The Comparison of SOM and K-means for Text Clustering

Yiheng Chen (Corresponding author)
School of Computer Science and Technology, Harbin Institute of Technology
PO box 321, Harbin, 150001, China
Tel: 86-451-8641-3683 E-mail: cyh@ir.hit.edu.cn

Bing Qin
School of Computer Science and Technology, Harbin Institute of Technology
PO box 321, Harbin, 150001, China
Tel: 86-451-8641-3683 E-mail: qinb@ir.hit.edu.cn

Ting Liu
School of Computer Science and Technology, Harbin Institute of Technology
PO box 321, Harbin, 150001, China
Tel: 86-451-8641-3683 E-mail: tliu@ir.hit.edu.cn

Yuanchao Liu
School of Computer Science and Technology, Harbin Institute of Technology
Harbin, 150001, China
Tel: 86-451-8641-6460 E-mail: lyc@hit.edu.cn

Sheng Li
School of Computer Science and Technology, Harbin Institute of Technology
PO box 321, Harbin, 150001, China
Tel: 86-451-8641-3683 E-mail: lis@ir.hit.edu.cn

Abstract

SOM and k-means are two classical methods for text clustering. In this paper some experiments have been done to compare their performances. The sample data used is 420 articles which come from different topics. K-means method is simple and easy to implement; the structure of SOM is relatively complex, but the clustering results are more visual and easy to comprehend. The comparison results also show that k-means is sensitive to initiative distribution, whereas the overall clustering performance of SOM is better than that of k-means, and it also performs well for detection of noisy documents and topology preservation, thus make it more suitable for some applications such as navigation of document collection, multi-document summarization and etc. whereas the clustering results of SOM is sensitive to output layer topology.

Keywords: Textt Clustering, Self organizing maps, K-means, Clustering algorithm

1. Introduction

With the explosive growth of many text documents such as Web news, electric books and E-mail documents on the internet, the task of how to organize and navigate these documents has become more and more important and urgent. As an unsupervised machine learning method, text clustering has attracted many researchers as it can organize text documents effectively (Ma, Shuai, Wang, TengJiao, 003)(Wang, aihua, Zhang, ming, 2001)(Wu, bin, fu, weipen, Shi, zhonzhi, 2002). The application of text clustering includes: ①text clustering can be used as the preprocess steps of other technology such as multi-document summarization (Vasileios Hatzivassiloglou, Judith L. Klavans, 2001); ② cluster the returned results of search engine, thus users can find what he need quickly(Cutting, D., Karger, D., Pedersen, J. and Tukey, J. W. 1992); ③ automatic organization of text

collection, such as the text navigation system Scatter/Gather developed by Cutting(Cutting, D., Karger, D., Pedersen, J. and Tukey, J. W. 1992); Text clustering can also be used to mine out the interest model of particular user by process the documents or web pages which the user is interested in, thus the narrowed but more related documents can be recommended and pushed to users. In addition, Microsoft's Ji-Rong Wen(JR Wen, JY Nie, HJ Zhang, 2001) has used text clustering technology to cluster the query logs of many users to update the FAQ of websites.

With comparison to other kinds of data, the documents in text clustering have many semantic features and usually are represented as vectors in high-dimension feature spaces. Through the clustering process of input documents, the topic structure hidden in many documents can be found, the documents with same topic or very near topic will be put into one same cluster, whereas documents with different topic will be separated into different clusters. The proof of text clustering is the famous clustering hypothesis(N. Jardine , C. J. van Rijsbergen, 1971): closely related documents belongs to one category and are related to one same query.

Among many text clustering methods, AHC(P. H. Sneath , R. R. Sokal, 1973), k-means, SOM seems to be three kinds of methods and many variant methods have emerged around these technologies. The clustering results of AHC usually are more fine,but the computation cost are also larger than other technologies, whereas the efficiency of k-means, SOM are very higher than AHC clustering technology. The fast increase of network information demands that the computation efficiency must be higher and the clustering results are easy to understand. So k-means and SOM are more popular than AHC method.

As text clustering has many application backgrounds, and different application has different demands for clustering quality, computation efficiency and the navigation ability. So it is necessary to choose suitable text clustering method according to actual application backgrounds.

In this paper, we first made an analysis of the basic principles of two text clustering methods: SOM and k-means. Then some experimental comparisons and the corresponding analysis have been made about the actual performance such as the sensitivity to initial distribution and the F measure in different situations of these two technologies. We anticipate that by our work the actual performance of these two popular text clustering methods can be shown clearly, and can provide some reference for related research.

2. k-means for text clustering

K-means is partition-based clustering method. When k-means is used for text clustering, all the documents will be put into k clusters randomly, and then the clustering partition will be adjusted according to some principles until the clustering results are stable. The basic principle of k-means for text clustering can be depicted as follows:

input:N documents to be clustered, the cluster number k

output:K clusters, and each document will be assigned to one cluster

1) Choose k documents randomly as the initial clustering document seeds;

2) Repeat the following two steps, if the partition is stable, then go to step 5);

3) According to the mean vector of all documents in each cluster, and assign each document into most similar cluster;

4) Update the mean vector of each cluster according to the document vector in it;

5) Output the generated clusters and the partition

3. SOM for text clustering

SOM(Self-Organizing feature Maps)is proposed by professor T.Kohonen(T.Kohonen, 1982).When sufficient training has been made, the output layer of a SOM network will be separated into different regions. And different neurons will have different response to different input samples. As this process is automatic, all the input documents will be clustered. Text documents written by natural language are high-dimension and have strong semantic features. It is hard to navigate many documents in high-dimension space. Whereas SOM can map all these high-dimension documents onto 2- or 1- dimension space, and their relations in the original space can also be kept. In addition, SOM are not very sensitive to some noisy documents and the clustering quality can also be assured. Due to these merits, SOM technology is suitable for text clustering, and has been used in some fields such as digital library(K.Lagus,T.Honkela, S.Kaski, and T.Kohonen, 1996). Many SOM variants also emerges(T.Kohonen, 1998)(D.Merkl, 1993).

The principle of SOM for text clustering can be summarized as follows:

1) Initialization. Assign some random number for all the neurons in the output layer. And normalized. The dimension number of neuron is same to the dimension number of all the documents;

2) Input the sample. Choose randomly one document from the document collection and send it the SOM network;

3) Find the winner neuron. Calculate the similarity between the input document vector and the neuron vector, the neuron with the highest similarity will be the winner;

4) Adapt the vectors of the winner and its neighbors. The adaptation can use the following formula:

$$m_i(t+1) = m_i(t) + \alpha(t) \bullet h_i(t) \bullet [x(t) - m_i(t)] \quad (1)$$

Where $x(t)$ is the document vector or time t, $m_i(t)$ is the original vector of neuron I, $m_i(t+1)$ is the neuron vector after adaptation. $\alpha(t)$ and $h_i(t)$ are the learning rate and neighbor rate respectively. $|x(t) - m_i(t)|$ represent the distance between neuron vector and document vector.

After the adaptation, the winner and its neighbors are more nearer to the input document vector, thus these neurons will be more competitive when similar documents are input again. Through the full training of the SOM network by input sufficient samples, the neurons on the output layers will be only sensitive to the documents of some topic, and their vector will become the mean vector of these topics.

4. The comparison between SOM and K-means

K-means is easy to realize and it usually has low computation cost, so it has become a well-known text clustering method and used by many fields (D.R. Cutting, J.O. Pedersen, D.R. Karger, and J.W. Tukey, 1992)(Daniel Boley, 1998). The shortcoming of k-means is that the value of K must be determined beforehand and the initial document seeds need to be selected randomly. And these initial setting will have impacts on the clustering results. K-means is greedy algorithm in essence; it is hard to attain the global optimum clustering results(R. Ng, J. Han, 1992).

When use k-means to cluster documents, there are some rules:

(1)after the initial settings(the value of k and the document seeds) have been determined, the clustering results will also be determined. But the clustering results will be different if the initial settings are different;

(2)when the initial settings(the value of k and the document seeds) have been determined, suppose the clustering result of the iteration n+1 is same as the iteration n, then the clustering result of iteration n+m will be same as iteration n(m>1). Thus whether the partition has varied can be used as the stop criterion of clustering iteration.

The neuron number of output layer in SOM network has close relation with the class number in the input document collection. If the neuron number is less than the class number, it will be not sufficient to separate all the classes, the documents from some closely-related class may be merged into one class. If the neuron number is more than the class number, the clustering results may be too fine. And the clustering efficiency and the clustering quality may also be adversely affected.

The computation complexity of k-means is O(KlN),where l is the iteration count, N is the document number. The computation complexity of k-means is SOM is O(kmN), where k is the neuron number, m is the training count. The computation complexity of these two methods is very near. They are both lower than ACH method.

5. Experimental results and dissuasion

The actual performance of these two methods was compared through experiments for text clustering. The document collection in experiments has 645 documents which are about different topic. Their basic properties are listed in table 1.

F measure (Michael Steinbach, 2002) is utilized as the evaluation function of system performance. For one generated cluster r and one predefined class s:

$$recall(r,s) = n(r,s)/n_s \quad (2)$$

$$precision(r,s) = n(r,s)/n_r \quad (3)$$

n(r,s) is document number of the intersection between r and s . n_r is document number of cluster r , n_s is document number of class s . F measure between cluster r and class s can be calculated as

$$F(r,s) = (2 * recall(r,s) * precision(r,s)) / ((precison(r,s) + recall(r,s)) \quad (4)$$

The overall F measure can be calculated as

$$F = \sum_i \frac{n_i}{n} \max\{F(i,j)\} \quad (5)$$

Where n is the number of all test documents. n_i is document number of class i . Generally speaking, the bigger value of F measure means better clustering result.

Firstly the impact of the training count on the performance of SOM is examined. Here we define the training count as the C times the size of input documents. For example, if the there are 100 documents to be clustered and C=10, we should set the training count as 100*10=1000. our experiments show that when C is low, the performance of SOM text clustering will grow quickly as the C value increases. When C is high, the performance of SOM is not very sensitive with C value.

One clustering result of SOM has been shown in table 2. the input are the documents from 4 classes: 1,2,3,4. The topology of SOM is rectangular and the output layer includes 2*2=4 neurons, C=20, In table 2, each row represents when C value is different, the number of documents that these 4 neurons TL(top left),TR(top right),DL(down left),DR(down right)has mapped. For example, "TL=29:0:0:5" denotes that neuron TL has mapped 29 documents from class 1, 5 documents from class 4, and no documents from class 2,3. in our experiments, we also found that when the training count is big enough, the purity of text clustering also improves, which will help improve the quality of some natural language process such as multi-document summarization, TDT(Topic detect and track) and etc.

As stated above, both SOM and K-means need a process of initialization. We compared if they are sensitive to the initial settings. We set the k value of SOM and K-means is equal, and compared their performance when k=4 and k=9. for SOM, the topology of its output layer is 2*2=4 and 3*3=9, C=20. when the training is over, each neuron in the output layer of SOM denotes documents from one class.

In table 3 and table 4 the average F measure of 20 running of both methods is shown. SOM is not sensitive to the initial settings. Whereas the clustering results of k-means is not stable and the iteration count is also different for each running. In fact, if suitable initial document seeds can be selected, k-means will converge quickly and better clustering quality can be achieved. As standard k-means usually select seeds randomly, the clustering quality will be affected adversely. Thus when k-means is used for text clustering, it is necessary to use some method to select suitable seeds(such as min-max principle, density-based method and etc.)

Our experimental results also proves that when the neuron number is more than the class number of input documents, as the training of SOM tend to utilize each neuron fully, some class may be represented by more than 2 neurons. In this situation, the documents from these classes will usually be mapped onto some neighboring neurons, as shown in table 5 and table 6. In table 5, there are 3*3=9 neurons in the output layer, and there are 6 classes in the input documents. Neuron N11, N33 can represent one class respectively, whereas N21, N31 actually represent one common class. In table 6, there are 2*4=8 neurons, and input documents have 5 classes. Neuron N21 itself can represent one class. Whereas neuron N11, N12, N22 actually represent one common class.

All these experimental results demonstrate that the topology of SOM has clear impacts on the clustering quality. However, the clustering results of SOM can provide good navigation ability, thus made the clustering meaningful and easy to understand. In the output layer of SOM, neighboring neurons usually maps similar documents. The documents from same topic or similar topic will be mapped onto same neuron or near neurons, Thus users can find the documents they need very quickly and the information access efficiency can be improved greatly. In many applications more neurons can be set (more than the possible cluster number) to cluster documents. In comparison, k-means need users to provide k value to start clustering. The unsupervised property of text clustering will be affected, as in most situations, users know little about the topic structure of input documents.

The clustering quality of SOM and k-means is compared directly in some situations (the number of neurons in the output layer of SOM is same to the k value in k-means). When the output layer of SOM is 2*2, 2*3, 2*4 and 3*3, F measure of both methods is shown in table 7. Each time 4 combinations of document class have been selected, and the mean value of 10 clustering results are utilized as the overall F measure. It can be seen that the overall clustering quality of SOM is better than K-means clearly. That suggests that when the setting of output layer of SOM is reasonable, I,e, the neurons in the output layer can be used fully, SOM can achieve better

clustering quality. The clustering performance of k-means is very sensitive to the initial settings, thus make its clustering quality is not stable and its F measure is less than SOM.

6. Conclusion

In this paper the performance of two text clustering method: SOM and k-means has been analyzed and compared by experiments. The experimental results demonstrate that k-means is very sensitive to the initial settings such as k value and document seeds. Whereas SOM can achieve better text clustering quality when the neurons in the output layer can be utilized fully. SOM also performs better in noise toleration and topology preservation and make it a text clustering method to be studied furthermore.

References

Cutting, D., Karger, D., Pedersen, J. and Tukey, J. W. (1992). Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. *Proceedings of the 15th Annual International ACM/SIGIR Conference*, Copenhagen, 1992:318-329.

D.Merkl. (1993). Structuring software for reuse: the case of self-organizing maps. In International Joint Conference on Neural Networks. Nagoya Congress Center, Japan, 1993,(3):2468-2471

D.R. Cutting, J.O. Pedersen, D.R. Karger, and J.W. Tukey. (1992). Scatter/gather: A cluster-based approach to browsing large document collections. In Proceedings of the ACM SIGIR, Copenhagen, 1992:318–329

Daniel Boley. (1998). Principal direction divisive partitioning. Data Mining and Knowledge Discovery. 1998, 2(4): 325-344

JR Wen, JY Nie, HJ Zhang. (2001). Clustering User Queries of a Search Engine. Tenth International World Wide Web Conference..Hong Kong, 2001:162-168.

K.Lagus,T.Honkela, S.Kaski, and T.Kohonen. (1996). Self-organizing maps of document collections: A new approach to interactive exploration. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, California. 1996:238-243.

M. Steinbach, G. Karypis, and V. Kumar. (2000). A comparison of document clustering techniques. In KDD Workshop on Text Mining, Boston, MA, USA.

Ma, Shuai, Wang, TengJiao. (2003). A Fast Clustering Algorithm Based on Reference and Density, *Journal of Software* .2003,14(6):1089-1095.

Michael Steinbach, George Karypis, Vipin Kumar. (2002). A Comparison of Document Clustering Techniques. Department of Computer Science and Engineering, University of Minnesota. Technical Report #00-034.

N. Jardine, C. J. van Rijsbergen. (1971). The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval.* 1971(7):217-240.

P. H. Sneath, R. R. Sokal. (1973). *Numerical Taxonomy*. Freeman, London, UK, 1973.

R. Ng, J. Han. (1994). Efficient and effective clustering method for spatial data mining. In Proc. of the 20th VLDB Conference, Santiago, Chile, 1994:144–155

T.Kohonen. (1982). Self-organized formation of topologically correct feature maps. Biological Cybernetics. 1982(43):59-69

T.Kohonen. (1998). Self-organization of very large document collections: State of the art. Proceedings of ICANN98, the 8th International Conference on Artificial Neural Networks, Springer, London. 1998(1):65-74.

Vasileios Hatzivassiloglou, Judith L. Klavans. (2001). SIMFINDER: A Flexible Clustering Tool for Summarization. NAACL Workshop on Automatic Summarization (Pittsburgh, PA), 2001:41-49.

Wang Aihua, Zhang Ming. (2001). PCCS:A Fast Clustering And Classification Method For Web Document. *Journal of Computer Research and Development*. 2001,38(4):415-421.

Wu Bin, Fu Weipen, Shi Zhonzhi. (2002). A Clustering Algorithm Based On Swarm Intelligence For Web Document. *Journal of Computer Research and Development.* 2002,39(11):1429-1435.

Table 1. Basic Information about Datasets

Category ID	Category Description	Documentation Number	Category ID	Category Description	Documentation Number
1	Crustacea	50	9	MBA	40
2	apple	50	10	MP3	40
3	sealing	50	11	game	40
4	Lu Yongxiang(Chinese name)	40	12	Jordan (English name)	40
5	Li Guojie (Chinese name)	45	13	Tsinghua University	50
6	Digital Cameras	50	14	Tourism	50
7	Joke	50	15	Lenovo	50
8	music	50	16	Health	50

Table 2. the impact of C value on the performance of SOM(input documents from class 1-4)

C	F measure	TL	TR	DL	DR
1	0.79	29:0:0:5	0:0:0:25	0:16:28:0	1:14:2:0
2--4	0.79	0:0:0:25	29:0:0:5	0:16:28:0	1:14:2:0
5--9	0.79	0:0:0:25	29:0:0:5	0:16:28:0	1:14:2:0
12	0.88	28:0:0:6	0:0:0:24	2:30:2:0	0:0:28:0
13	0.88	0:0:0:24	29:0:0:6	0:0:28:0	1:30:2:0
14	0.88	30:0:0:6	0:30:2:0	0:0:0:24	0:0:28:0
15	0.90	0:0:28:0	1:30:2:0	0:0:0:26	29:0:0:4
16	0.90	0:0:28:0	28:0:0:3	2:30:2:0	0:0:0:27
17--19	0.92	3:0:0:29	27:0:0:1	0:0:29:0	0:30:1:0
20	0.93	29:0:0:4	1:30:2:0	0:0:0:26	0:0:28:0
50	0.93	29:0:0:5	0:0:30:0	0:0:0:25	1:30:0:0
100	0.93	30:0:0:6	0:30:2:0	0:0:0:24	0:0:28:0

Table 3. The impacts of initial settings on the F measure of SOM and k-means
(input: documents of class 1-4; dimension number:645; SOM output layer:2*2)

No.	SOM (C =20)	k-means (iteration count)	No.	SOM (C =20)	k-means (iteration count)
1	0.93	0.87(7)	11	0.93	0.87(6)
2	0.93	0.65(10)	12	0.93	0.92(7)
3	0.93	0.92(7)	13	0.93	0.65(7)
4	0.93	0.67(7)	14	0.93	0.66(6)
5	0.93	0.85(10)	15	0.93	0.72(7)
6	0.93	0.92(7)	16	0.93	0.85(9)
7	0.93	0.65(2)	17	0.93	0.65(9)
8	0.93	0.65(2)	18	0.93	0.90(6)
9	0.93	0.94(6)	19	0.93	0.92(7
10	0.93	0.65(7)	20	0.93	0.87(6)

Table 4. The impacts of initial settings on the F measure of SOM and k-means
(input: documents of class A,B,C,D,E,F,G,H,I; dimension number:912; SOM output layer:3*3)

No.	SOM (C =20)	k-means (iteration count:7-15)	No.	SOM (C =20)	k-means (iteration count:7-15)
1	0.91	0.92	11	0.91	0.88
2	0.91	0.75	12	0.91	0.75
3	0.91	0.90	13	0.91	0.75
4	0.91	0.89	14	0.91	0.98
5	0.91	0.91	15	0.91	0.88
6	0.91	0.75	16	0.91	0.75
7	0.91	0.89	17	0.91	0.75
8	0.91	0.91	18	0.91	0.93
9	0.91	0.88	19	0.91	0.88
10	0.91	0.90	20	0.91	0.75

Table 5. The Class Division Phenomenon When Node Number Is Greater Than Class Number
(Input: Six Classes Including A,B,C,D,E,F; SOM Output Layer: 3*3)

	Column 1	Column 2	Column 3
Line 1	0:0:30:0:0:0(N11)	15:1:0:4:2:0(N21)	14:0:0:0:0:0(N31)
Line 2	0:0:0:0:0:13(N12)	1:1:0:8:0:0(N22)	0:0:0:18:0:0(N32)
Line 3	0:12:0:0:0:17(N13)	0:16:0:0:0:0(N23)	0:0:0:0:28:0(N33)

Table 6. The Class Division Phenomenon When Node Number Is Greater Than Class Number
(Input: Five Classes Including 1-5; SOM Output Layer: 2*4)

	Column 1	Column 2	Column 3	Column 4
Line 1	0:0:0:5:0(N11)	0:0:0:0:30(N21)	0:0:4:0:0(N31)	30:0:0:0:0(N41)
Line 2	0:0:0:19:0(N12)	0:0:0:6:0(N22)	0:0:26:0:0(N32)	0:30:0:0:0(N42)

Table 7. The Performance Comparisons between SOM and K-means (K Value is known)

classes	F measure		classes	F measure	
	SOM(output layer)	K-means		SOM(output layer)	K-means
1-4	0.93(2*2)	0.76	1-8	0.92(2*4)	0.78
5-8	0.91(2*2)	0.81	3-10	0.87(2*4)	0.78
9-12	0.84(2*2)	0.78	5-12	0.86(2*4)	0.81
13-16	0.93(2*2)	0.80	7-14	0.92(2*4)	0.90
1-6	0.86(2*3)	0.73	9-16	0.91(2*4)	0.83
3-9	0.92(2*3)	0.81	1-8	0.81(3*3)	0.70
6-12	0.86(2*3)	0.71	5-12	0.85(3*3)	0.80
12-16	0.87(2*3)	0.79	9-16	0.89(3*3)	0.84