A Chinese Word Clustering Method Using Latent Dirichlet Allocation and K-means

Qiu Lin, Xu Jungang
School of Computer and Control Engineering
University of Chinese Academy of Sciences
Beijing, China
qiulin10@mails.ucas.ac.cn, xujg@ucas.ac.cn

***Abstract*—Word clustering is a popular research issue in the field of natural language processing. In this paper, Latent Dirichlet Allocation algorithm is used to extract the topics from nouns in the text, and the highest probability noun of each topic is selected as the centroids of the k-means algorithm. Experimental results show that this method can get better effects than the graph-based word clustering algorithms using a web search engine.**

Keywords-word clustering; latent dirichlet allocation; k-means; word similarity

I. INTRODUCTION

In the field of natural language processing, word clustering is a widely studied subject [1]. And it is important for automatic thesaurus construction, text classification, and word sense disambiguation [2]. A typical word clustering task is described as follows: given a set of words (nouns), cluster words into groups so that the similar words are in the same group (cluster).

Currently, there are a lot of researches on word clustering. The text-oriented word clustering has been widely studied and used in document clustering [3], document classification [4], and large-scale class-based language modeling in machine translation [5].

In this paper, a Chinese word clustering method that combines Latent Dirichlet Allocation (LDA) [6] and k-means algorithms. K-means [7] is an unsupervised method for clustering, which cannot produce unique clustering result because the initial clusters are chosen randomly. In order to solve this problem, we extract the topics from nouns of each sentence through LDA algorithm, and then choose the highest probability noun of each topic as the centroids of k-means algorithm, and then use k-means algorithm to cluster all the words in the text.

The remainder of this paper is organized as follows. Section 2 is the related work. Section 3 introduces the k-means algorithm, the LDA algorithm and Chinese word similarity calculation method. Section 4 describes the Chinese word clustering method using LDA and k-means algorithms. Section 5 is experimental results. Finally, section 6 is the conclusions.

II. RELATED WORK

Word clustering is an important branch of natural language processing techniques, and which causes wide concerns.

The basic idea of word clustering method based on corpus is grammatical features that extracted from the target word in the context, and the words which have similar grammatical features can be in the same cluster [8]. Farhat et al [9] proposed a formal representation of the target word in the context, in which the target word is represented as a binary random variable, the Kullback-Leibler is used to calculate the distance between two words.

The word clustering based on semantic features usually rely on a semantic knowledge base. These semantic resources are directly coded by language experts, so this method belongs to the rule-based method [8]. Hang Li [10] proposed a method that combining the Minimum Description Length principle and the disambiguation method to derive a disambiguation method that makes use of both automatically constructed thesauruses and hand-made thesaurus.

Pragmatic feature refers to the feature of the word that shows in the specific application. Yutaka Matsuo [11] proposed an unsupervised algorithm for word clustering based on a word similarity measure by web counts.

III. THE EXISTING METHODS OF CHINESE WORD CLUSTERING AND SIMILARITY CALCUTION

A. K-means

The k-means [7] algorithm finds locally optimal solutions for minimizing the sum of distance between each data point and its nearest cluster centroid. Note that k-means is defined over numeric (continuous-valued) data since it requires the ability to compute the mean. On the other hand, words cannot be represented with numeric data, so the distance between the words cannot be measured by Euclidean distance. In this paper, we use the results of LDA algorithm to get the centroids of k-means, and use the method of Chinese word similarity calculation to measure the distance between the words, Chinese word similarity calculation method will be discussed in the following subsection.

B. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [6] is a generative probabilistic model for processing collections of discrete data such as text corpus, which has quickly become one of the most popular probabilistic text modeling techniques. LDA uses the bag of words model, which considers each document as a word frequency vector. The graphical model of LDA is shown in Fig. 1.

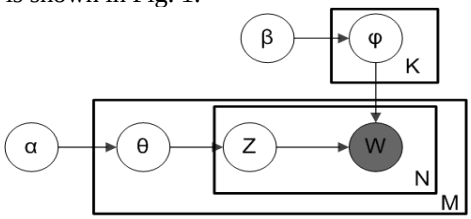

Figure 1. The graphical model of LDA

LDA could be described as follows [12]:

- Word: a basic unit defined to be an item from a vocabulary of size *W*.
- Document: a sequence of *N* words denoted by $d=(w_1,..., w_n)$, where w_n is the n^{th} word in the sequence.
- Corpus: a collection of *M* documents denoted by $D=\{d_1, ..., d_m\}$.

Given *D* documents is expressed over *W* unique words and *T* topics, LDA outputs the document-topic distribution θ and topic-word distribution ϕ. This distribution can be obtained by a probabilistic argument or by cancellation of terms in (1):

$$p(z_i = j \mid z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{\sum_{w'}^{W} n_{-i,j}^{(w')} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{\sum_{j}^{T} n_{-i,j}^{(d_i)} + T\alpha} \quad (1)$$

Where $\sum_{w'}^{W} n_{-i,j}^{(w')}$ is a count that does not include the current assignment of Z_j. The first ratio denotes the probability of w_i under topic *j*, and the second ratio denotes the probability of topic *j* in document d_i. Critically, these counts are the only necessary information for computing the full conditional distribution, which allow the algorithm to be implemented efficiently by caching the relatively small set of nonzero counts. After several iterations for all the words in all documents, the distribution θ and distribution ϕ are finally estimated using (2) and (3).

$$\phi_j^{(w_i)} = \frac{n_j^{(w_i)} + \beta}{\sum_{w'}^{W} n_j^{(w')} + W\beta} \quad (2)$$

$$\theta_j^{(d_i)} = \frac{n_j^{(d_i)} + \alpha}{\sum_{j}^{T} n_j^{(d_i)} + T\alpha} \quad (3)$$

C. Similarity Calculation Based on HowNet

In order to calculate the distance between words in k-means algorithm, we use the word similarity instead of the Euclidean distance. The higher the similarity between two words is, the closer the two words are.

HowNet [13] is an on-line common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese and their English equivalents. The concept can be divided into a number of sememes, which are the smallest basic semantic units that cannot be reduced further.

The calculation formula of the similarity between two sememes [14] is described in (4).

$$Sim(p_1, p_2) = \frac{2 \times Spd(p_1, p_2)}{Dsd(p_1, p_2) + 2 \times Spd(p_1, p_2)} \quad (4)$$

Where $Spd(p_1, p_2)$ refers to the path length of the parent node which the two sememes p_1 and p_2 share in the sememes hierarchy system. $Dsd(p_1, p_2)$ describes the length of the shortest path where p_1 and p_2 move upward gradually along parent nodes until they reach the second shared node. And the concept similarity formula [14] between two concepts C_1 and C_2 is described in (5).

$$Sim(C_1, C_2) = \beta_1 Sim_1(C_1, C_2) + \sum_{i=2}^{4} \beta_1\beta_i Sim_i(C_1, C_2) \quad (5)$$

Where $\beta(1 \le i \le 4)$ denotes the similarity of the first basic sememes, the other basic sememes, the relational sememes and the signal semems respectively, which are the adjustment parameters limited by

$$\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1, \beta_1 \ge \beta_2 \ge \beta_3 \ge \beta_4 > 0.$$

IV. One Improved Chinese Word Clustering Method

A. The definition of the Algorithm

As everyone knows, the traditional k-means clustering algorithm is a supervised learning algorithm, and the different choice of initial centers will bring great impact on the results. So, in this paper, a new method that combines the Latent Dirichlet Allocation (LDA) algorithm and k-means clustering algorithm is proposed. The main steps of this algorithm are listed in Fig. 2.

Algorithm: Improved Chinese word clustering method

1: **for** each d_i in document set **do**
2: SS←sentence segment
3: WS←word segments for each sentence
4: NV←vectors expressed by nouns in each sentence
5: NL←noun list including the different nouns in the text
6: SeedSet←extract the topics by LDA algorithm ($NV_{i1}, NV_{i2}, \ldots, NV_{in}$)
7: Cluster←k-means (centroids initialized by SeedSet, NL_i)
8: end **for**

Figure 2. The main steps of the algorithm

For each document in the document set, some pre-process work need to be done firstly (Line 2 - Line 5). Secondly, and then some topics are extracted by LDA algorithm and the highest probability nouns of each topic are chosen as the centroids of k-means (Line 6). At last, k-

means algorithm is used to cluster all words in the text (Line 7).

B. Baseline Algorithms

Yutaka Matsuo [11] proposed an unsupervised clustering algorithm for word clustering based on a word similarity measured by web counts. Each pair of words is queried to a search engine, which can result in a co-occurrence matrix. By calculating the similarity of words, a word co-occurrence graph is created. Girvan and Newman (GN) [11] algorithm is a new king of graph clustering algorithm. GN algorithm emphasizes betweenness of an edge and identifies densely connected sub-graphs.

The method to measure semantic similarity between words or entities using Web search engines has been introduced by many papers. Web search engines provide an efficient interface to this vast information. And the most common methods of semantic similarity between words based on Web search engines are PMI, Jaccard, Overlap and Dice, which are defined as in [15].

C. Improved Chinese Word Clustering Method

LDA can be used to convert word dimension of document into topic dimension. So LDA could be re-defined as follows:

- Word: a basic unit defined to be an item from a vocabulary of size *W*.
- Sentence: a sequence of *N* words denoted by $s=(w_1,..., w_n)$, where w_n is the n^{th} noun in the sentence.
- Document: a collection of *M* sentences denoted by $D=\{s_1, ..., s_m\}$.

LDA includes a process of generating the topics in each document, which greatly reduces the number of parameters to be learned and provides a clearly-defined probability for arbitrary documents. In collapsed Gibbs sampling, only z_{ij} is sampled, and the sampling is done conditioned on α, β and the topic assignments of other words z_{ij}.

As the output of LDA algorithm, *k* topics for the text can be got, and then the highest probability nouns of each topic are chosen as the initial cluster centroids of k-means algorithm.

V. EXPERIMENTATION

A. Data Set

We select 516 People's Daily editorials from year 2008 to 2010 as the experimental date set. These 516 editorials are manually divided into five categories: politics (183 editorials), economy (114 editorials), culture (56 editorials), people's livelihood (91 editorials), science and technology (72 editorials).

B. Experimental Design

We labeled 10% of the data sets, which includes 18 editorials of politics, 11 editorials of economy, 5 editorials of culture, 9 editorials of people's livelihood and 7 editorials of science and technology. *k* (topic number) is chosen as valued from 5 to 15, and the precision for different *k* can be obtained, which is shown in Fig. 3. We find that the best precision can reach 61.2% when k is 8. α and β are set as 0.2 and 0.1 respectively.

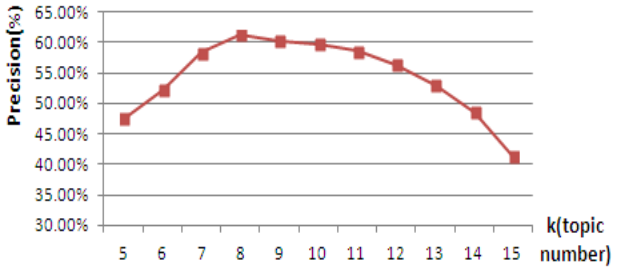

Figure 3. The precision for different value k

HowNet is used to calculate the average similarity of every cluster. And the average similarity of clusters is defined in (6), where *n* stands for the number of clusters.

$$AveSim = \frac{\sum_{i=1}^{n} average\ similarity\ of\ cluster\ i}{n} \quad (6)$$

And the average similarity calculation equation of cluster *i* is defined in (7), where *m* stands for the number of words.

$$ASC_i = \frac{\sum_{i=1}^{m}\sum_{j=i+1}^{m} similarity\ between\ word_i\ and\ word_j}{m*(m-1)/2} \quad (7)$$

In order to calculate the similarity between clusters, we choose the centriods that represent every cluster in Improved K-means using LDA (IKL) and k-means. And for GN algorithms, we choose a word (we also call the word as centroid) from every cluster and guarantee that the average similarity between the centroid and the other words in the same cluster is the highest. The calculation of the similarity between clusters is simplified to the calculation of the similarity between centroids. And the average similarity between clusters is defined in (8), where *n* stands for the number of clusters.

$$ASBC = \frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n} similarity\ between\ centroid_i\ and\ centroid_j}{n*(n-1)/2} \quad (8)$$

C. Experimental Results

Fig. 4 shows the average similarity of IKL, K-means and four GN algorithms used in a Web search engine.

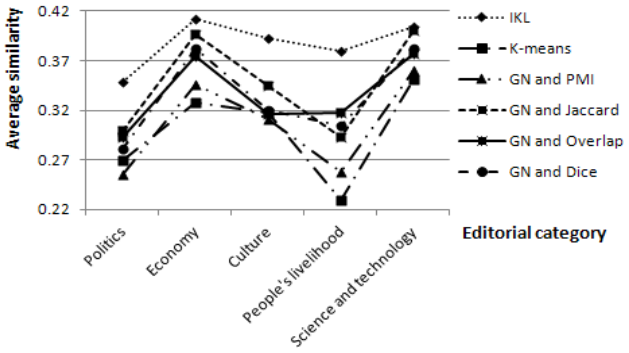

Figure 4. The average similarity of six algorithms

From Fig. 4, we can see that the average similarity calculated by IKL is higher than the other algorithms, which means that using IKL can get better clustering effect within one cluster than the other algorithms.

The result of the average similarity between clusters (centroids) is shown in Fig. 5.

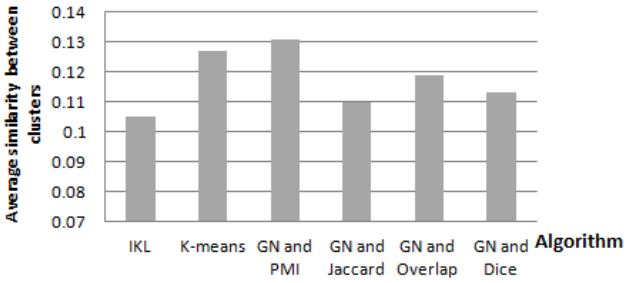

Figure 5. The average similarity between clusters

From Fig. 5, we can see that the average similarity between clusters of IKL is lower than the other algorithms, which means that using IKL can get better discrimination effect between the clusters than the other algorithms.

Also, the time consumption of GN algorithm using Web search engine is enormous. We set the nouns after pre-processing on a server[1] to obtain the web counts for each word and the web counts for pairs of words using a search engine. It takes about ten days to get all web counts. And on the other hand, IKL takes only about 4 minutes to get the result.

VI. CONCLUSIONS

In this paper, one Chinese word clustering method is proposed. LDA algorithm is used to choose the initial centroids of k-means, and the result of the centroids of clusters obtained by LDA algorithm is relatively close to the final result of k-means clustering, thus it can improve the average similarity of the final k-means clustering. Based on the People's Daily editorial data set, experimental results show that the average similarity of the proposed algorithm is better than k-means algorithm and four graph-based word clustering algorithms.

[1] The server's configuration: 2.03 GHz E5606 Intel(R) Xenon(R) processor, 4GB DDR3 RAM and 500 GB SATA Hard Disk.

ACKNOWLEDGMENTS

This work is supported in part by the National Key Technology R&D Program of China under Contract No. 2012BAH23B03.

[1] M.S. Sun, Z.P. Zuo, and B.K.Tsou, "Part-of-speech identification for unknown Chinese words based on K-Nearest-Neighbors strategy," Chinese Journal of Computers, vol. 23, Feb. 2000, pp. 166-170.

[2] A. Khaled, M. David, and S. Nathan, "An effcient k-means clustering algorithm: analysis and implementation," IEEE Transactions on Pattern Analysis and Machine Intellgence, vol. 24, Jul. 2002, pp. 43-48.

[3] S. Noam and T. Naftali, "Document clustering using word clusters via the information bottleneck method," Proceeding of the 23rd Annual ACM SIGIR conference (Athens, Greece, July 24-28, 2000), pp. 208-215.

[4] H. Li and A. Naoki, "Word clustering and disambiguation based on co-occurrence data," Proceedings of the 17th International Conference on Computational Linguistics (Montreal, Quebec, Canada, August 10-14, 1998), pp. 749-755.

[5] Y. Wen, C.F. Yuan, and C.N. Huang, "Clustering of Chinese adjectives-Nouns based on compositional pairs," Journal of Chinese Information Processing, vol. 14, Jun. 2000, pp. 45-50.

[6] M.B. David, Y.N. Andrew, and I.J. Michael, "Latent dirichlet allocation," Journal of Machine Learning, Mar. 2003, pp. 993-1022.

[7] Z. Zhang, J.X. Zhang, and H.F. Xue, "Improved k-means clustering algorithm," Proceedings of 2008 International Congress on Image and Signal Processing (Sanya, Hainan, China, May 27-30, 2008), pp. 169-172.

[8] H.E. Guo, L.J Zhu,.and S Xu, "A survey on word clustering technique," Digital Library Forum, May. 2010, pp. 14-18.

[9] A. Farhat, J.F. Isabelle, and D. O'Shaughnessy, "Clustering words for statistical language models based on contextual word similarity," Proceedings of the Fourth International Congerence on Spoken Languge Processing (Philadelphia, USA, October 03-06, 1996), pp. 180-183.

[10] H. Li and A. Naoki, "Word clustering and disambiguation based on co-occurrence data," Proceedings of the 17th International Conference on Computational Linguistics (Montreal, Quebec, Canada, August 10-14, 1998), pp. 749-755.

[11] M. Yutaka and S. Takeshi, "Graph-based word clustering using a web search engine," Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (Sydney, Australia, July 22-23, 2006), pp. 542-550.

[12] L.G. Thomas and S. Mark, "Finding scientific topics," Proceedings of the National Academy of Sciences, Apr. 2004, pp. 5228-5235.

[13] Z.D. Dong and Q. Dong, "Hownet," http://www.keenage.com/, 2011-04-10.

[14] Q Liu and S.J. Li, "Word similarity computing based on Hownet," Computational Linguistics and Chinese Language Processing, vol. 7, Aug. 2002, pp. 59-76.

[15] B. Danushka, M. Yutaka, and I. Mitsuru, "Measuring semantic similarity between words using web search engines," Proceedings of the 16th International Conference on World Wide Web (Banff, Alberta, Canada, May 08-12, 2007), pp. 757-766.