes the elements-associated eigenvalue mentioned in table 2.

Wagstaff etc [15] had introduced two kinds of pairwise constraints the first time, say must-link and cannot-link constraints to facilitate clustering search.

According to the rules of heuristic judgment and integrating the relation among documents elements with the adjustment matrix S'_{ik} .Define the elements-associated feature constraints (including the “must-link” and “cannot-link”) shown in formula (6) as the supervision constraint information, and then do the correction of the initial similarity matrix S based on the “must-link” and “cannot-link” constraints to obtain the final similarity matrix S''_{ik} .

$$\begin{cases} S''_{ik} = S''_{ki} = 1, \left(S'_{ik} \geq 0.5\right) \Leftrightarrow \left(x_i, x_k\right) \in must-link \\ S''_{ik} = S''_{ki} = 0, \left(S'_{ik} = 0\right) \Leftrightarrow \left(x_i, x_k\right) \in cannot-link \\ S''_{ik} = S_{ik} \left(0 < S'_{ik} < 0.5\right) \end{cases} \tag{6}$$

For formula (6), it is represented that it has the same category at least or that the previous four eigenvalues are all 1 in the two new-pages together are a “must-link” constraint. The same reason is that all elements-associated eigenvalues are all 0 when $S'_{ik} = 0$.

4 Experiments and Analysis

4.1 Experiment Data Preparation

In this paper, different templates were formulated using HtmlUnit to gather the experimental data from different sites. The gathered data were stored in MongoDB database. The use of HtmlUnit to gather news data can obtain more accurate structured data of the news webpage, such as titles, URL links, keywords, category, etc.

In the experiment, the test set was selected 4 times. Each time, 10 news texts in each of 10 fields (politic, economy, culture, etc.) were extracted from the database and these 100 texts were used as the test set. 1000 high-frequency words were chosen based on word frequency statistics method and their feature weight were calculated using TF-IDF method. Then 1000 dimensional feature space was constructed. In the experiment, the number of the feature constraints of the relation among the elements of the page ranged from 0 to 400. The feature constraint of the relation among the page elements was randomly generated from the training data set. The randomly generated constraints were used to guide the analysis of all the data with the clustering algorithm. At last, the clustering result of the test data sets was evaluated according to the clustering evaluation index.

4.2 Experimental Evaluation

Here the F-value will be employed. The F-value is defined as below:

$$\Pr e = \frac{T_p}{T_p + F_p}, \operatorname{Re} c = \frac{T_p}{T_p + F_n}, F = \frac{2 \Pr e \times \operatorname{Re} c}{\Pr e + \operatorname{Re} c} \tag{7}$$

Where T_p is the number of new-page documents that the two documents together in one cluster are classified correctly, F_p represents the number of new-page documents that the two should not be placed in one cluster are divided into one falsely, F_n is the number of new-page documents that the two should not be separated are parted wrongly.

4.3 Experimental Design and Analysis

Experiment 1 was performed to evaluate the influence of different number of relation characteristic constraints on clustering. For any given number of constraints, the experiment was done 30 times. The average output results represented the performance of the Affinity Propagation model integrating relation among page elements and with a fixed number of constraints to cluster some test set. Each test set had 200 texts, of which 50 were used as the test set, and the rest were used as the training set. So figure

1 reveals the changes of the F-value indexes on the four test sets with the different numbers of associated constraints.

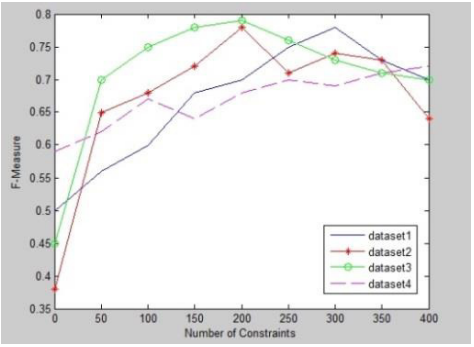

Fig. 2. The impact of different numbers of associated constraints on the F-values indexes of the four test sets

It can be seen from figure 1 that the F-value index obtained by our approach show upward trend as a whole with the increase in the number of constraints, when the number of constraints is between 150 and 300, the corresponding F-value reached the maximum and the number of clusters obtained at this time is optimal.

Experiment 2 is the clustering performance comparison of the AP clustering method integrating the relation among page elements of news (T_1) and without constraint information (T_2). There are results of experiment 2 given in table 3.

Table 3. Results of clustering of news topics with different algorithms

datasets	T_1			T_2		
	P（%）	R（%）	F（%）	P（%）	R（%）	F（%）
TestSet1	67.67	50.75	58.00	83.00	92.00	87.26
Test Set 2	76.00	84.00	79.8	85.64	82.35	83.97
Test Set 3	71.64	75.00	73.28	83.65	80.01	81.78
Test Set 4	55.56	73.45	63.26	86.17	83.57	84.86

Seen from Table 3, the P, R and F-value obtained by using the AP, which integrating the relation among page elements are significantly improved than those without constraint information. Therefore, it can be known that adding constraint information can largely enhanced the clustering performance of the entire data set.

5 Conclusions

In this paper, an Affinity Propagation clustering method integrating the relation among page elements of news is proposed. By this method, the characteristic weight

obtained using the relation among page elements of news was made full use of to modify the initial page similarity matrix, through which large change took place of data similarity matrix. The contrast experiments demonstrate that the use of the relation among page elements of news as priori constraint is effective and it can play a guiding role in generating the Affinity Propagation clustering algorithm. The further work is to take the initiative to provide affinity propagation clustering algorithm rich constraint information, so that the news topic clustering effect can be improved.

Acknowledgments. This work was supported by National Natural Science Foundation of China, Contract No.61472168.

References

1. Makkonen, J., Ahonen-Myka, H.: Simple Semantics in Topic Detection and Tracking. Information Retrieval 7(3-4), 347-368(2004)
2. Wei, C., Lee, Y., Chiang, Y.: Discovering Event Episodes from News Corpora: A Temporal-based Approach. In: Proceedings of ACM ICEC2009, Taipei, Taiwan, pp.72-80(2009)
3. Ahmed, A., Ho, Q., Eisenstein, J.: Unified analysis of streaming news. In: Proceedings of the 20th international conference on World Wide Web. ACM, pp.267-276(2011)
4. Newman, D., Chemudugunta, C., Smyth, P.: Statistical entity-topic models. In Kdd, New York, NY,USA, pp.680-686(2006)
5. Ogilvie, P.: Extracting and Using Relationships Found in Text for Topic Tracking(2000)
6. Carthy, J.: Lexical Chains versus Keywords for Topic Tracking. In: Proceedings of the 5th International Conference on Intelligent Text Processing and Computational Linguistics (2004)
7. Liu, M., Wang, X., Liu, Y.: Research of Key-Phrase Extraction Based on Lexical Chain. Chinese Journal of Computers(33),1246 – 1255(2010)
8. Frey, B., Dueck, D.: Clustering by passing messages between data points. Science315(5814),972-976(2007)
9. Frey, B., Dueck, D.: Response to Comment on "Clustering by passing messages between datapoints". Science319(1151268),726(2008)
10. Xiao, Y., Yu, J.: Semi-Supervised Clustering Based on Affinity Propagation Algorithm. Journal of Software19(11),2803-2813(2008)
11. Guan, R., Shi, X.: Text Clustering with Seeds Affinity Propagation. IEEE Trans, on Knowledge and Data Engineering23(4),627-637(2011)
12. Luo, W., Liu, Q., Cheng, X.: Development and Analysis of Technology of Topic Detection and Tracking . Association for Compu-tational Linguistics(2004)
13. Liu, D., Wang, L.: Extraction Algorithm Based on Semantic Expansion Integrated with Lexical chain. Computer Science40(12),264-269(2013)
14. Tian, J., Zhao, W.: Words Similarity Algorithm Based on Tongyici Cilin in Semantic Web Adaptive Learning System. Journal of Jilin University28(6), (2010)
15. Wagstaff, K., Cardie, C.: Clustering with instance-level constraints. In: Langley P, Ed. Proc. of the 17th Int'l Conf. on Machine Learning. Morgan Kaufmann Publishers, pp.1103−1110(2000)