Clustering of Words in Texts Using Fuzzy Neighborhood

Zhang Canlun, Sadaaki Miyamoto
Department of Risk Engineering
University of Tsukuba
Tsukuba, Ibaraki, Japan
{s1320646, miyamoto}@risk.tsukuba.ac.jp

***Abstract*—In this paper, we study the clustering of keywords in documents. We consider a model of fuzzy neighborhood for measuring similarity between words. Fuzzy neighborhoods lead to positive-definite kernels. By using the methods of kernel-based fuzzy c-means and affinity propagation, we show the results of clustering for Chinese documents on the Web. Moreover, Rand index is used to compare robustness of hard and fuzzy c-means algorithms with respect to different initial values.**

Keywords- kernel-based clustering; fuzzy neighborhood; c-means; keywords in texts.

I. INTRODUCTION

With the development of information and communication technology, the Web became a large and complex resource of information. However, by processing of the data from Internet having non-numeric attributes of huge volume, it is difficult to use them effectively and easily, although there is a huge amount of important data being used widely in daily life. In order to provide a better usability of Web data, a technique called text mining has attracted attention of many users. In China, with the spread of Internet, the demand and opportunity to handle a large number of documents, blogs, Twitter, and news Web pages is increasing very rapidly. This has attracted attention for text mining.

The purpose of this study is to show the effectiveness of clustering [1] using a fuzzy neighborhood model [2] [3] that regards a document as a sequence of words/terms. By extracting words of interest from text data on the Web, we calculate the similarity of words based on the neighborhood model and we obtain clusters of words using this similarity. A feature of this model is that the similarity satisfies the property of positive-definite kernel function [4]. This method is applied to a set of Chinese documents on the Web, where the methods of hard c-means, fuzzy c-means [1], and affinity propagation [5] [6] are used. A problem of c-means is the dependence of clusters on initial values. We use Rand index [7] to compare the degrees of dependence by different algorithms.

The rest of this paper is organized as follows: Section 2 describes the methods that are used for clustering, which is the main part of this paper. The methods of kernel-based fuzzy c-means as well as kernel-based affinity propagation are described. Section 3 shows the application of these methods to a Chinese document. Finally, Section 4 concludes the paper.

II. MODEL FOR CLUSTERING

The purpose here is to generate clusters of words or terms in a text. When we have a set of texts, they are handled as a single long text by concatenating them.

A. Fuzzy Neighborhood Model

A fuzzy neighborhood is a model for text data analysis. It does not just consider the numbers of word appearances, but also consider the location of the words in the text. We assume that when two words are near, a similarity between these words is higher.

All the items of fuzzy neighborhood model used in this study are defined as follows:

A word set or term set $T = \{t_1, \dots, t_m\}$ is the objects for clustering.

An occurrence set $o, o', o_i \dots \in O$ is assumed. An occurrence *o* is different from a term *t* in the sense that a term can appear more than once in a same document.

A text is a sequence of occurrences: $oo'o''$ ..., where terms are occurring several or many times; the relation between an occurrence and a term is described by the next relation.

A fuzzy relation between *T* and *O* is assumed: when *t* and *o* imply the same term, $R(t, o) = 1$; when the t and o are different term, $R(t, o) = 0$, when *t* and *o* have fuzzy relation, $0 < R(t, o) < 1$.

D is the distance between occurrences *o* and o' in a text, which is defined as follows:

$$D(o, o') = \{\text{number of occurrences between } o \text{ and } o'\} + 1 \quad (1)$$

Another fuzzy relation *N* of $O \times O$ is defined by using the distance, which is called a *fuzzy neighborhood*. $N(o_i, o_j)$ means the degree of nearness between terms o_i and o_j;

A function f defines a particular form of *N*, $f: \{0,1,2, \dots\} \rightarrow [0,1]$; *f* satisfies the next conditions:

1. $f(0) = 1$,
2. $lim_{x \rightarrow \infty} f(x) = 0$,
3. f is monotone non-increasing.

Using *f* and *D*, *N* is defined as follows:

$$N(o,o') = f(D(o,o')) \ . \quad (2)$$

The fuzzy neighborhood has symmetry and reflectivity, which is shown as follows:

$$N(o,o') = N(o',o) \quad (3)$$

$$N(o,o) = 1. \quad (4)$$

The formula shown below is a typical example of fuzzy neighborhood defined by the distance:

$$N(o,o') = \max\{0, 1 - \frac{D(o,o')}{L}\}, \quad (5)$$

where L is a positive constant.

Figure 1 shows an example of a functional form of fuzzy neighborhood. The center is an occurrence of a term.

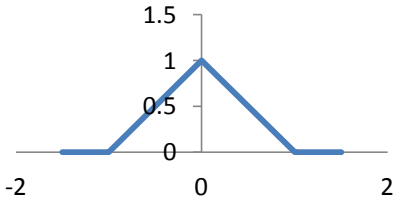

Figure 1. An example of fuzzy neighborhood

A relation $p(t,t')$ is a similarity measure between two words, which is defined as follows:

$$p(t,t') = \sum_{a\in O}\sum_{b\in O} R(t,a)N(a,b)R(t',b) \quad (6)$$

Another relation $s(t,t')$ is a normalized measure derived from $p(t,t')$; $s(t,t')$ is defined as follows:

$$s(t,t') = \frac{p(t,t')}{\sqrt{p(t,t)p(t',t')}} \quad (7)$$

It can be proved that similarity $p(t,t')$ using (5) becomes a positive-definite kernel function [4]. More generally, when function f is convex, $p(t,t')$ is positive-definite. See [2] or [3] for the proof. It is not difficult to see that $s(t,t')$ is also positive-definite.

B. *Kernel fuzzy c-means*

The method of kernel c-means can be applied to data with clusters having nonlinear boundaries; without a kernel, only linear boundaries can be obtained by c-means [1]. In this study, the similarity calculated by fuzzy neighborhood is a kernel function.

Clustering using kernel function uses a high-dimensional feature space, which is the main difference from ordinary fuzzy c-means [1].

The objective function of kernel-fuzzy c-means (K-FCM) is as follows:

$$J_{kfcm}(U,W) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m \|\emptyset(x_k) - W_i\|^2 \quad (8)$$

where W_i is the center of cluster i in a high-dimensional feature space, and can be obtained by minimizing (8) with respect to W_i:

$$W_i = \frac{1}{N_i}\sum_{k=1}^{n}(u_{ik})^m \emptyset(x_k). \quad (9)$$

Note that $\emptyset(x_k)$ is called a high-dimensional mapping. The functional form of $\emptyset(x_k)$ is generally unknown but its inner product is known and given by a kernel function $K(x_j,x_k) = (\emptyset(x_j),\emptyset(x_k))$. Note that either $p(t,t')$ or $s(t,t')$ can be used as the kernel function.

We cannot calculate (9) directly in c-means clustering. Instead, distance d_{ik} is calculated by using the kernel function $K(x_j,x_k)$ as follows:

$$d_{ik} = \|\emptyset(x_k) - W_i\|^2 = \left\|\emptyset(x_k) - \frac{1}{N_i}\sum_{j=1}^{n}(u_{ij})^m \emptyset(x_j)\right\|^2$$
$$= K(x_k,x_k) - \frac{2}{N_i}\sum_{j=1}^{n}(u_{ij})^m K(x_j,x_k)$$
$$+ \frac{1}{(N_i)^2}\sum_{j=1}^{n}\sum_{l=1}^{n}(u_{ij})^m(u_{jl})^m K(x_j,x_l) \quad (10)$$

Fuzzy membership value u_{ik} is given as follows:

$$u_{ik} = \left[\sum_{p=1}^{c}\left(\frac{d_{ik}}{d_{pk}}\right)^{\frac{1}{m-1}}\right]^{-1} \quad (11)$$

The algorithm of Kernel Fuzzy c-Means (K-FCM) is as follows:

K-FCM1. Give an initial value of membership degree $\bar{U}$;

K-FCM2. Calculate the distance d_{ik} of each individual to each cluster center in a high-dimensional space;

K-FCM3. Solve the minimization problem $min_{U\in M_f} J_{kfcm}$ (U, W) and make $\bar{U}$ as an optimal solution;

K-FCM4. If the solution is convergent, stop. Else go to **K-FCM2**.

End of K-FCM.

The hard c-means (K-HCM) has also been well-known [1]. We omit the detail of K-HCM algorithm.

C. *Kernel Affinity Propagation*

A remarkable point of affinity propagation [5] [6] is that the number of clusters is defined by the method itself. In this method, all the data points exchange messages between each other to decide which cluster to subordinate and which point to be the exemplar.

There are two messages exchanged: responsibility $r(i,k)$ and availability $a(i,k)$. Data point i sends a message of responsibility $r(i,k)$ to a candidate exemplar point k to show how point k is suited to serve as the exemplar for point i. The candidate exemplar point k sends the message of availability $a(i,k)$ to point i to reflect the accumulated evidence of the degree of appropriateness for point i so that point k is chosen as its exemplar [5] [6].

Then, responsibility $r(i,k)$ is defined by the following rule:

$$r(i,k) \leftarrow s(i,k) - \max_{k'\neq k}\{a(i,k') + s(i,k')\}. \quad (12)$$

As the above responsibility is updated by (12), availability is updated by the following rule:

$$a(i,k) \leftarrow min\ \{0, r(k,k) + \sum_{i'\notin\{i,k\}} max\{0, r(i',k)\}\}. \quad (13)$$

But availability of point to itself is different:

$$a(i,k) \leftarrow \sum_{i' \notin k} max\{0, r(i',k)\} \quad (14)$$

The similarity $s(i,k)$ between two data points is defined using the kernel function: For point x_i and x_k, we have

$$s(i,k) = -\big(K(x_i,x_i) + K(x_k,x_k) - 2K(x_i,x_k)\big). \quad (15)$$

For data point i, the point k that maximizes $a(i,k) + r(i,k)$ is chosen to be the exemplar of data point i, in other words, data point i belongs to cluster k.

D. Rand Index

The Rand index [7] is a measure of similarity between two data clustering results. The definition is as follows.

Given N points $S = \{X_1, \ldots, X_N\}$, and two clusters of them $Y = \{Y_1, \ldots, Y_s\}$ and $Y' = \{Y'_1, \ldots, Y'_k\}$, we define:

- m: the number of pairs of elements in S that are in the same set in Y and Y',
- m': the number of pairs of elements in S that are in different sets in Y and Y'.

Then the Rand index R is:

$$R = \frac{m+m'}{\binom{N}{2}} \quad (16)$$

Assume that we use K-FCM. The result is fuzzy clusters. In order to use the Rand index, we make clusters hard by using the maximum membership rule: *allocate an object to the cluster where the membership takes its maximum value*.

When we perform M times of **K-FCM** algorithm with different initial values, we have M results. Then, we use the Rand index to calculate each pair of resulting clusters: the combination of all the pairs of results is $\binom{M}{2}$. Then the average of the Rand index value is calculated. It will show the degree of dependence of clusters on the initial values: when we have a larger value of the Rand index, it means less dependency on the initial values. When we compare the averaged Rand indexes for different methods, we can judge which method has the least dependency on the initial values, in other words, robustness with respect to initial values.

In this paper, we will show two Rand index values of two methods of K-HCM and K-FCM.

III. Application to a Document in Chinese

A document < Company Culture Management > on the Web site [8] introduces the importance of company culture and how to management the company culture. The data cannot be easily clustered by human eyes. Nouns and adjectives of which the numbers of occurrences were more than 2 were extracted; the number of objects for clustering was 36.

Figures 2 and 3 show clusters using kernel fuzzy c-means and kernel affinity propagation for this document. The kernel principal component [4] was used for the visualization. The number of clusters for kernel c-means was assumed to be three, while the number of clusters was automatically decided in the affinity propagation. The number of clusters obtained by the latter method can be varied according to the maximum number of iterations, but generally the number is far larger than three.

Investigating the contents of this document, we observe that the meanings of the clusters are as follows:

- Cluster 1 (blue): the role of corporate culture and its development history;
- Cluster 1 (red): excellent cases of business or companies, such as United States and Japan, outstanding business success stories and economic outcomes or impact;
- Cluster 3 (green): the way we should build the socialist modern company culture in the market economy, and the importance of modern corporate culture and its role, or the regulatory requirements to business leaders, the masses of workers and others.

The details of the keywords in these clusters are omitted here, but the three clusters shown by different colors by fuzzy c-means correspond well with the three prolonged groups on the plane in Figure 2, while geometrical separation of clusters is not found in Figure 3. Note that the configuration of points in these figures was derived from the kernel principal component and is independent of clusters.

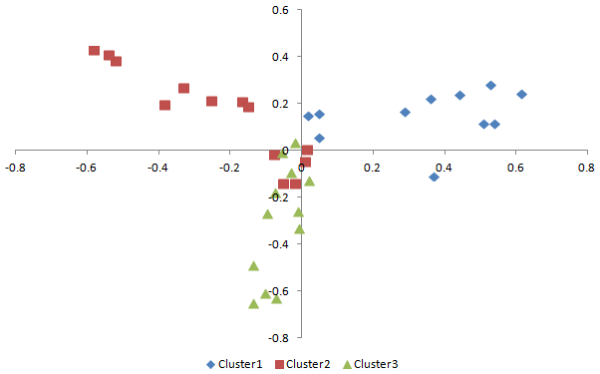

Figure 2. Clusters by kernel fuzzy c-means.

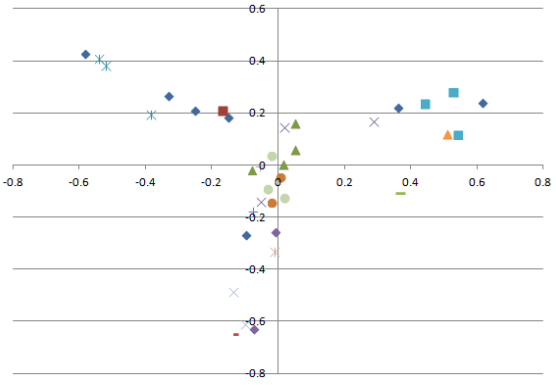

Figure 3. Clusters by kernel Affinity Propagation

TABLE I. RAND INDEX VALUES BY HARD AND FUZZY C-MEANS

	K-HCM	K-FCM
Rand index	0.716	0.830

Table 1 shows the values of the Rand index from K-FCM

and K-HCM. The value of the Rand index was higher when we used K-FCM than that of K-HCM, which means that the fuzzy method was more robust with respect to the difference of initial values than the hard method of c-means clustering.

IV. CONCLUSION

The model of fuzzy neighborhood was shown and the kernel based methods of clustering were developed. The developed methods were hard and fuzzy kernel c-means as well as kernel affinity propagation. The results of application to a Chinese document showed interpretable clusters by fuzzy c-means, while the result of affinity propagation did not show good clusters. Using the Rand index, we showed the method of fuzzy c-means gave more robust results than the hard c-means. To evaluate the combination of fuzzy neighborhood and kernel affinity propagation quantitatively, we need far more examples.

The document is a real Chinese document on the Internet. Although the volume is not enough now, further studies of Chinese documents or Social Networking Service data seem to be promising.

ACKNOWLEDGMENT

This study has partly been supported by the Grant-in-aid for Scientific Research, JSPS, Japan, No. 26330270.

REFERENCES

[1] S. Miyamoto, H. Ichihashi, K. Honda, Algorithms for Fuzzy Clustering, Springer, 2008.

[2] S. Miyamoto, S. Suzuki, “Clustering in Tweets Using a Fuzzy Neighborhood Model”, Proc. of WCCI 2012.

[3] S. Miyamoto, Y. Komazaki, S. Suzuki, S. Takumi, “Analysis of Disaster Information on Twitter Using Different Methods of Clustering Based on A Fuzzy Neighborhood Model”, Proc. of ISCIIA 2012.

[4] B. Scholkopf, A. J. Smola, Learning with Kernels, MIT Press, 2002.

[5] B. J. Frey and D. Dueck, “Clustering by Passing Messages Between Data Points”, Science, vol.315, pp.972-977, 2007.

[6] I. E. Givoni, B. J. Frey, “A Binary Variable Model for Affinity Propagation”, Neural Comput., vol.21, pp.1589-1600. doi: 10.1162/neco.2009.05-08-785, 2009.

[7] W. M. Rand, “Objective criteria for the evaluation of clustering methods”, Journal of the American Statistical Association, vol. 66, pp.846–850, 1971.

[8] http://wenku.baidu.com/view/e599acf2f242336c1fb95e36.html

Copyright (c) IARIA, 2015. ISBN: 978-1-61208-408-4