A feature representation method for biomedical scientific data based on composite text description[①]

SUN Wei[1,2]

[1] National Science Library, Chinese Academy of Sciences, Beijing 100190, China
[2] Graduate School, Chinese Academy of Sciences, Beijing 100049, China

Received Sep. 3, 2009
Revised Oct. 28, 2009
Accepted Nov. 3, 2009
Translated with a permission from *New Technology of Library and Information Service* (in Chinese), 2009(5):22

Abstract Feature representation is one of the key issues in data clustering. The existing feature representation of scientific data is not sufficient, which to some extent affects the result of scientific data clustering. Therefore, the paper proposes a concept of composite text description (CTD) and a CTD-based feature representation method for biomedical scientific data. The method mainly uses different feature weight algorisms to represent candidate features based on two types of data sources respectively, combines and finally strengthens the two feature sets. Experiments show that comparing with traditional methods, the feature representation method is more effective than traditional methods and can significantly improve the performance of biomedcial data clustering.

Keywords Composite text description, Scientific data, Feature representation, Weight algorism

1 Introduction

Data clustering is to divide some unknown distributed data sets into several groups or categories, making data in the same category have the highest similarity and data between different categories have the lowest similarity[1]. Despite the increasing volume of scientific research performed each year and the resultant scientific data increasing at an exponential rate, the scale of scientific data sharing and standardization is only expanding gradually. As a result, scientific data clustering technology is becoming more and more important, particularly in relation to such applications as scientific data analysis, scientific data retrieval, and industry trend forecasting.

Feature representation of data is a prerequisite for data clustering. The current feature representation methods of scientific data are divided into scientific data sources and scientific literature source. The feature representation method based on

[①] This project is supported by the Agridata, the sub-program of National Science and Technology Infrastructure Program (Grant No. 2005DKA31800).
Correspondence could be addressed to Sun Wei (E-mail: sunwei@mail.las.ac.cn).

CJLIS
Vol. 2 No. 4, December 2009
pp 43–53
National Science Library, Chinese Academy of Sciences

scientific data sources is mostly using the scientific database to find the model and structure of the data or make a hierarchical description of data sets to describe data objects.[2] This method focuses on the essential features of data but lacks detailed description of their peripheral features. The feature representation method based on scientific literature sources mainly uses the knowledge in the literatures to describe data[3–4]. The data description, only based on the relations between data and literature or between literatures, focuses on the peripheral features of data but takes no consideration into essential features of the data according to the relations between data. Obviously, neither of the two methods gives sufficient represention of the data features, thus weakening the clustering effects to some extent.

To solve the above problems, the paper proposes a feature representation method for biomedical scientific data based on composite text description. The method is built on the two data resources—scientific data source and literature source, and combines two feature weight algorisms—tf~idf and Zscore. Not only does it express the essential features of scientific data, but also adds more descriptions of their peripheral features. Comparing with traditional feature representation methods, the new method performs much better in relation to feature representation. With this method, the clustering effect of biomedical scientific data is also improved remarkably.

2 Composite text description

Scientific data are usually large in quantity, uneven, irregular and high-dimentional[5]. As text can describe any type of scientific data in a unified format, the paper proposes a concept of composite text description (CTD). CTD means the text sets which come from two data sources (scientific data source and literature source) and can highly summarize both essential and peripheral features of the same piece or the same type of data object.

There are close relations between data and between literatures. As the feature representation method for the biomedical scientific data proposed in the paper is based on the text description and relations between data and between data and literature (semantic relations), the method is limited to biomedical scientific data.

The CTD concept is as shown in Fig. 1, in which a piece of scientific data corresponds to a group of related literatures (RT_1,RT_2,...,RT_n) and a group of related data (RD_1,RD_2,…,RD_n)[6]. Each piece of literature RT_i in the literature database has a text t_i (eg. title, abstract) which can highly summarize its core content; each piece of scientific data in the scientific database has a text-based content description d_i (eg. text attribute field), which can highly summarize the essentianl features of the data (eg. functions and performance). Therefore, the description set "T" in Fig. 1, based on the relations between literatures and between literature and data, can be

National Science Library, Chinese Academy of Sciences

used to represent the peripheral features of the data “D” (called the text description “T”). Similarly, the description set “D” in Fig. 1, based on the relations between data, can be used to represent the essential features of the data “D” (called the text description “D”). The two text descriptions are called as a whole the composite text description (CTD) (see the shadow part in Fig. 1).

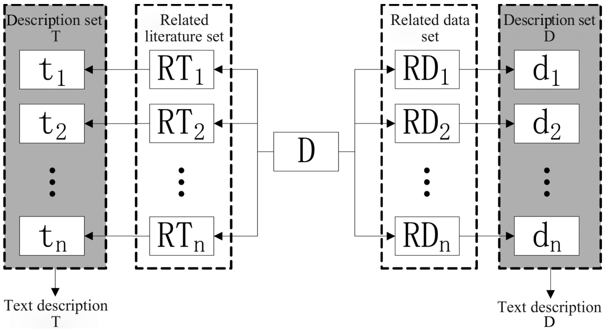

Fig. 1 Diagram of composite text description

3 Feature weight algorism of scientific data

The feature representation method proposed in the paper uses a vector space model, that is, creating feature vectors with special terms (including non-login terms) and their weights. First of all, to segment the terms in the composite text descriptions and filter out stop words; then, calculate the weight of each term in the pre-processed text, and take the top “K” terms with the highest weight as feature terms. The common traditional weight algorism include: Boolean, Term Frequency, tf ~ idf and so on. The paper uses the combination of two methods—tf ~ idf and Zscore[7]. The feature representation method based on text description “T” adopts Zscore and broadens the range of its variables as it is necessary. Specific formulas are shown as follows:

$$Z_d^a = \frac{F_d^a - \overline{F^a}}{\sigma^a} \tag{1}$$

$$F_d^a = \frac{W_d^a}{S_d} \tag{2}$$

$$\overline{F^a} = \frac{\sum_{i=1}^{N} F_i^a}{n} \tag{3}$$

$$\sigma^a = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(F_i^a - \overline{F^a}\right)^2} \tag{4}$$

In Formula 2, W_d^a is the number of literatures containing term "*a*" in the relevant literature set of scientific data "D", and S_d is the total number of relevant literatures of scientific data "D". So, F_d^a is the frequency of term "*a*" in all special terms representing scientific data "D", and is also called the frequency "*a*" of scientific data "D". $\overline{F^a}$ is the mean of F_d^a in the scientific data training set, and σ_a is the variance between F_d^a and its mean. Zscore method is chosen because text description "T", based on the relations between data and between data and literature, describes the data with peripheral relations of data and lays more emphasis on peripheral features of scientific data. Moreover, from the above formula we can see that, the Zscore weight algorism takes into account the proportion of the frequency of special terms to the frequency of its background terms for both a single piece of data and the whole data set so that it can describe the peripheral relations in a more vivid and comprehensive manner.

In addition, tf~idf weight algorism is used in the feature representation based on text description "D", as shown in the following formula:

$$W\left(t,\vec{d}\right)=tf\left(t,\vec{d}\right)\times lg\left(\frac{N}{n_t}+0.01\right) \tag{5}$$

In Formula 5, $W\left(t,\vec{d}\right)$ is the weight of term "t" in the scientific data "D", $tf\left(t,\vec{d}\right)$ is the frequency of term "t" in the text description set of scientific data "D", N is the total number of data in the scientific data training set, and n_t is the number of scientific data whose description sets "D" contain term "t". As a result, the feature vector of every piece of scientific data is shown as follows:

$$V\left(d\right)=\left(W\left((t_1,d),W\left(t_2,d\right),\ldots,W\left(t_n,d\right)\right)\right) \tag{6}$$

The tf~idf weight algorism is used here because text description "D", different from text description "T", is only built on the relations between data, and text description "D" is to describe the essential features of data. Therefore, tf~idf weight algorism is relatively easier and can calculate the feature term set of scientific data by taking into account the proportion of the frequency of special terms to the frequency of background terms for the whole data training set alone.

4 CTD-based feature representation method for biomedical scientific data

As mentioned in the *Introduction*, the current feature representation method for biomedical scientific data did not fully represent the essential and peripheral features of scientific data, which weakened the clustering effect of scientific data to a certain extent. The CTD-based feature representation method proposed in the paper is based on text descriptions “T” and “D”, combines tf~idf and Zscore—the two feature weight algorisms, converges the feature term sets figured out by the two algorisms. With this method, the feature representation effect is being improved to a large scale, so is the clustering effect of scientific data.

4.1 Pre-processing

In this step, terms are segmented and stop words are removed for all text descriptions “D” and “T” of scientific data training set. As each and every piece of scientific data belongs to a certain specialty, the clustering effects of data can be enhanced if scientific data are represented by different special terms. Therefore, the segmentation method here combines the N-gram algorism and the greatest matching algorism. With this method, special terms and non-login terms can be properly segmented.

4.2 Extraction of candidate features

Extraction of candidate features is designed to screen out terms which could be served as features from the original feature space (stop words already removed). The main purpose of the step is to remove the noise.

Either tf ~ idf or Zscore weight algorism relies on low-frequency terms to some degree, which is bound to reduce the effect of feature representation, so it is necessary to remove low-frequency terms in the original feature space. Based on this, the candidata features are respectively extracted from the tf ~ idf and Zscore weight algorisms, that is to say, all special terms in the original feature space are removed when $\overline{F^a} = 0$. The feature vectors $V_d(d_i)$ and $V_t(d_i)$ of data d_i based on the text descriptions “D” and “T” are expressed as follows:

$$V_d\left(d_i\right)=\left(W\left(t_1,d_i\right),W\left(t_2,d_i\right),W\left(t_3,d_i\right),...,W\left(t_n,d_i\right)\right) \tag{7}$$

$$V_t\left(d_i\right)=\left(Z_{d_i}^{a_1},Z_{d_i}^{a_2},Z_{d_i}^{a_3},...,Z_{d_i}^{a_n}\right) \tag{8}$$

4.3 Combination of features

As tf ~ idf differs from Zscore in the weight value fields of feature terms, it is necessary to normalize the weight values of the two kinds of feature representation vectors $V_d(d)$ and $V_t(d)$ respectively if combining the two features. The formula for normalization is as follows:

$$U_p=\left(p-\min\left(p\right)\right)/\left(\max\left(p\right)-\min\left(p\right)\right) \tag{9}$$

In formula 9, p and U_p are values before and after the normalization, while Max (p) and Min (p) are the maximum and the minimum values of the sample. For this, the feature vector $V_d(d_i)$ and $V_t(d_i)$ of data d_i based on text descriptions “D” and “T” are defined as follows respectively:

$$V_d\left(d_i\right)=\left(U_{W\left(t_1,d_i\right)},U_{W\left(t_2,d_i\right)},U_{W\left(t_3,d_i\right)},\ldots,U_{W\left(t_n,d_i\right)}\right) \tag{10}$$

$$V_t\left(d_i\right)=U_{Z_{d_i}^{a_1}},U_{Z_{d_i}^{a_2}},U_{Z_{d_i}^{a_3}},\ldots,U_{Z_{d_i}^{a_n}} \tag{11}$$

Secondly, it is required to combine the weights of the same feature vectors (the same special terms) representing the same data in the two feature sets: to assign weight “a” to terms in the feature representation method based on the essential feature description (text description “D”), and weight $(1-a)$ to that based on the peripharal feature description (text description “T”), in which $a=60\%$. The formula of the combination is as follows:

$$U\left(d_i\right)=U_{W\left(t_k,d_i\right)}\times a+U_{Z_{d_i}^{a_l}}\times\left(1-a\right)\quad\left(t_k=a_l\right) \tag{12}$$

The feature vector of every piece of data after normalization and combination is expressed as follows:

$$V\left(d_i\right)=\left(U_{W\left(t_1,d_i\right)},U_{W\left(t_2,d_i\right)},\ldots,U_{W\left(t_k,d_i\right)},U_1\left(d_i\right),U_2\left(d_i\right),\ldots,U_l\left(d_i\right),U_{Z_{d_i}^{a_1}},U_{Z_{d_i}^{a_2}},\ldots,U_{Z_{d_i}^{a_m}}\right) \tag{13}$$

4.4 Filtering of candidate features

To cluster scientific data, it is required to generate a matrix of data × special terms. Therefore, the special term sets used to represent the features of every piece of data in the scientific data training set should be the same. Since there are differences among term sets after the combination of features, it is necessary to filter out the candidate features by calculating the mean weight $\overline{U}$ for U values of all terms in formula 13 and choosing the top K terms with the highest $\overline{U}$. The formula for the mean weight of terms is as follows:

$$\overline{U}=\sum_{i=1}^{N}U\left(d_i\right) \tag{14}$$

In formula 14, N is the total number of data in the training set, and $U(d_i)$ is the weight value of terms representing data d_i. In this way, the final feature vectors of CTD-based scientific data are established.

5 Experiment and analysis

The biological information data will be taken as examples below to testify the effectiveness of aforesaid CTD-based feature representation method through experiments.

5.1 Selection of data

Table 1 shows the four gene data sets (including 26 genomes as mentioned in the literature[8]) categorized by functions and made by experts. To facilitate the comparison and verification of the methods proposed in the paper on the effects of data clustering, the experiment chooses the same 26 genomes, and retrieves relevant gene data of the 26 genomes in the PubMed database of Entrez; to save time and make less impact on the results, the experiment retrieves relevant reviews of the 26 genome in the Gene database of Entrez. For results, please see Table 2.

Table 1 The manual clustering list of 26 genomes categorized by functions

Group No.	Genome	Functions
1	GluR1, GluR2, GluR3, GluR4, GluR6, KA1, KA2, NMDA-R1, NMDA-R2A, NMDA-R2B	Glutamate receptor channels
2	Tyrosine hydroxylase, DOPA decarboxylase, Dopamine beta-hydroxylase, Phenethanolamine N-methyltransferase, Monoamine oxidase A, Monoamine oxidase B, Catechol-O-methyltransferase	Catecholamine synthetic
3	Actin, Alpha-tubulin, Beta-tubulin, Alpha-spectrin, Dynein	Cytoskeletal proteins
4	Chorismate mutase, Prephenate dehydratase, Prephenate dehydrogenase, Tyrosine transaminase	Enzymes in tyrosine and phenylalanine synthesis

Table 2 List of retrieved data and literature related to 26 genomes

No.	Genome	Number of retrieved data	Number of retrieved literature	No.	Genome	Number of retrieved data	Number of retrieved literature
1	Actin	13,006	5,372	14	GluR6	35	9
2	Alpha-spectrin	28	343	15	KA1	395	4
3	Alpha-tubulin	1,836	1,224	16	KA2	24	9
4	Beta-tubulin	738	1,224	17	Monoamine oxidase A	27	2,514
5	Catechol-O-methyltransferase	84	425	18	Monoamine oxidase B	24	2,514
6	Chorismate mutase	4,115	17	19	NMDA-R1	4	3
7	DOPA decarboxylase	433	206	20	NMDA-R2A	0	1
8	Dopamine beta-hydroxylase	51	339	21	NMDA-R2B	0	1
9	Dynein	3,914	487	22	Phenethanolamine N-methyltransferase	0	65
10	GluR1	53	37	23	Prephenate dehydratase	818	6
11	GluR2	55	48	24	Prephenate dehydrogenase	788	7
12	GluR3	18	11	25	Tyrosine hydroxylase	192	709
13	GluR4	24	13	26	Tyrosine transaminase	11	104

The alias, name and introduction of every piece of data in the experiment are taken as a text description “D”, while the title and abstract of relevant literatures are taken as a text description “T”. After that, the experiment pre-processes the text description sets “D” and “T” respectively. Besides, the vocabulary of the experiment include the terminologies of biomedical science, bioinformatics and gene engineering, with a total of 10,632 special terms.

5.2 Evaluation method

To evaluate the effects of data clustering, the paper uses some common methods: Precision (P), Recall rate (R) and Value F1, as shown in the formulars[9] below:

Precision of J-class: $$R_j = \left(l_j / m_j\right) \times 100\% \tag{15}$$

In the formula 15, l_j is the number of data clustered correctly in the J-class, and m_j is the number of data which is clustered into J-class actually by the clustering system.

The recall rate of J-class: $$R_j = \left(l_j / n_j\right) \times 100\% \tag{16}$$

In the formula 16, l_j is the number of data clustered correctly in the J-class, and n_j is the number of data which is clustered into the J-class by hand.

Value F1 of J-class: $$F1_j = \frac{P_j \times R_j \times 2}{P_j + R_j} \tag{17}$$

As the clustering system has a number of classes, the paper uses two algorisms—micro average and macro average to calculate the Precision, Recall rate and Value F1. They are defined as follows:

Macro average precision: $$MacroP = {}^{1}\!/_{n} \sum_{j=1}^{n} P_j \tag{18}$$

Macro average recall: $$MacroR = {}^{1}\!/_{n} \sum_{j=1}^{n} R_j \tag{19}$$

Macro average F1: $$MacroF1 = \frac{MacroP \times MacroR \times 2}{MacroP + MacroR} \tag{20}$$

Micro average precision: $$MicroP = \sum_{j=1}^{n} l_j \Big/ \sum_{j=1}^{n} m_j \tag{21}$$

Micro average recall: $$MicroP = \sum_{j=1}^{n} l_j \Big/ \sum_{j=1}^{n} n_j \tag{22}$$

Micro average F1: $$MicroF1 = \frac{MicroP \times MicroR \times 2}{MicroP + MicroR} \tag{23}$$

5.3 Analysis on the experimental result

K-mean clustering method is used in the experiment to compare the clustering effect of the three feature presentation methods—the method proposed in the paper,

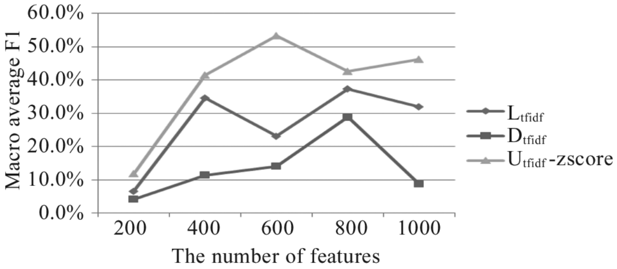

Fig. 2 Macro averages of the three methods

Note: In Fig. 2, L_{tfidif} is tf~idf feature representation method based on scientific literature source, D_{tfidf} is tf~idf feature representation method based on scientific data source, and U_{tfidf}-zscore is the method the paper propsed.

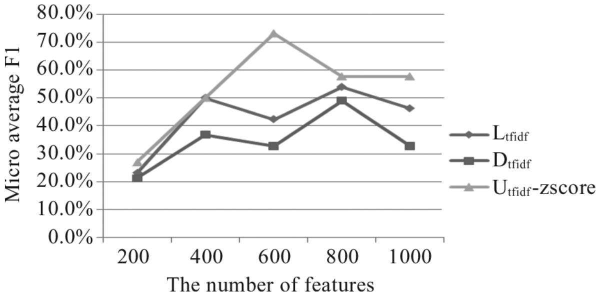

Fig. 3 Micro average of the three methods

Note: In Fig. 3, L_{tfidif} is tf~idf feature representation method based on scientific literature source, D_{tfidf} is tf~idf feature representation method based on scientific data source, and U_{tfidf}-zscore is the method the paper propsed.

tf~idf feature representation method based on scientific literature source and tf~idf feature representation method based on scientific data source. Experimental results are shown in Fig. 2 and Fig. 3. The experiment compares all the micro average F1 and macro average F1 of clustering effect of the three methods when the 200,400,600,800,1000 top $\overline{U}$ feature terms are kept in each method. The experimental result shows (in Fig. 2 and Fig. 3) that the method propsed in the paper is better than the other two methods for different number of the feature terms. Table 3 shows the macro average F1 and micro average F1 of the three methods mentioned above when 600 feature terms are kept.

Table 3 Macro average F1 and micro average F1 of the three methods when 600 feature terms are kept

	L_{tfidf}	D_{tfidf}	U_{tfidf}-zscore
Macro average F1	23.0%	14.0%	53.2%
Micro average F1	42.3%	32.7%	73.1%

From the experiment we can see that, the micro average F1 values and macro average F1 values of the method proposed in the paper can reach up to 50% or above when 600 feature terms are kept, while those of the other two methods are very low. The reasons are as follows:

- In the new method proposed by the paper, candidate features are extracted to remove noises to a certain extent. Moreover, after pre-processing, there are only 6,400 CTD-based candidate feature terms left, a relatively small number, which further reduces the effect of noise.
- The new method uses the Zscore weight algorism, and extends the range of variables in the Zscore. As it takes into account the proportion of the frequency of special terms to the frequency of its background terms for both a single piece of data and the whole scientific data training set, the peripheral features of scientific data are described in a more adequate manner.
- The new method combines the two weight algorisms (tfidf and Zscore), strengthens both the essential features and the peripheral features of scientific data, highlights the importance of essential features by assigning different weights to different features. When the experiment combines the two feature term sets of the 26 gene data, the average number of the same feature terms of the 26 gene data is only a little more than 320, which means, the percentage of terms in the composite text description that can represent both the essential features and peripheral features of scientific data in the text descriptions “D” and “T” is very small. This further proves that the new method is more effective.
- In the new method, the filtering of candidate features solves the problem of sparse data to a certain extent, reduces some noises, and enhances the density of features.

6 Summary and future work

As the current feature representation methods did not fully represent the scientific data and the clustering effect of data was not good enough, the paper proposes a method based on composite text description (CTD), which can represent and strengthen the essential features and peripheral features of scientific data. Experiments show that with this method, the features of scientific data can be described in a more comprehensive way and the clustering effect of scientific data is improved significantly, which means, the method proposed in the paper is effective and feasible. Our next step is to optimize the segmentation of special terms and non-login terms, based on which a large quantity of scientific data will be tested to perfect the method and further improve the clustering effect.

National Science Library, Chinese Academy of Sciences

References

1 Jiao, L. C., Liu, F., & Gou, S. P. Intelligent data mining and knowledge discovery (in Chinese). Xi’an: Xidian University Press, 2006:16.

2 Deng, X. B. Study on data extraction model and algorithm for complex data sources (in Chinese) (thesis). Shanghai: Fudan University (2005).

3 Masys, D. R., Welsh, J. B., & Fink, J. L., et al. Use of keyword hierarchies to interpret gene expression patterns. Bioinformatics, 2001, 17(4):319–326.

4 Liu, Y., Brandon, M., & Navathe, S., et al. Text mining functional keywords associated with genes. Studies in Health Technology and Informatics, 2004, 107(Pt 1):292–296.

5 Li, X. Y., & Fu, Y. Clustering in scientific data mining based on grid and iterative method. Chengdu University of Information Engineering College (in Chinese), 2006, 21(3):327–330.

6 Sun, Z. R, Han, T., & Yang, W. The analysis to the relationship between scientific data and scientific literature in bioinformatics. Library and Information Service (in Chinese), 2008, 52(2):88–91.

7 Liu, Y., Ciliax, B. J., & Borges, K., et al. Comparison of two schemes for automatic keyword extraction from MEDLINE for functional gene clustering. In Proceedings of IEEE Computational Systems Bioinformatics Conference. Stanford, CA : IEEE Computer Society, 2004:394–404.

8 Liu, Y., Navathe, S. B., & Civera, J., et al. Text mining biomedical literature for discovering gene-to-gene relationships: A comparative study of algorithms. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2005, 2(1):62–76.

9 Liu, H. F., Wang, Y. Y., & Zhang, X. R. An improved feature selection method in text classification. Information Science (in Chinese), 2007, 25(10):1534–1537.

(*Copy editor*: *Ning LI*)