High Quality Algorithm for Chinese Short Messages Text Clustering Based on Semantic

Fengxia Yang
Department of Computer Science
Cangzhou Normal University
Cangzhou , China
szyfx@163.com

***Abstract*—Existing data clustering method lacks considering of latent similar information existing among words,and it leads to unsatisfactory clustering result.Aiming at Chinese short message text clustering,this paper proposes a clustering algorithm based on semantic.It offers Chinese concept,and the measuring methods to calculate the similarity degree about words and Chinese short message text.It completes the clustering of Chinese short messages text through fission downwards and mergence of twos upwards.Experimental results show that this algorithm has better clustering quality than traditional algorithm.**

Keywords-short messages text;semantic;concept similarity

I. INTRODUCTION

Text clustering is an un-supervising machinery learning. By analyzing the text content, the text shall be divided into many meaningful classifications, in which the similarity of the same classification shall become as high as possible, and the similarity of the different classification shall become as low as possible. Now, the common text clustering algorithms are mainly hierarchical clustering method represented by G-HAC algorithm and flat division method represented by K-means algorithm. There are many achievements on text clustering at home and abroad. For example, text clustering algorithm based on semantic filtering model in literature[1]; text clustering algorithm based on fuzzy concepts in literature[2]; text clustering algorithm based on swarming intelligence Web in literature[3]; text clustering algorithm based on semantic inner space in literature[4]; achieving a high efficient text clustering algorithm by the chain fission downward and the two-two merging upward, based on the up-down relationship of primitive, constructing a primitive concept tree in literature[2] and so on. In literature[6] based on HowNet model, the author put forward a similarity calculation algorithm, but this algorithm only can apply to the similarity calculation between words and concepts and does not provide the text similarity calculation analysis. This article analyzes the text from the perspective of semantics, making semantic disambiguation firstly[7], expressing the texts as a keyword set, calculating the similarity of words with the similarity of non-weak primitives, and calculating the similarity of texts with the similarity of words. This algorithm analyzes the similarity among texts from the perspective of semantics, so the results better fit for people's institution.

II. SIMILARITY CALCULATION OF CHINESE SHORT-MESSAGE

A. HowNet

HowNet is an online common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts such as connoting of Chinese lexicons and their English equivalents(http://www.keenage.com).As a knowledge base,the knowledge structured by HowNet is a graph rather than a tree.It is devoted to demonstrating the general and specific properties of concepts.For instance, 'human being'is the general property of'doctor'and'patient'.The general properties of'human being'are documented in the main features of concepts.Being the agent of cure is the specific attribute of'doctor',whereas being the experiencer of sickness is the specific attribute of'patient'.Be it the rich or the poor,or the beauty or the ugly,they all share the same attribute of being a human being,though each takes a distinct attribute-value,that is,rich,poor,beautiful and ugly.The HowNet knowledge dictionary is the heart of the whole system.In this dictionary,every concept of a word or phrase and its description form one entry.Regardless of language types,an entry will comprise four items.These items are arranged in the following sequence:

W_X=word/phrase form
G_X=word/phrase syntactic class
E_X=example of usage
DEF=concept definition

Ref.[8]proposed a method to calculate the semantic similarity,which was applied well in text classification and text retrieval because of its accurate measurement of the similarity between words.

In HowNet, the meaning is defined as the combination of various primitive, which is used to describe the concepts of knowledge. Primitive is the smallest semantic unit in describing concepts. There are 1618 primitives adopted in HowNet which are classifies into ten classes: event, entity, attribute, value, quantity, q-value, secondary feature, syntax, event role and event feature. Among primitives, there exist complicated relationships in which up-down is the most important. According to up-down, primitive can be organized into a tree-diagram hierarchic system, which is the basis of calculating the similarity of words. Here are two figures. Figure1 is the primitive hierarchic system of event and figure2 is the primitive hierarchic system of entity.

Published by Atlantis Press, Paris, France.

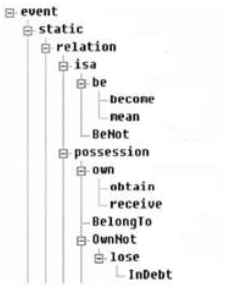

Figure 1. Primitive hierarchic system tree of event

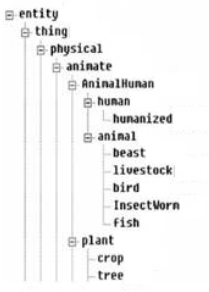

Figure 2. Primitive hierarchic system tree of entity

B. Similarity calculation of primitive

Applying the rich semantic information of describing every concept in HowNet, the author of this article defines the similarity of primitives through the data, structure and knowledge describing language in HowNet.

Definition 1: the distance between pi and parent(p_i):

$$d(p_i, parent(p_i)) = m - D_i \cdot n \quad (1)$$

D_i is the depth of the primitive pi, m is the initial threshold value of the distance, n is a positive real number to satisfy the inequality max(D)<m/n.

Definition 2: the definition of the arbitrary two primitives pi and pj is as follow:

$$d(p_i, p_j) = \omega_k.[m - \max(D_i, D_j).n] \quad (2)$$

In the above formula, ω_k refers to the corresponding weights to the k relationship, usually taking $\omega_k \geqslant 1$.

From formula (1) and formula (2), we can see that the deeper the primitive classifies in the hierarchic tree-diagram, the smaller and the more similar their distance is. It is consistent with people's visual pression.

For the arbitrary path t between pi and pj, the $d_t(p_i,p_j)$ is calculated according to formula (1) and formula (2), so that the smallest distance between the arbitrary two primitives is $d_{min}=\min(d_t(p_i,p_j))$.

Definition 3: the similarity of primitive

$$Sim(p_i, p_j) = \frac{\alpha}{d_{\min}(p_i, p_j) + \alpha} \quad (3)$$

In the above formula, p_i and p_j express the arbitrary two primitives, and $d_{\min}(p_i, p_j)$ is the shortest distance between pi and pj, α is an adjustable parameter.

C. Similarity calculation of words

Every word's meaning is consisted of many primitives which are not equal. The first primitives of many words are always abstract and repeated that has little help to get the semantic information of these words. In the example of "friendship" and "characteristic", the primitive of "friendship" is "attribute, relatedness and human", but the primitive of "characteristic" is "attribute, property and entity". From the example, we can see that the first primitive of the two words is same, but the following two primitives are different. However, the first primitive attribute is on the top of primitive structure, the calculation result of semantic similarity is 0.747 that is far from people's direct judge-ment. Therefore, the author of this article put forward the method of calculating the similarity of two words by comparing the non-weak primitives, thus the calculation distance effect has to be greatly improved that is closer to people's direct judge-ment.

Definition 4: Non-weak primitive refers to the primitive which has more help to get the words' semantic information.

Definition 5: Supposing that the word w_1 has n primitives of $p_{11}, p_{12}, \cdots, p_{1n}$, w_2 has m primitives of $p_{21}, p_{22}, \cdots, p_{2m}$, so the similarity of w_1 and w_2 is defined as the maxim of every non-weak primitive similarity. That is

$$sim(w_1, w_2) = \max\{sim(p_{1i}, p_{2j}) \mid p_{1i} \in w_1, p_{2j} \in w_2\} \quad (4)$$

HowNet is a lexical base with rich semantic information. It uses 1 618 sememes to describe words. The related words have the same sememe when they are described by the HowNet. the paper proposed an improved algorithm to compute the similarity between the related words. It also introduced concept about weak sememes and excluded such sememes' interference when they appeared in the computation of the word's similarity. The experiment proves the improved word similarity computation meets the peoples' intuition and text mining better.

D. Calculation of text similarity

Text contains the basic semantic units of words and phrases, in other words, text contains the semantic concepts of words or phrases.This article picks the unsupervised words characteristic selection algorithm, and with TFIDF, text can be transferred as words that the similarity of text can be calculated with the similarity of words.

Definition 6: Supposing the similarity between Chinese message text $SM_i(w_{i1}, w_{i2}, \cdots, w_{in})$ and Chinese message text $SM_j(w_{j1}, w_{j2}, \cdots, w_{jn})$ is

$$Sim(SM_i, SM_j) = \sqrt{(w_{i1} - w_{j1})^2 + (w_{i2} - w_{j2})^2 + \cdots (w_{in} - w_{jn})^2} \quad (5)$$

, in which $w_{ik}, 1 \leq k \leq n$ and $w_{jk}, 1 \leq k \leq n$ respectively represent the weights of word's vector of short-message text SM_i and SM_j.

Definition 7: The similarity between a short-message text SM_i and a text set V is defined as the minimum of similarity between this short-message text and every text of this set. That is

$$Sim(SM_i, V) = \min\{Sim(SM_i, SM_j) \mid SM_j \in V\} \quad (6)$$

Definition 8: The similarity between the text sets S and V is defined as the minimum of similarity among texts in these two sets. That is

$$Sim(S,V) = \min\{Sim(SM_i, SM_j) \mid SM_i \in S, SM_j \in V\} \quad (7)$$

Definition 9: The quality of clustering set S is defined as

$$Density = \sqrt{\sum_{SM_i, SM_j \in S} Sim^2(SM_i, SM_j) / |S|} \quad (8)$$

III. HIGH EFFICIENT CHINESE SHORT-MESSAGE CLUSTERING ALGORITHM BASED ON SEMANTIC

The basic idea of high efficient Chinese short-message clustering algorithm based on semantic is : the whole Chinese short-message set can be seen as merged collection in which every text is a separate class and the quality of every subclass is judged. If it exceeds the given threshold value, we will delete it and add the clustering results. Otherwise, the similarity of subclasses has to be calculated, and if it is beyond the given threshold value, we will merge the similarity among the subclasses until the similarity is less than the given threshold value.

High efficient Chinese short-message clustering algorithm based on primitive.

Input: Chinese short-message text set: $Set = (S_1, S_2, \cdots S_n)$, given value t; clustering quality value g.

Output: Clustering Result Set: $\mathrm{Re}\,sultSet = (S_1^{'}, S_2^{'}, \cdots S_m^{'})$

(1)Initialization Set and clustering Result Set;

/*seeing every short-message text in the Set as an initialization class and initializing Result Set as an empty set*/

(2)for $\forall S_i \in Set$

(3)Calculating the quality of S_i according to formula (8)

(4)if(Density(S_i)>g)

(5)Deleting S_i from the Set and adding it into the clustering Result Set;

(6)then for $\forall S_j \in Set, S_j \neq S_i$

(7)Calculating the similarity of the arbitrary sub-classes S_i and S_j with each of the formulas (3) and (4) and (5).

(8)if($Sim(S_i, S_j)$ >t)

(9)Combining S_i and S_j as a class, at the same time deleting S_i and S_j from the Set. Seeing （S_i, S_j）as a new sub-class which will be added into the Set;

(10)endif

(11)endfor

(12)endif

(13)endfor

(14) All sub-classes in the Set shall be added into the clustering Result Set;

(15) Inputting clustering Result Set= $(S_1^{'}, S_2^{'}, \cdots S_m^{'})$.

IV. EXPERIMENT AND RESULT ANALYSIS

The algorithm is achieving by Visual C++6.0, using the database of MS SQL Server 2000, experimenting in the brand of Founder computer equipped with memory 2.0 G, clocked dual-core 2.0 GHz and Windows XP operating system. The initial parameters of this experiment are α =0.5, t=0.45, g=0.40.

Experiment 1: Similarity calculation of words. The result of this experiment is showed in Figure 1.

TABLE I. SIMILARITY COMPARING OF WORDS

Words 1	Words 2	Document【6】Methods and Results	Methods and Results in this article
bread	apples	0.186047	0.683152
everything	space	0.444444	0.371121
fruit	vegetables	0.444444	0.531118
eat	bread	0.074074	0.165912
bread	chocolate	1.000000	1.000000
biology	insects	0.377559	0.458759
apples	knife	0.285714	0.249890
substances	insects	0.285714	0.300018
cut	fruit	0.074074	0.100849

From figure1, we can see that according to the calculation methods in document【6】, the similarity of “everything”and “space”, “fruit”and “vegetables” are both “0.444444”; the similarity of “apples”and “knife”, “substances”and “insects” are both “0.285714”, that are irrational obviously. According to the concept of words’ similarity (In different contexts, two words can be used interchangeably without changing the syntactic semantic structures of text), the similarity of “fruit”and “vegetables” are more. Therefore, the results of using the methods of this article fit better for people’s intuition.

Experiment 2: Comparing of clustering quality with different algorithms. In this article, the algorithm is Div-Mer, and the purpose of this experiment is mainly comparing Div-Mer algorithm with Div-Mer algorithm without considering similar semantics (called Div-MerN). The overall clustering quality refers to the ratio of average quality and clustering quantity.

TABLE II. CLUSTERING QUALITY COMPARING IN DIFFERENT SPECIES

Algorithm	Total clustering	Average quality/(%)	Time/s	Overall quality
Div-Mer	14	76.3	2.85	5.45
Div-MerN	89	69.4	0.77	0.78

From the results of this experiment, we can see that the overall clustering quality with Div-Mer algorithm is obviously superior to Div-MerN algorithm, and is seven times. But from the needing time, Div-Mer algorithm is more than Div-MerN algorithm. It is because Div-Mer algorithm adds the calculation of Chinese short-message text similarity that can increase the processing capability by improving the hardware’s performance and adopting the parallel processing method.

V. CONCLUSION

The author of this article put forward a Chinese short-message text processing algorithm based on semantics. According to the conceptual structure in HowNet, through primitive description concept, concept description words and word description Chinese short-message text, the author of this article defined primitive, word, similarity of Chinese short-message text successively, and calculated the similarity of two words by comparing the non-weak primitives of two words. The results fit better for people's intuition so that the efficiency of calculating the distance among words is greatly increased. This algorithm is applicable to the clustering general texts, whose word vector's dimension in transforming is too high, meanwhile the calculating quantities are too much, so it will take too much time to calculate them with this algorithm. From the analysis, in the premise of ensuring the high quality of clustering, the next step is to improve the efficiency of executing this algorithm.

REFERENCES

[1] Olsen R A,Kroes G J,"Comparison of Methods for Finding Saddle Points Without Knowledge of the Final States,"Journal of Chemical Physics,2004,121(20):9776-9792..

[2] Ning Chen, An Chen, Longxiang Zhou, Weijia Luo and Sanding Luo. "Text Clustering based on Fuzzy Concepts and its application in Web ," Journal of Software, 13(8):1598-1605,2002.

[3] Ken Chen,"Saddle-point Based Separation of Touched Objects in 2-D Image,"Journal of Electronics,2006,23(3):452-456.

[4] Jing Peng, Dongqing Yang, Shiwei Tang, Yan Fu and Hankui Jiang, "A Text Clustering Algorithm based on Semantic Inner Space Model," Journal of Computer, 30(8):1354-1363,2007.

[5] Jinling Liu.,"High Efficient Chinese Short-message Text Clustering Algorithm based on Semantic ," Computer Engineering, 5(10):201-205,2009.

[6] Yan Xu, Hui Zhang,"An Improved Algorithm for Vessel Centerline Tracking in Coronary Angiograms,"Computer Methods and Programs in Biomedicine,2007,88(2):131-143

[7] Erhong Yang, Guoqing Zhang and Yongkui Zhang.," Chinese Semantic Excluding Method based on Co-occurrence of Primitives ,". Computer Research and Development, 38(7):833-838,2001.

[8] Q Liu, S J. Li,"The computation of semantic similarity based on HowNet,"Proceedings of 3rd Chinese Word Semantics Workshop,May 1?4,2002,Taibei,China.2002:59-76(in Chinese)

Published by Atlantis Press, Paris, France.