PROCEEDINGS OF SCIENCE

# Online Comment Clustering Based on Golden Section Method

**Ping Zhang**[1][2]
*College of Fundamental Education*
*Sichuan Normal University*
*Chengdu, 610068,China*
*E-mail:* 835148187@qq.com

**Jianzhong Wang**
*College of Fundamental Education*
*Sichuan Normal University*
*Chengdu, 610068, China*

Based on the existing study of online comment on the short text paradigm and the golden section method as introduced, an improved method that calculates the short text distance based on the Golden section method is hereby proposed. The new method considers the text distance as the combination between the semantic distance and the form distance on the basis of the importance weight and the preferable optimum experimental results are deprived with the new method based on the Golden section method. The simulation experiments indicate that the newly improved method of text distance calculation based on the Golden section method is better than the traditional methods with the clustering performance indeed improved.

*CENet2015*
*12-13 September 2015*
Shanghai, China

[1] Speaker
[2] In this paper, the research was sponsored by the Department of Education Project of Sichuan Province.

## 1. Introduction

With the popularity of mobile internet services and mobile phone as the internet terminal usage rate for improvement, the short text paradigm brings new challenges and opportunities for Chinese information processing. The network reviews based on short text usually begin with an event or topic, spread fast and have a wide impact, reflecting the public's attitude towards public event, expressing content which has strong subjectivity. It not only expresses the reviewer's own point of view, but also affects other participants. Researches on short text clustering comments can timely grasp the views and positions of various hot topics, which is of great significance for the state, the enterprises and the society.

In the study of network comment clustering, naive Bayesian model is used [1]. A feature extraction method [2] and a short text semantic similarity measurement method [3] have been proposed. Short text semantics is extracted by using LSA + ICA so as to avoid the short text [4]. The above method does not fully consider the particularity of short text processing, but deals with this short text by using the same method that was used to handle the traditional long text successfully or avoid text length problem. With the short text features in the study of clustering for network review and increasingly complex processing purpose of users, the method based on the comparison of text form and semantic similarity is again brought to the attention of the researchers. The method of text similarity is proposed by comparing the text containing words and word order so as to calculate the semantic similarity between short texts [5]. The method of text semantic similarity is proposed by using ontology and other semantic resources to compute the semantic similarity of short text [6-7]. It gives an improved semantic distance measure, an integrated representation of form distance and unit of semantic distance [8]. The paper [9] presents a method which improves the text similarity calculation by using LDA model and carrying out the text similarity matrix clustering experiments to assess the effect of clustering. In all, the combination between the form distance and the semantic distance in the above research work was less considered or its inner link was not considered by just simply adding the two methods; thus on the basis of the above study and in view of the points of network comment short text, both the form distance and semantic distance of short text are considered; and on the basis of the golden section method for studying the clustering, the network clustering method based on the golden section method is proposed.

## 2. Existing Text Distance Calculation

Text semantics is a complex problem involved in computer, artificial intelligence, psychology, cognitive science and many other disciplines; however, from the point of view of statistical linguistics, we can only calculate the text differences by using statistical characterization. In the existing research of text distance calculation, the form distance of texts is calculated based on the form comparison method [5], while the semantic distance of texts is calculated based on the text semantic method [6-7], the text distance expressing the semantic distance and form distance.

Suppose that $\Sigma$ is word list, $\Phi \in \Sigma$ is empty character, $\Sigma^*$ is the sentence collection that is made up of $\Sigma$, sentencesis $R,S \in \Sigma^*$ represented as follows: $R = r_1 r_2 \dots r_m$, $S = s_1 s_2 \dots s_n$. Sentence length is m and n, the $r_i$ $s_i$, says the word i in the sentence. Define the distance of short text $R,S$ as a combination of form distance $d1(R,S)$ and the semantic distance $d2(R,S)$. The formula is as follows [8]:

$$D(R,S)=d1(R,S)+d2(R,S) \tag{2.1}$$

The form distance $d1(R,S)$ is achieved by calculating operations in the form of alignment and the semantic distance $d2(R,S)$ calculation that is based on the knowledge of linguistics in Formula (2.1).

### 2.1 Role of Semantic Calculation Method in the Short Text Similarity Calculation is Better than That of Long Text

Based on the above assumptions in the previous section, suppose that this text is generated in a random manner according to the word code Σ, namely, any character $c_i \in \Sigma$ with probability $P_{c_i}$ appearing in the sentence and statistically independent characters appear, the probability of character $c_i$ in random text $R,S$ at the same time is

$$P(c_i)=\sum_{k=1}^{m} C_m^k P_{c_i}^k (1-P_{c_i})^{m-k} \sum_{r=1}^{n} C_n^r P_{c_i}^r (1-P_{c_i})^{n-r} \tag{2.2}$$

Suppose the semantic equivalence class of characters $c_i$ defined as

$$c_i+=\{c_j|c_j\in\Sigma, sim(c_i,c_j)\geqslant\varepsilon\} \tag{2.3}$$

Representation of semantically similar character sets with character $c_i$, $sim(.,.)$ says the semantic approximation measurement. Suppose that this text is generated according to the word code Σ in a random way, namely the equivalence class $c_i+$ of any character $c_i$ to appear as probability $Q_{c_i}$ in the sentence, then the probability of equivalence $c_i+$ appearing in the random text $R,S$ at the same time is

$$P(c_i+)=\sum_{k=1}^{m} C_m^k Q_{c_i}^k (1-Q_{c_i})^{m-k} \sum_{r=1}^{n} C_n^r Q_{c_i}^r (1-Q_{c_i})^{n-r} \tag{2.4}$$

As a result, $P(c_i)$ can be used to measure the similar probability of sentence $R,S$ without introducing semantic computation, $P(c_i)+$ can use metrics to introduce semantic computation, sentence $R,S$ similar probability. Apparently $Q_{c_i}>P_{c_i}$, $Q_{c_i}=P_{c_i}+\Delta$, then P(ci) and P(ci+) difference is $\delta$.

$$\delta=P(C_i+)-P(C_i)=[1-(1-P_{C_i}-\Delta)^m][1-(1-P_{C_i}-\Delta)^n]-[1-(1-P_{C_i})^m][1-(1-P_{C_i})^n] \tag{2.5}$$

First of all, we look at the relationship of the sentence difference δ generated by introducing semantic computing and Δ limit, derivative and elastic coefficient

$$\lim_{\Delta\to\infty}\delta=0 \tag{2.6}$$

$$\frac{\partial\delta}{\partial\Delta}=2\mathrm{m}[1-(1-p-\Delta)^m](1-p-\Delta)^{m-1}>0 \tag{2.7}$$

$$\rho=\frac{\partial\delta}{\partial\Delta}\cdot\frac{\Delta}{\delta}\approx1>0 \tag{2.8}$$

Here, in order to simplify the analysis, supposing that m = n, then $n!\approx\sqrt{2\pi n}(\frac{n}{e})^n$. From Formula (2.2) ~ Formula (2.8), δ refers to the monotone increasing function of Δ, and the elastic coefficient $\rho>0$ indicates positive correlation exists between the δ and Δ. Usually, Δ is small in the actual language. If Δ is bigger, it shows that language itself is an ambiguity; secondly, when the sentence length m and n are bigger, that is, when $R,S$ are long texts.

$$\lim_{m,n\to\infty}\delta=0 \tag{2.9}$$

$$\frac{\partial\delta}{\partial m}=2\log(1-p-\Delta)[(1-p-\Delta)^m-1](1-p-\Delta)^m-2\log(1-p)[(1-p)^m-1](1-p)^m \tag{2.10}$$

Formula (2.9) and (2.10) show that although there is no monotonous relationship between the sentence length m and n and the semantic difference δ, whatever $P_{c_i}$ and Δ value, with the increase of text length m and n, $\delta \to 0$ As to the long text with the increase of the length of the text, the difference by the semantic calculation diminishes gradually in text length. Especially for the actual language and Δ which are relatively smaller, the trend is more obvious. Again, when the length difference between sentence m and n is bigger, namely, when in $R,S$ one is a long text, the other for short text, the semantic differences between the texts are dominated by longer text, $\delta \to 0$ namely: $\lim_{m\to\infty}\delta=0$ , $\lim_{n\to\infty}\delta=0$ ; but only if $R,S$ is for short text, the difference δ is dominated by Δ. In this short text similarity calculation, a semantic measure effect is more significant.

In conclusion, the role of semantic calculation method in the short text similarity calculation is better than that of long text. Semantic computing affected by the length of the text, especially when the length of the text is not the same, the long text plays a more significant role. In order to alleviate the calculation error caused by the difference in length of the short text for passage of the comments of the network analysis, the essay distance calculation method based on a distinct word count punishment [8] is used. It uses the original text length to punish distance and hope eliminate the influence caused by different sentence lengths. This method is to demonstrate the effectiveness of the amazing in the practical application. It is used in various fields [10].

## 3. Golden Section Method

0.618 method is also called the golden section method that it is designed according to the principle of golden section [11]. It is the classical algorithm of the optimum seeking method and known as the simple algorithm of significant effect. It is the foundation of many optimization algorithms. During the optimization trying point is put on the golden points to find the optimal choice. 0.618 method was put forward by American mathematician Jack Kiefer in 1953 and China's famous mathematician Luogeng Hua in the 1960s and 1970s, complemented and carried on the promotion in our country. Nowadays, it has been widely applied in various fields. The optimum seeking method is a method of optimization problem. If the experimental point is taken in the interval of 0.618 and the number of experiments will be greatly reduced. The interval of 0.618 places as a testing method is the one dimensional optimum seeking method, also called the 0.618 method. Practice has proved that to a factor of the problem, the use of "0.618" in 16 trials can be completed by the "Bisection method" to carry out 2500 experiments. The basic idea of the golden section method is in reference to the "go rid of bad and keep good" principle, the principle of symmetry and constriction of the geometric principle to gradually narrow the search.

## 4. Improved Text Distance Calculation by Introducing Golden Section Method

According to the existing text distance calculation and the short text distance calculation method based on the word length punishment, for passage analysis of the network comments, the text similarity and semantic text unit should not be ignored. According to the linguistic knowledge, the importance of semantic distance and form distance may be the same or different in the clustering analysis. When it is different in the roles of semantic distance and form distance, which one is more important, the importance weights can be determined by the clustering results through experiments. Thus in this paper, the golden section method is introduced to measure different importances of semantic distance and form distance, gradually better clustering results as optimized.

### 4.1 Improved Formula , Calculation Process and Algorithm

The distance of short text $R,S$ is defined as the synthesis of the form distance $d1(R,S)$ and the semantic distance $d2(R,S)$:

$$D(R,S)=w1*d1(R,S)+w2*d2(R,S) \tag{4.1}$$

Among them, according to the importance of form distance $d1(R,S)$ and the importance of the semantic distance $d2(R,S)$, the initial conditions are respectively set up, such as *w1 > w2*, *w1<w2*. According to the golden section method, the value of the *w1* and *w2* is preliminarily determined. After many tests according to the golden section method the *w2, w1* value is gradually adjusted, finally a better cluster analysis result in respect of network comment can be achieved.

The text distance calculation process is shown in Fig. 2.

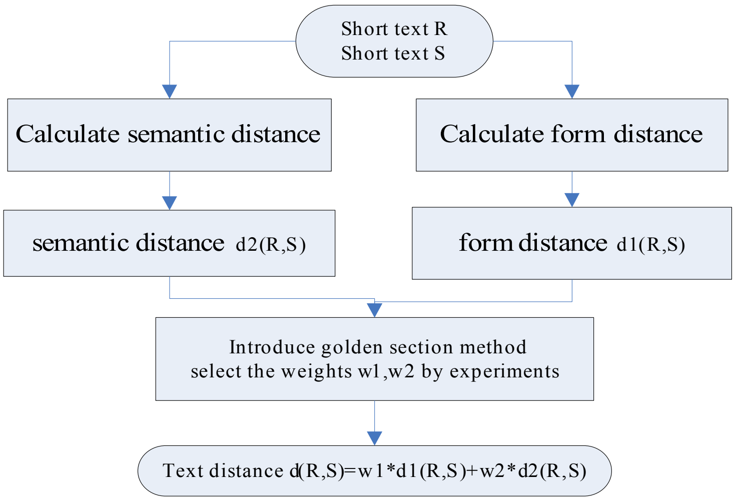
Short text R
Short text S
Calculate semantic distance
Calculate form distance
semantic distance d2(R,S)
form distance d1(R,S)
Introduce golden section method
select the weights w1,w2 by experiments
Text distance d(R,S)=w1*d1(R,S)+w2*d2(R,S)

**Figure 1:** Text Distance Calculation Process

Text distance calculation algorithm is shown as follows:

Input short text $R=r_1r_2...r_m$, $S=s_1s_2...s_n$ .

The form distance $d1(R,S)$ is achieved by calculating operations in the form of alignment. The method is that match the semantic alignment of arbitrary words in the two short texts one-to-one without repeatition according to the words largest similarity, adjust the phrase word order and make the phrase structure and form of long sentences in the maximum form similarity. Similarity between words is calculated by using the synonym word Lin extended edition [12-13]. Assume $r_i,s_j$ either word meaning for $M(r_i)$ and $M(s_j)$ in the synonyms Lin extended edition collection. The degree of similarity between $a\in M(r_i)$ and $b\in M(s_j)$ is defined as $Sim(a,b)=n1(N+1)$, n for the compilation of code beginning with different levels, N for coding digits. So the similarity degree of the two words $r_i,s_j$ is defined as $sim(r_i,s_j)=max_{a\in M(r_i),b\in M(s_j)}sim(a,b)$ .

The semantic distance $d2(R,S)$ calculation is based on the knowledge of linguistics. The sentence meanings are scientifically divided into expression of the main meaning unit and unit of the notional seasoning. In order to distinguish contribution, a solid righteousness unit and a non-solid righteousness unit are given with different weights. All nouns, pronouns, verbs, adjectives are of a solid righteousness unit at the Chinese short text [8], other parts words such as numerals and quantifiers, adverbs etc. are of non-solid righteousness unit while using the edition distance calculate on a word. Three editing operations: insertion, deletion and replacement and different editing operations are assigned to different weights whether it is solid righteousness unit or semantic similarity. As to the specific calculation formula, see $d_2=\omega_1*a_1+\omega_2*a_2+\gamma_1*b_1+\gamma_2*b_2+\theta*c$, in which, $\omega_1/\omega_2$ is to insert or delete solid righteousness or non-solid righteousness unit operating weights, $\gamma_1/\gamma_2$ is to replace real righteousness or non-solid righteousness unit operating weights, θ is the weights of synonyms for replace, $a_1/a_2$, $b_1/b_2$, $c$ correspond to the number of editing operation, $d2$ to word as a unit to transform a sentence for another sentence the required minimum number of edit operation. Different weights express different understandings of the semantics. The experiment of weight value is set according to the following principles. The semantic which is more important than the form indicate the meanings of alignment cost being less than the cost of insert or delete, $\varphi_1>\lambda\ \omega_2>\lambda$. Semantic replacing outweigh is greater than the increase or decrease of semantics, $\gamma_1>\omega_1,\gamma_2>\omega_2$. Solid righteousness unit operation is greater than the solid righteousness unit, $\omega_1>\omega_2,\gamma_1>\gamma_2$. Synonyms for operating cost is less than the price of the synonyms, $\lambda>\theta,\omega_1>\theta$, $\omega_1>\theta,\omega_2>\theta,\gamma_1>\theta,\gamma_2>\theta$. Form distance, the unit of semantic distance weighting normalization processing, $\lambda+\omega_1+\omega_2+\gamma_1+\gamma_2+\theta=1$. It refers to the number of operations that is normalized by using the short text length. Suppose that d is operation times, $|R|$ , $|S|$ respectively text length of $R,S$ , the normalized number of operations d is defined as. $\bar{d}=\frac{2\quad d}{|R|+|S|}$. The contribution of the unit of semantic information is greater than the information structure; therefore, the semantic operation weight of unit is greater than form structure in the experiment. The weights are set in combination with the weight selection principle and the experience value such as: $\lambda=0.04,\omega_1=0.28,\omega_2=0.05,\gamma_1=0.52,\gamma_2=0.09,\theta=0.02$.

Based on the normalization of *d1* and *d2*, the initial $w1,w2$ are set up with the introduction of the golden section method, the short text $D(R,S)$ is defined as a combination of the product of form distance d1 and the weight w1 and the product of semantic distance d2 and the weight w2. Upon many experimental optimizations, the more optimal value of w1 and w2 is ultimately determined.

Output short text $R,S$ distance $D(R,S)$.

### 4.2 Complexity of the Analysis

Sentence $R,S\in\Sigma^*$ , $|R|,|S|$ expresses the length of sentence, so the time computational complexity in the classical edit distance is $O(|R|*|S|)$, namely, the $O(n^2)$[14]. The improved text distance algorithm of time computational cost consists of two parts in this paper. It is A of form distance calculation cost and B of unit semantic distance calculation cost. A says alignment operands between sentence R and S according to the semantic similarity of all words. The time computational complexity of which can be expressed as$O(|R|*|S|*T)$, namely the$O(Tn^2)$. T is the calculated time complexity of two words semantic degrees based on the synonym words books. B says the time complexity based on the sentence edited after the alignment at differentiate solid righteousness unit and non-solid righteousness unit according to different weights assigned with the different editing operation semantics, and thus it can be approximated as $O(|R|*|S|)$, namely the $O(n^2)$. Above all, the time complexity of text distance to improve the golden section method is$O(Tn^2)+O(n^2)$; but it is important to note that, the improved algorithm is mainly applied to the short text. When n is relatively small, the overall cost of time can meet the requirement of practical applications.

## 5. Experiment Stuying

The experiment of this paper used 863 text classification data sets and online reviews data sets.

In the simulation experiments of this paper, another short text data set as adopted is a subset of the AK and ADK data sets in 863 text classification evaluation data set, the same short text data set as that [8]. The data classification represents maximum, Leninism, Mao Ze-dong thought, Deng Xiao-ping theory, D classification represents politics, law, and K classification represents history, geography. Randomly part of the experimental samples was selected, of which 100 text were randomly selected in the two kinds of text A, K, totally 200 making up AK data set. Respectively 100 text in these three kinds of text A, D, K were randomly selected, a total of 300 making up ADK data set. Affinity Propagation algorithm [15] using short text clustering model had a clustering analysis of the text. The entropy clustering performance evaluation index E clustering assessment [16] was used. The smaller the entropy value is, the better the performance of clustering will be. Three kinds of distance calculation models in clustering performance of the group 2 test set are used in this experiment. The experimental results are shown in Table 1.

The ISD considers the form structure and the semantic differences between sentences by adjusting the weights to adjusting the size of the form structure and semantic unit weight in text semantic content. ISD-I set the information contribution of unit semantic greater than form structure information, namely the right of the operation of the unit semantic value is greater than the operating right of value of the form structure, and its weight is as follows: $\lambda = 0.04,\ \omega_1 = 0.28, \omega_2 = 0.05, \gamma_1 = 0.52, \gamma_2 = 0.09, \theta = 0.02$. This paper determines the initial value of w1 and w2, for example, w1=0.382*2, w2=0.618*2, through the contrast experiment, it can be concluded that the initial value w1<w2, then a lot of better results attained w1=0.3244*2, w2=0.6756*2 upon optimization. The experiment result shows that comparing text distance based on the improved golden section method put forward in this article with the text distance [8], a certain advantage is shown and the clustering performance is improved.

| Data set | The method of literature [8] | | The method of this paper |
|---|---|---|---|
| | ISD-1 | ISD-2 | |
| AK<br>average the lengths of sentences=6.22 | 0.614931±0.0431233 | 0.621599±0.043741 | 0.592327±0.0393526 |
| ADK<br>average the lengths of sentences=7.39 | 0.560775±0.028748 | 0.559006±0.031459 | 0.534583±0.0215492 |

**Table 1:** Experimental Results Are Entropy Clustering Performance

The online reviews of the clustering system consist of extraction module, data processing module, text distance calculation module and cluster module. The online reviews extraction module uses two ways that push and pull to get online reviews from the Internet. The data preprocessing module through the Unicode data on online reviews pretreatment and eliminate the influence of different character encodings of semantic distance calculation. Text distance calculation module is used to calculate the text after normalizing the distance and gets the text distance matrix through comprehensive analysis of form distance and semantic distance between text units. The online comment cluster module using clustering algorithm to aggregate analysis of the text distance matrix gets the theme of different clusters. Finally, the online comments are presented to the user in accordance with the theme of cluster. The online review data set obtains the online short text comment data by online reviews extracted module. The news comment information of one day is extracted on Sina micro-blog (http://weibo.com). Two classes of the 240 text were randomly selected in the information as the experimental corpus. The cluster theme content in descending order includes shopping, entertainment, travel, health, children learning, sports, holidays and other. Upon analysis of corpus experimental characteristics which can be seen in the experimental corpus set covering most of the network users of the popular comment information, to a certain extent, this method may be applicable to the real network application scenarios.

## 6. Conclusion

The network clustering methods based on golden section method is proposed under the condition of no change in time complexity. It is superior to the traditional method and the clustering performance has been improved to a certain extent. The future research direction is to further improve the algorithm with the feature extraction method, apply the algorithm to the network real scene, optimize the clustering analysis and serve the society.

## References

[1] K. Nishida,T. Hoshide, K. Fujimura. *Improving tweet stream classification by detecting changes in word probability*. In: Proc. of the ACM SIGIR 2012 (Proceeding SIGIR '12 Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval.) ACM New York, NY, USA ©2012 . 971−980. [doi: 10.1145/2348283.2348412]

[2] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu , M. Demirbas , *Short text classification in twitter to improve information filtering*.In: Proc. of the ACM SIGIR 2010( Proceeding SIGIR '10 Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval.) ACM New York, NY, USA ©2010. 841−842. [doi: 10.1145/1835449.1835643]

[3] R. Mihalcea, C. Courtney, C. Strapparava. *Corpus-Based and knowledge-based measures of text semantic similarity*. In: Proc. of the AAAI 2006(Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference), July 16-20, 2006, Boston, Massachusetts, USA . 775−780.

[4] Q. Pu, G. W Yang. *Short-Text classification based on ICA and LSA*. In: Proc. of the ISNN 2006. LNCS 3972, Heidelberg: Springer-Verlag, Chengdu, China. 2006. 265−270. [doi: 10.1007/11760023_39]

[5] N. Chatterjee. *A statistical approach for similarity measurement between sentences for EBMT*. In: Proc. of Symp. on Translation Support Systems.STRANS-2001. Kanpur: Indian Institute of Technology, 2001.

[6] Y. P. Liu, Li S, T. J. Zhao, *System combination based on wsd using WordNet*. Acta Automatica Sinica, 2010,36(11):1575−1580 (in Chinese with English abstract). [doi: 10.3724/SP.J.1004.2010.01575]

[7] Z. Yang, K. F. Fan, J. J. Lei, Guo J. *Text manifold based on semantic analysis*. Acta Electronica Sinica, 2009,37(3):557−561 (in Chinese with English abstract). [doi: 10.3321/j.issn:0372-2112.2009.03.024]

[8] Z. Yang, L. T. Wang, Y. X. Lai, *Online comment clustering based on an improved semantic distance*. Ruan Jian Xue Bao/Journal of Software, 2014,25(12):2777−2789 (in Chinese). http://www.jos.org.cn/1000-9825/4729.htm

[9] Z. Z. Wang , M. He ,:DU YP. *Text Similarity Computing Based on Topic Model LDA*, Computer Science ,Dec 2013 Vo1.40 No.12 229-232

[10] Y. J. Li , B. Liu, *A normalized levenshtein distance metric*. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2007,29(6):1091−1095. [doi: 10.1109/TPAMI.2007.1078]

[11] http://baike.baidu.com/link?url=RxxTAYB15Ymn5Elfo2osW8Ji1vORODm0z5Pep4BYpensu3Y5Cmjk89-ldk-Ubv3fyqRd9qnGESh_FZ7F6r-J7K baidu baike. 0.618method .2015

[12] W. X. Che, Z. H.Li , T. Liu, *LTP: A Chinese language technology platform*. In: Proc. of the Coling 2010( Proceeding COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations Association for Computational Linguistics Stroudsburg), PA, USA ©2010 13−16.

[13] S. J. Li, *Research of relevancy between sentences based on semantic computation*. Computer Engineering and Applications, 2002,38(7):75−76 (in Chinese with English abstract). [doi: 10.3321/j.issn:1002-8331.2002.07.025]

[14] E. S. Ristud , P. N. Yianilos , *Learning string-edit distance*. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1998,20(5):522−532. [doi: 10.1109/34.682181]

[15] B. J.Frey , D. Dueck, *Clustering by passing messages between data points*. Science, 2007,315(5814):972−976. [doi: 10.1126/science.1136800]

[16] Y. Zhao, G. Karypis, *Empirical and theoretical comparisons of selected criterion functions for document clustering*. Machine Learning, 2004,55(3):311−331. [doi: 10.1023/B:MACH.0000027785.44527.d6]