7/10 2007 - 6 pages - 2021.12.29 - supplementary - formular inspiration

Available online at www.sciencedirect.com
ScienceDirect
Neurocomputing 71 (2008) 857–862
NEUROCOMPUTING
www.elsevier.com/locate/neucom

ConSOM: A conceptional self-organizing map model for text clustering

Yuanchao Liu[a,b,*], Xiaolong Wang[a], Chong Wu[b]

[a]Department of Computer Science and Technology, Harbin Institute of Technology, 150001 Harbin, PR China
[b]Department of Administration, Harbin Institute of Technology, 150001 Harbin, PR China

Received 2 July 2006; received in revised form 26 January 2007; accepted 8 March 2007
Communicated by E.W. Lang
Available online 14 April 2007

Abstract

In the novel conceptional self-organizing map model (ConSOM) proposed for text clustering in this paper, neurons and documents can be represented by two vectors: one in extended concept space, and the other in traditional feature space, and weight modification of neuron vector is guided by combination of similarities in both traditional and extended spaces. Experimental results show that by utilizing concept relevance knowledge effectively, ConSOM performs better than traditional "SOM plus VSM" mode in text clustering due to its semantic sensitivity.

Keywords: SOM; Neuron vector; Weight modification; HowNet; Text clustering

1. Introduction

Self-organizing map (SOM) proposed by professor Kohonen [7] is a powerful nonlinear projection model that can produce a 2- or 1-dimension similarity graph of high-dimension input data in an orderly fashion [13]. Organizing texts according to their similarities can be achieved using SOM when texts are characterized by the words they contain [8,9]. Under traditional "SOM plus VSM" mode (using Vector Space Model to represent documents and use Self-organizing map to cluster documents automatically), all neurons and documents can be represented by vectors in a high-dimension space according to frequency information [6,10].

For a particular dimension (usually one word) in a traditional feature space, a document will have a positive weight value on this dimension if the document has this word, otherwise the weight value on this dimension will be zero. TFIDF formula is often calculated from the term frequency in the document multiplied by the inverse document frequency and is used to assign a value to a term (word). The similarity between vectors is usually calculated as the cosine of the angle between them [2]. Therefore, if the similarity of two vectors is great, then in most situations it is because they share more words. Whereas if two documents share few words, it usually means that they will be assigned by text clustering system into different clusters. Due to complexity of natural language, many different terms may have the same sense or senses which are very near to each other. Traditional "SOM plus VSM" clustering mode cannot be used to grasp rich natural language information hidden in documents and the clustering results are not ideal sometimes.

In the training process of standard SOM, quite often similarities between an input document and all the neurons are calculated. Only the neuron with the highest similarity and its nearest neighbors have the chance to modify their vector weights. It is an important step to calculate similarities for self-organizing maps. SOM researchers usually focus on how to adapt network structure while training [1,5], and pay little attention to construction of feature space and similarity calculation. In fact, if the result of similarity computation is not consistent with people's anticipation, modification direction of neuron vector will be adversely affected, and the cluster quality will be poor.

Therefore, a novel conceptional self-organizing map model (ConSOM) is proposed for text clustering. In ConSOM, all input documents and neurons are represented by two vectors: one in traditional feature space and

*Corresponding author. Tel.: +86 451 86413322; fax: +86 451 86413309.
E-mail address: lyc@insun.hit.edu.cn (Y. Liu).

0925-2312/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2007.03.006

the other in extended concept space. Concept space is constructed using concept words of all the documents, while the concept words of each document are formed by extending document vectors. Weight modification of neuron vector is guided by the combination of similarities in both traditional feature space and the extended concept space, and consequently ConSOM can be effectively trained and relevant documents can be assigned into one cluster despite some of them sharing only few words.

2. Constructing ConSOM based on SOM and concept relevance model

While documents are clustered under standard "SOM plus VSM" mode, the dimensions in feature space come from all documents and the vector of each document is formed by the words in that document. For a particular dimension (usually a word), a document will have zero value if it does not contain this word. The similarity between two documents about the same topic is usually small if they share few words, and they will be assigned wrongly to different clusters. However, in most cases people prefer to cluster documents by semantics. Even when many documents have few words in common, people may still think they are about the same topic.

For example, of the documents in a collection about different topics, such as politics, economics, sports, computer, military, etc., the documents about the same topic may share few words. When documents are clustered under standard mode, it is hard to bring the documents about same topic together. However, we can notice an interesting characteristic: the documents about the same topic are very relevant to each other and many words in these documents are also very relevant to each other. The calculation of relevance between documents can be simplified by examining if the words in two documents are relevant to each other. For example, "将军"(general in Chinese) and "士兵"(soldier in Chinese) are very relevant to each other, whereas "医生"(doctor in Chinese) 和 "士兵"are not relevant to each other. In order to effectively calculate similarity between two documents, a document vector can be extended into a concept vector; the similarity can also be calculated in the extended concept space. In order to decrease computation cost, only a few high-frequency words (with stopwords removed) in documents can be selected and extended.

It is therefore necessary to generate an extended concept vector for each document. The first thing to do is find relevant words for some high-frequency words in each document. We use *HowNet* [4] to find relevant words for a particular word. *HowNet* is a common-sense knowledge base that can be used to unveil the inter-conceptual relations and inter-attribute relations of concepts in both English and Chinese languages, and it can be used to demonstrate general and specific properties of concepts. There are 153, 070 sense records in *HowNet* and each sense record has a unique id. One word may have different sense records if it has different meanings. Each sense record is composed of some sememes (1500 terms have been selected as sememes in *HowNet*, which refer to the smallest basic semantic unit that cannot be reduced further), as shown in its DEF entry. One word will be relevant to another if they share some sememes. For example, the sense records of words "将军"and "士兵"are listed in Table 1, they share some sememes in DEF section of the *HowNet* entries. *Hownet* 2004 has three levels for word relevance (level 0, level 1 and level 2). Level 2 is the least relevant level, whereas level 0 is the most relevant level and is used in this paper. The concept extension algorithm for one document is depicted as follows:

Algorithm 1. Concept extension algorithm

Input: Document d
Output: Concept vector V_d

1. Segment document d into a bag of words, remove stopwords and obtain the document vector V';
2. Count the frequency for each word in V', and push $\eta^*|V'|$ words that are most frequent into S';
3. **While** S' is not null, pop out word w_i from S';
   (a) Find w_i in *HowNet*, get sense record set $S(w_i)$,
   (b) If($|S(w_i)|>0$);

Table 1
The sense records of Chinese word "将军"and "士兵"in *HowNet*

将军(general)	士兵(soldier)
NO.=055253	NO.=096817
W_C=将军	W_C=士兵
G_C=N	G_C=N
E_C=	E_C=
W_E=general	W_E=privates
G_E=N	G_E=N
E_E=	E_E=
DEF={human| 人 :belong={army| 军队 },modifier={official| 官 },{fight| 争斗:agent={~},domain={military|军}}}	DEF={human|人:belong={army|军队},{fight|争斗:agent={~},domain={military|军}}}

i. **For** every sense record r_{ij} in $S(w_i)$ $(0 \leqslant j \leqslant |S(w_i)|-1)$
   1. Find all words which are relevant to r_{ij}, and put them into S_{ij};
   2. $I_{ij} = S_{ij} \cap S'$;

ii. **End For**

(c) Find possible sense record m, $|I_{im}| = \max(|I_{i0}|, |I_{i1}|, \ldots |I_{ij}|, \ldots |I_{|S(w_i)|-1}|)$;

(d) If $|S_{im} \cap S'| \geqslant N_\theta$, Push all words in S_{im} into S';

4. **End While**;
5. Count the word frequency in S'' and output concept word vector V_d;

In Algorithm 1, |*| means the number of elements in set * η denotes the ratio of high-frequency words in document d, η is set to be 0.1 by experiences to reduce dimension number so that the clustering quality will not be adversely affected. After finding w_i in *HowNet* in Step 3(a) in Algorithm 1, the returned sense record set is denoted as $S(w_i)$. We need to select only one possible sense record of word w_i in its original document by comparing the number of elements for all the intersections $|I_{i0}|, |I_{i1}|, \ldots, |I_{ij}|, \ldots, |I_{|S(w_i)|-1}|$, find the biggest value $|I_{im}|$ and use sense record m to retrieve relevant words in *HowNet*, I_{ij} means the intersection between the relevant words of one sense record r_{ij} and the words in S'.

For each word in S', if the intersection of relevant words and high-frequency words has less than N_θ words ($N_\theta = 3$ in this paper), this word need not be extended. In addition, as some relevant words returned by HowNet are not words in essence, they are phrases such as "grain crops", "cotton field", "wheat flour", etc.; these phrases are usually composed of two or more words. They are filtered and would not be added to the concept vector. Concept space can be constructed after all documents have been processed, and concept vector V_C for both input document and neuron can then be formed in this space.

3. Similarity computation between neuron and document

After both extended concept space and traditional feature space are constructed, all documents and neurons are represented by two vectors: traditional vector V_F purely formed by word frequency and extended concept vector V_C, as shown in Fig. 1. Concept vector can incorporate more semantic factors into ConSOM clustering model. Even though the documents about the same topic share few words, the similarity between them can still be big enough. While we calculate the overall similarity, V_F is still kept to prevent ConSOM from over-extending of concepts. For example, if concept vectors of two different documents d_m and d_n are exactly the same, their similarity in the extended concept space will be 1.0. For two documents d_i and d_j which are the same (d_i is a copy of d_j), it is apparent that they have the same concept vector and their similarity in concept space is also 1.0. It is not difficult to understand

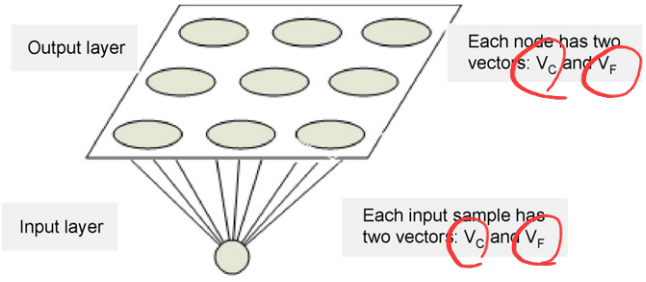

Fig. 1. Basic model of ConSOM.

that the overall similarity between d_i and d_j should be bigger than that between d_m and d_n. Therefore after concept extension, it is still necessary to keep V_F as a document feature. While the model is trained, a general formula that combines the roles of V_F and V_C is used to calculate the overall similarity between input document d and neuron n_i.

$$sim(n_i, d) = \alpha^* sim_1(n_i, d) + (1-\alpha)^* sim_2(n_i, d), \quad 0 \leqslant \alpha \leqslant 1, \tag{1}$$

where n_i is one of the neurons in the ConSOM output layer, d is an input document. sim_1 is the similarity between d and n_i in the extended concept space, sim_2 is the similarity between d and n_i in the standard feature space, α is the impact ratio of concept relevance knowledge. When $\alpha = 0$, $sim(n_i,d)$ is equal to sim_2, i.e. the traditional "SOM plus VSM" mode. As α increases, the similarity in the concept space plays a more and more important role in overall similarity computation.

sim_1 and sim_2 can be calculated as cosine value of the angle between vector [11]. For example, the cosine between two vectors A and B can be calculated as

$$\cos(A, B) = \frac{\sum_{J=1}^{n} w_{A_j} w_{B_J}}{\sqrt{\sum_{J=1}^{n} w_{A_j}^2} \sqrt{\sum_{J=1}^{n} w_{B_j}^2}}. \tag{2}$$

It should be noted that the feature spaces used for computing sim_1 and sim_2 are different. Winner n_{win} is the neuron with the maximum $sim(n_i,d)$.

$$sim(n_{\text{win}}, d) = \max(sim(n_i, d)), \quad 0 \leqslant i \leqslant c-1, \tag{3}$$

where n_i represents a neuron, winner n_{win} and its neighbors have the chance to modify their vector weight. After modification, these neurons move nearer toward the input sample.

Other issues in ConSOM, such as the decrease of learn rate and the neighborhood range is the same as the setting of a standard SOM implementation SOMlib [3], which is an open-source project (http://www.ifs.tuwien.ac.at/~andi/somlib/download/index.html), whereas the difference lies in the construction of feature space and the similarity computation between a neuron and a document, as shown in Sections 2 and 3.

Table 2
Dataset

Dataset	Description of Data set	Feature vector dimension	Concept vector dimension	Source
D1	Wheat, grain, ship, trade	494	723	Reuters 21578
D2	Corn, wheat, grain, ship	441	652	Reuters 21578
D3	Space, Auto, guns, medicine	612	716	20 newsgroups
D4	Space, baseball, Christian, medicine	568	629	20 newsgroups
D5	教育 (education), 军事 (martial), 交通 (traffic), 计算机 (computer)	575	791	http://news.sina.com
D6	政治 (politics), 经济 (economics), 文化 (culture), 军 事 (martial)	742	915	http://news.sina.com

4. Experiment results

The widely used evaluation criterion for text clustering is F measure [12]. For generated cluster r and predefined class s, the corresponding recall and precision can be calculated as follows:

$$precision(r,s) = n(r,s)/n_r, \tag{4}$$

$$recall(r,s) = n(r,s)/n_s. \tag{5}$$

$n(r,s)$ is the document number of the intersection between cluster r and class s. n_r is the number of documents in cluster r, and n_s is the number of documents in class s. F measure between cluster r and class s is as shown below calculated:

$$F(r,s) = (2^{*}recall(r,s)^{*}precision(r,s))/((precison(r,s) + recall(r,s)). \tag{6}$$

The overall F measure can be given by

$$F = \sum_i \frac{n_i}{n}\max\{F(i,j)\}, \tag{7}$$

where n is the number of input documents and i is each predefined class. The performance of ConSOM was evaluated using different data sets. Each data set has 160 documents about different topics, and contains information as shown in Table 2.

In order to directly compare clustering qualities of ConSOM and standard "SOM plus VSM" mode, the number of output layer neurons is set equal to the class number of input documents. Our objective is to examine the performance improvement after importing concept relevance knowledge, so there must be no other factors involved and the other settings of both the models must be kept as same as possible. The relations between F measure and impact factor α in these data sets are shown in Figs. 2–7. The F measure in each figure is the mean value of 10 runs with different random seeds. When "$\alpha = 0$", concept relevance knowledge is not used (the standard "SOM plus VSM" mode), the clustering quality is poor in all these data sets. F measure increases as α increases, and especially when α is 0.6~0.7, our clustering system has achieved the best performance; after this, the value of F measure begins to decrease. It can be seen from the analysis

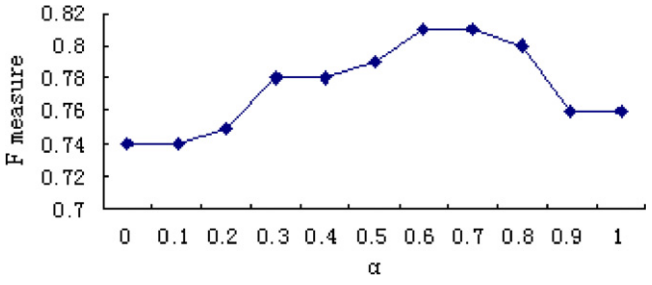

Fig. 2. Relation between F measure and impact factor α in data set 1.

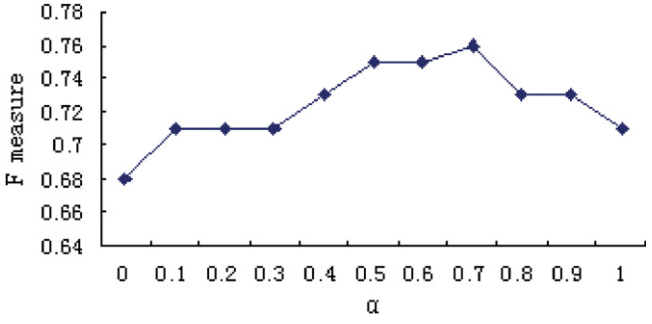

Fig. 3. Relation between F measure and impact factor α in data set 2.

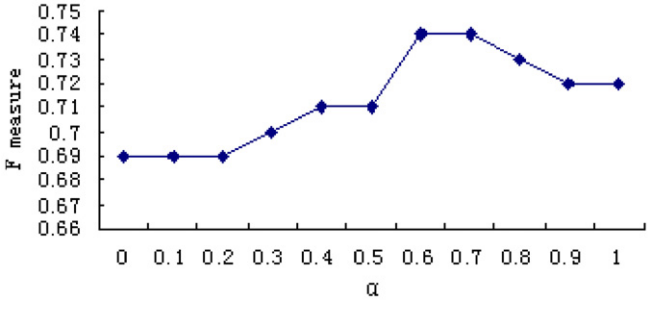

Fig. 4. Relation between F measure and impact factor α in data set 3.

of all these curves that high success can be achieved by combining more concept relevance knowledge, whereas it is still necessary to keep sim_2 in the overall similarity computation formula to attain the best cluster quality.

It is apparent that for all these data sets, the clustering quality of ConSOM is better than standard "SOM plus

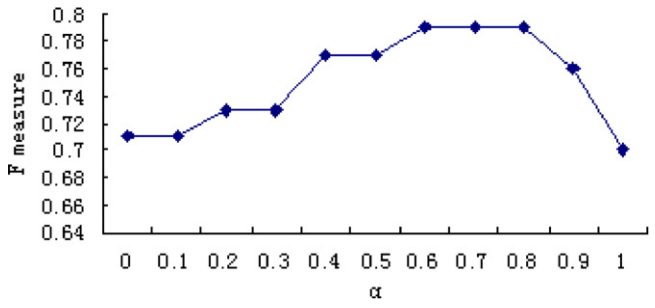

Fig. 5. Relation between F measure and impact factor α in data set 4.

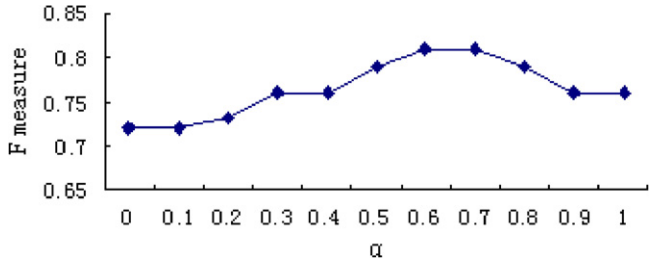

Fig. 6. Relation between F measure and impact factor α in data set 5.

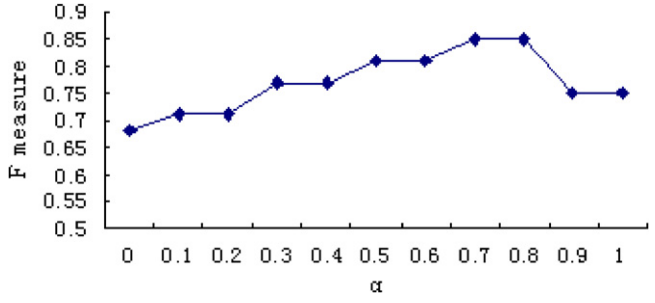

Fig. 7. Relation between F measure and impact factor α in data set 6.

VSM'' mode (when $\alpha = 0$), because apart from traditional word frequency factor, there are more semantic factors involved in ConSOM. Although each class in the data set has one common topic (see description of the data set in Table 2), there are usually few common words among the documents of the same class. This is the reason why we propose conceptional SOM model to cluster documents.

5. Conclusion

In this paper, a novel text clustering model (referred to as ConSOM) that combines the advantages of both SOM and concept relevance knowledge is proposed. In ConSOM, all neurons and documents are represented by two vectors: one is traditional vector that is based purely on word frequency; the other is based on concept relevance model. Weight modification of neuron vector is guided by overall similarity between neuron and document. In this way, the documents about same topic can be assigned into one cluster even though they share few words. By importing concept relevance knowledge, ConSOM are more sensitive to semantics and can achieve better performance than traditional ''SOM plus VSM'' mode.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (key program: 60435020) and the jointed key laboratory foundation of MOE of China and Microsoft Corporation (01307620).

References

[1] J. Blackmore, R. Miikkulainen, Incremental grid growing: encoding high-dimensional structure into a two-dimensional feature map, in: Procceedings IEEE International Conference Neural Networks (ICNN'93), Piscataway, IEEE, New York, 1993; pp. 450–455.
[2] D.R. Cutting, J.O. Pedersen, D.R. Karger, J.W. Tukey, Scatter/gather: a cluster-based approach to browsing large document collections, in: Proceedings of the ACM SIGIR. Copenhagen, 1992; pp. 318–329.
[3] M. Dittenbach, D. Merkl, and A. Rauber. The growing hierarchical self-organizing map, in: Proceedings of the International Joint Conference on Neural Networks 2000, IJCNN'2000, Como, Italy, 2000; pp. 24–27.
[4] Z. Dong, Q. Dong, *HowNet* knowledge database 〈http://www.keenage.com/.keenage.com〉, 2003.
[5] B. Fritzke, Growing grid: a self-organizing network with constant neighborhood range and adaptation strength, Neural Process. Lett. 2 (5) (1995) 9–13.
[6] D.J. Harper, M. Mechkour, G. Muresan. Document, Clustering for mediated information access, in: Proceedings of the 21st Annual BCS-IRSG Colloquium, 1999; pp. 92–107.
[7] T. Kohonen, Self-organized formation of topologically correct feature maps, Biol. Cybern. 43 (1) (1982) 59–69.
[8] K. Lagus, T. Honkela, S. Kaski, T. Kohonen, WEBSOM for textual data mining, Artif. Intell. Rev. 13 (1999) 345–364.
[9] L. Niklasson, M. Bodén, Ziemke, Self-organization of very large document collections: state of the art, in: Proceedings of ICANN98, The 8th International Conference on Artificial Neural Networks 1998; pp. 65–74.
[10] G. Salton, A. wong, C. Yang, A vector space model for automatic indexing, Commun. of the ACM 18 (11) (1975) 613–620.
[11] A. Strehl, J. Ghosh, R. Mooney, Impact of similarity measures on web-page clustering, in: AAAI 2000 Workshop on AI for web Search. Austin, TX USA, 2000; pp. 58–64.
[12] B.B. Wang, R.I. McKay, H.A. Abbass, et al., A comparative study for domain ontology guided feature extraction, in Procceedings of 26th Australian Computer Science Conference (ACSC2003), Australian Computer Society Inc, Darlinghurst, Australia: 2003; pp. 69–78
[13] M.-F. Yeh, K.-C. Chang, GraySOFM network for solving classification problems, Neurocomputing 67 (2005) 281–287.

Yuanchao Liu was born in People's Republic of China in 1971; he received his B.S. and M.S. degrees from Harbin Institute of Technology in 1995 and 1999, respectively; his Ph.D. degree in Computer Science and Technology from Harbin Institute of Technology in 2006. He is now an associate professor in both the key laboratory of MOE of China and Microsoft Corporation in Harbin Institute of Technology. His current research interests include data mining, machine learning and text clustering.

Xiaolong Wang received his B.S. degree from Harbin Institute of Electrical Technology, China, and his M.S. degree from Tianjin University, China, in 1982, and 1984, respectively, and his Ph.D. degree in Computer Science and Engineering from Harbin Institute of Technology, China, in 1989. He joined Harbin Institute of Technology, China, as an Assistant Lecturer in 1984; followed by an Associate Professor in 1990; followed by a senior research fellow at the polytechnic University from 1998 to 2000. He is now a Professor of computer Science at Harbin Institute of Technology, China, and his research interest includes network information processing, artificial intelligence, natural language processing, computational molecular biology, and business intelligence.

Chong Wu was born in People's Republic of China in 1971; He received his B.S. and Ph.D. degrees in Mathematics from Harbin Institute of Technology in 1993 and 1998, respectively. He is now a professor in Department of Administration in Harbin Institute of Technology. His current research interests include data mining, machine learning.