J Inf Process Syst, Vol.14, No.6, pp.1438~1444, December 2018
https://doi.org/10.3745/JIPS.04.0097

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Semantic-Based K-Means Clustering for Microblogs Exploiting Folksonomy

Jee-Uk Heu*

Abstract

Recently, with the development of Internet technologies and propagation of smart devices, use of microblogs such as Facebook, Twitter, and Instagram has been rapidly increasing. Many users check for new information on microblogs because the content on their timelines is continually updating. Therefore, clustering algorithms are necessary to arrange the content of microblogs by grouping them for a user who wants to get the newest information. However, microblogs have word limits, and it has there is not enough information to analyze for content clustering. In this paper, we propose a semantic-based K-means clustering algorithm that not only measures the similarity between the data represented as a vector space model, but also measures the semantic similarity between the data by exploiting the TagCluster for clustering. Through the experimental results on the RepLab2013 Twitter dataset, we show the effectiveness of the semantic-based K-means clustering algorithm.

Keywords

Cluster, K-means, Microblog, Semantic, TagCluster

1. Introduction

Recently, with the development of Internet technologies and propagation of smart devices, users are able to easily access various information. In the Web 2.0 environment, users can create and share various media such as images and videos through a social media service, and the quantity of those content has been increasing daily. In particular, users have been exposed to a flood of information because of the rapid growth of microblog services such as Twitter, Facebook, Instagram, and Tumblr. Twitter is one of the most famous microblogs, where people express their thoughts or opinions using a short message with the limitation of 140 characters, referred to as a tweet. The user can freely express their thoughts or feelings about any topic by creating a tweet and constructing their own social network. Twitter is used by various celebrities, such as singers, athletes, actors, and politicians, as well as ordinary people. According to research, on December 31, 2013, there were already more than 240 million active users per month, spanning nearly every country in the world. Approximately 500 million tweets are created every day, and the number of created tweets is continually increasing.

It is not easy to find meaningful information from a large group of tweets after users are offline for a

Manuscript received July 17, 2018; accepted August 25, 2018.
Corresponding Author: Jee-Uk Heu (jeeukheu@gmail.com)
* Dept. of Computer Science and Engineering, Hanyang University, Seoul, Korea (jeeukheu@gmail.com)

while because newly posted content is continuously generated on their timelines. Therefore, microblog users have to check all of the newly posted content to find the most interesting information. To address this problem, many cluster algorithm techniques have been proposed for arranging microblog contents. These technologies have become especially important with the advent of smart devices and wearable devices. However, microblog contents generate less data than general documents for clustering because the limited information. Therefore, it is essential to overcome this by exploiting an external knowledge base from the collective intelligence of folksonomy when analyzing and detecting microblogs.

In this paper, we propose a semantic-based K-means clustering algorithm that not only measures the similarity between the tweets represented by a vector space model, but also measures the semantic similarity between the tweets using TagCluster of Flickr for clustering a large number of tweets. The rest of this paper is organized as follows: Section 2 introduces related work on clustering techniques for tweets. Section 3 presents our proposed system. Experimental results are presented in Section 4. Finally, Section 5 concludes the paper by discussing the direction for future work.

2. Related Work

To cluster microblogs, document clustering algorithms have been used in many cases. Generally, in such algorithms, the words in a document are expressed by a vector space, and the similarity of documents is measured by calculating the distance between their vectors. However, a tweet does not include sufficient information to analyze it due to its limitation of 140 characters. Therefore, previous clustering algorithms handled multimedia, such as documents, images, and videos, which provide a lot of data for analysis. However, these algorithms are not sufficient for microblog data, which lack sufficient contextual information. To overcome the above limitation, recent studies [1-7] have considered the features of microblogs for analyzing and clustering tweets.

Wang et al. [5] proposed the Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model (GSDMM) with latent feature LDA (LFLDA) as a two-stage hierarchical topic modeling system for the twitter topic model. As the first stage, the authors used GSDMM for topic modeling. Then, they assigned every tweet to its corresponding cluster and aggregated each cluster to form a virtual document. Finally, in the second stage, they apply the LFLDA topic modeling algorithm for clustering the generated virtual documents. Although the LDA-based model performs well, the probability method employed requires complex computation and long processing time.

Overcoming the lack of information in a short text, Zheng et al. [6] proposed a corpus-based enrichment approach for short text clustering. The authors inferred hypothetical topic proportion vectors from the short texts using the LDA algorithm and generated virtual words from these vectors for enrichment. However, the proposed method did not consider the semantic similarity between words when the collected tweets are clustered.

Dhuria et al. [7] proposed an integrated natural language processing (NLP) and ontology-based clustering TVC algorithm that generates semantically meaningful concepts stored in cluster terms. However, these approaches require a high processing cost and a long time during clustering because external knowledge such as ontology is too extensive to measure the similarity.

To overcome these problems, we propose a semantic-based K-means clustering algorithm. Our approach measures semantic similarity between words in tweets by exploiting TagCluster and the similarity of tweets by the vector space model for clustering a large volume of microblogs.

3. Semantic-Based K-Means Clustering

Fig. 1 shows the architecture of our proposed system, which consists of two modules: pre-processing and tweet clustering. The overall workflow of the proposed system can be described as follows. First, the pre-processing module extracts useful data from the input tweets and refines it.

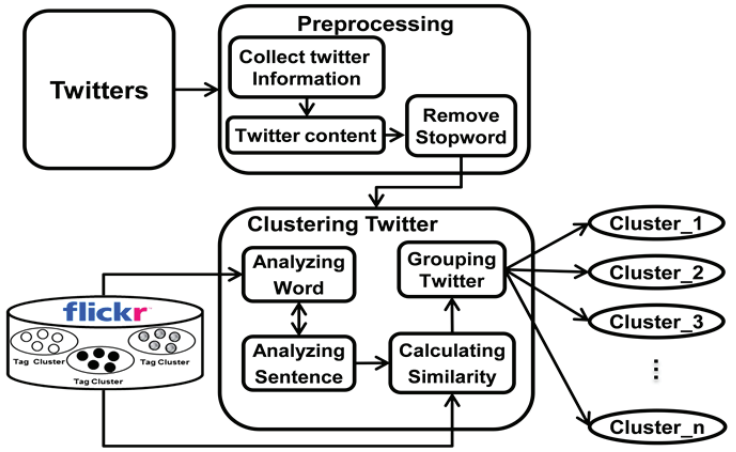

Fig. 1. System architecture.

The extracted texts are clustered by the tweet clustering module that calculates the similarity of tweets using TagCluster from Flickr.

The clustering module analyzes tweet data, which are received from the pre-processing module for calculating the similarity of each tweet and clustering the given tweets. We propose a semantic-based K-means clustering algorithm that is a modified version of the K-means algorithm to cluster tweets. The K-means algorithm is one of the most popular clustering algorithms, which uses a vector space model of input data and the required number of k clusters to iteratively partition the data into k clusters based on a distance function such as cosine similarity. However, the K-means algorithm does not consider semantics when calculating similarity, instead using cosine similarity. The algorithm views the given tweet data as points in the vector space and measures how closely they are related. Furthermore, tweets can only be 140 characters and so are not easy to analyze for clustering. This leads to low accuracy when clustering a large amount of tweets. To overcome this drawback, we exploit external knowledge for more accurately analyzing tweet semantics for determination of similarity.

To obtain external knowledge, we exploit TagCluster from Flickr, a folksonomy system that consists of collective intelligence. Flickr is a popular collaborative tagging application for pictures, and it provides TagCluster to group related tags [6]. Moreover, TagCluster consists of many proper nouns and newly coined words such as the names of people and products. These properties make it easy to analyze the semantics in each tweet. Here, $T = \{t_1, t_2, t_3, \dots, t_n\}$ denotes a set of given tweets, and the information of each tweet is represented as follows:

$$t_i = \{w_1, w_2, w_3, \dots, w_n\}, \tag{1}$$

where t_i is the i-th tweet of a given tweet set T, and w_i is the i-th word in each tweet. Then, the expanded information of each tweet found by TagCluster from Flickr can be defined as

$$t_i' = t_i \cup TC_{t_i}, \tag{2}$$

where t_i' is an updated tweet of t_i by TagCluster, and TC_{t_i} is the TagCluster of t_i. t_i' consists of original words of t_i and new words in TC_{t_i}. As a result, we can calculate the semantic similarity of the expanded tweets using the Jaccard Similarity:

$$SemanticSimilarity(t_i', t_j') = \frac{|t_i' \cap t_j'|}{|t_i' \cup t_j'|}. \tag{3}$$

Jaccard Similarity is used for comparing the similarity and diversity of a sample set. It is one of the simplest methods and requires low cost for calculating the similarity of given sets. Finally, similarity of tweets can be calculated by summing the cosine similarity and the semantic similarity as follows:

$$Similarity(t_i, t_j) = CosineSimilarity(t_i, t_j) + SemanticSimilarity(t_i', t_j'). \tag{4}$$

The semantic-based K-means clustering algorithm considers the semantics as well as the occurrence and frequency of words in each tweet for measuring the similarity between tweets t_i and t_j for K-means clustering.

4. Performance Analysis

We used the RepLab2013 Twitter dataset to empirically evaluate our proposed methods consisting of two main modules for semantic-based K-means clustering. The RepLab2013 dataset was collected from June 2012 to December 2012 and consists of four domains: automotive, banking, university, and music. The dataset was manually labelled with the topic, topic priority (important or non-important), and polarity (positive or negative) of each tweet by annotators who were trained and guided by experts. The dataset provides a training set (45,671 tweets) and a test set (105,099 tweets). In these datasets, we selected the music domain and performed experiments to evaluate the performance of the semantic-based K-means clustering algorithm. To cluster the large amount of tweets, we set k at 20 for clustering the 20 musicians from the RepLab2013 dataset.

The detailed results of semantic K-means clustering on the Raplab2013 dataset are presented in Table 1, showing precision, recall, and F-measure results of each semantic K-means cluster. Our proposed clustering method achieves generally good performances except for the 18th cluster. We will explain later why this cluster returned a low result. The precision results of the 5th and 10th clusters were 100%. In terms of recall, the 4th, 5th, and 14th clusters showed 100%. As the results, the tweets of 20 musicians on the Replab2013 dataset were properly distributed with 82.1% precision, 89.07% recall, and 87.36% F-measure.

To show the superiority of our proposed clustering algorithm, we compared results to the original K-means clustering algorithm on the RepLab2013 dataset. Fig. 2 shows the comparison of the clustering results of the semantic-based K-means and the original K-means clustering algorithms. In most cases, the proposed method outperformed the original.

Table 1. Results of semantic-based K-means clustering (K=20)

Cluster no#	Precision (%)	Recall (%)	F-measure (%)
1	89	97.05	92.85
2	94	75.9	83.99
3	84	94.1	88.76
4	98	100	98.99
5	100	100	100
6	84	96	89.60
7	90	90	90
8	61	95.6	74.48
9	92	98.2	95
10	100	95.8	97.85
11	97	94.7	95.84
12	75	90	81.82
13	84	91.1	87.41
14	97	100	98.48
15	79	83.9	81.38
16	87	97.6	92
17	75	77	75.99
18	37	37.3	37.15
19	95	94.7	94.85
20	83	100	90.71

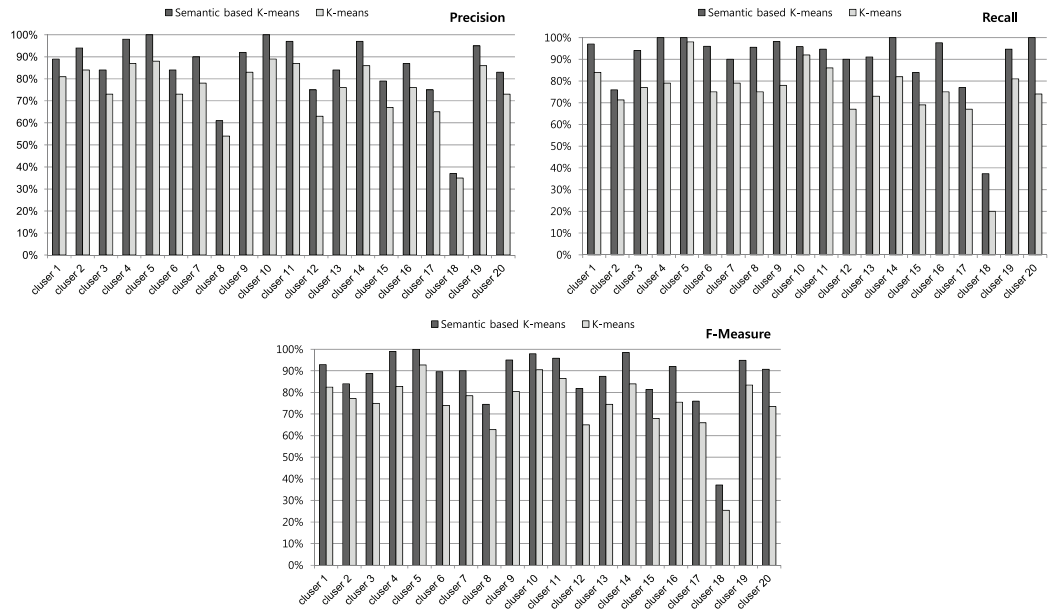

Fig. 2. Comparison of semantic-based and original K-means results.

The precision and recall of our proposed clustering algorithm were improved from 7.52% to 14.47% (10.73% on average) compared with that of the original K-means. In terms of recall, the proposed method showed better performance by an average of 13.87%. The 16th cluster showed outstanding results for both precision and recall. Finally, the precision, recall, and F-measure of the proposed clustering algorithm were higher than those of the original algorithm. This means that the proposed

semantic-based K-means algorithm can more effectively perform clustering for the enormous number of tweets.

For a detailed description, Fig. 3 shows each cluster result as determined by our proposed tweet cluster algorithm on the RepLab2013 dataset.

Each cluster presented a list of musicians as a topic of the cluster, and the percentages indicate the percentage of tweets about the musicians in each region. For example, the 3rd cluster is composed of 84% 'Lady Gaga', 8% 'PSY', 5% 'the wanted', and 2% 'Madonna'. The topic of this cluster is 'Lady Gaga'. As explained above (see Table 1), most of the cluster results showed good performance. However, the 18th cluster showed an unsatisfactory result. In this cluster, the tweets were more evenly composed of 37% 'the wanted', 25% 'Led Zeppelin', 25% 'AC/DC', and 13% 'Aerosmith'. To find the reason for this low precision result, we analyzed the detailed distributions of all cluster results.

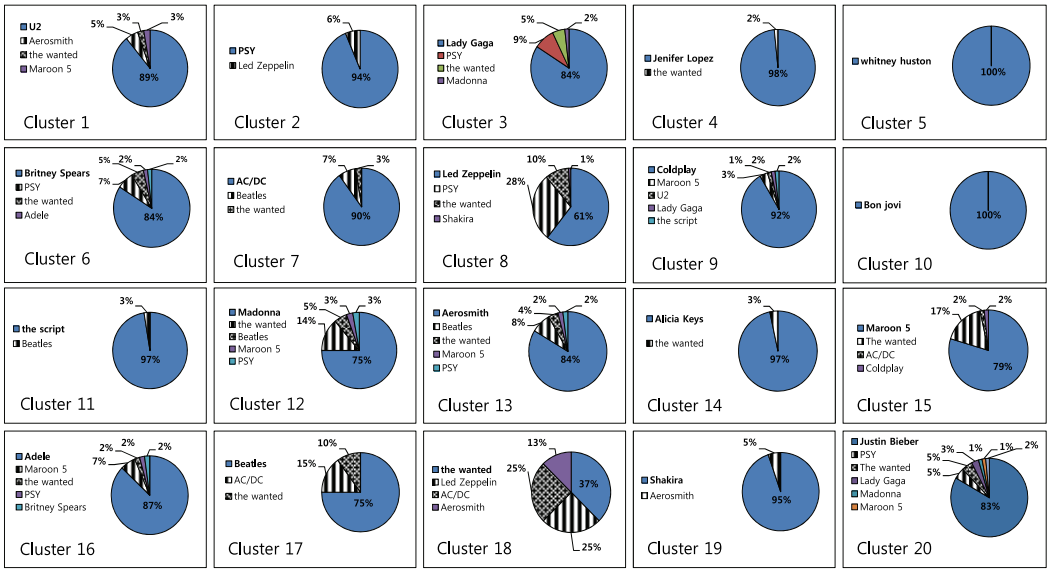

Fig. 3. Ratio of each semantic-based K-means cluster result (K=20).

5. Conclusion

In this paper, we proposed a semantic-based K-means clustering algorithm that efficiently clusters a large amount of microblogs such as Twitter using TagCluster from Flickr, a folksonomy system. The semantic-based K-means clustering algorithm considers not only the similarity between words of a tweet represented by a vector space model, but also the semantic similarity between words of a tweet by semantic expansion using TagCluster. Since a tweet has limited information, semantic expansion to add similar words to a set is very important for more accurately analyzing the tweet. Through experiments on the RepLab2013 Twitter dataset, our proposed tweet clustering performs better than the previous method. In future work, we will explore a more effective microblog clustering algorithm that does not depend on a various domain and topic.

References

[1] M. S. C. Sapul, T. H. Aung, and R. Jiamthapthaksin, "Trending topic discovery of Twitter Tweets using clustering and topic modeling algorithms," in *Proceedings of 2017 14th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, Nakhon Si Thammarat, Thailand, 2017, pp. 1-6.

[2] K. H. Lim, S. Karunasekera, and A. Harwood, "ClusTop: a clustering-based topic modelling algorithm for twitter using word networks," in *Proceedings of IEEE International Conference on Big Data*, Boston, MA, 2017, pp. 2009-2018.

[3] J. Xu, B. Xu, P. Wang, S. Zheng, G. Tian, and J. Zhao, "Self-taught convolutional neural networks for short text clustering," *Neural Networks*, vol. 88, pp. 22-31, 2017.

[4] L. Kotlerman, I. Dagan, and O. Kurland, "Clustering small-sized collections of short texts," *Information Retrieval Journal*, vol. 21, no. 4, pp. 273-306, 2018.

[5] B. Wang, M. Liakata, A. Zubiaga, and R. Procter, "A hierarchical topic modelling approach for tweet clustering," in *Social Informatics*. Cham: Springer, 2017, pp. 378-390.

[6] C. T. Zheng, C. Liu, and H. S. Wong, "Corpus-based topic diffusion for short text clustering," *Neurocomputing*, vol. 275, pp. 2444-2458, 2018.

[7] S. Dhuria, H. Taneja, and K. Taneja, "NLP and ontology based clustering: an integrated approach for optimal information extraction from social web," in *Proceedings of 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, India, 2016, pp. 1765-1770.

Jee-Uk Heu https://orcid.org/0000-0002-1666-1626

He received a B.S. from Hallym University in 2007. He received M.S. and Ph.D. degrees from the School of Computer Science and Engineering at Hanyang University in 2009 and 2016, respectively. Since February 2018, he has been with the Vice Provost for Teaching and Learning (VPTL) at Stanford University as a researcher. His research interests mainly include data mining, semantic analysis, and big data.