Fuzzy Set Based Web Opinion Text Clustering Algorithm

Hongxin Wan[1, a], Yun Peng[2, b]

[1]College of Mathematics & Computer Science, Jiangxi Science & Technology Normal University, Nanchang 330013, China;

[2]College of Computer and Information Engineering, Jiangxi Normal University,

Nanchang 330022, China.

[a]wanhongxin@126.com, [b]pengyunmail@126.com

Keywords: Opinion text, Clustering analysis, Fuzzy set, Membership function

Abstract. With the development of social media, people like to express their views on the Web. Because people express their views casually, which makes the opinion text contain a lot of uncertain and unstructured contents, and it is difficult to cluster the text by normal classification methods. An algorithm of opinion text clustering based on fuzzy set is proposed, which adopts the key words as classifying attributes and calculate the membership degree by fuzzy function, thus can deal with the uncertain and unstructured contents well. Also the algorithm proposed can improve the time and space efficiency, and increase the robustness compared with other classification algorithms.

Introduction

Most web opinions are expressed in the form of text, which contain a lot of uncertain contents, and fuzzy set can deal with those data more efficiently[1]. To mine classification information from the text, it is important to cluster opinion texts and get the classification knowledge. The text clustering analysis should divide the text corpus into some clusters and the likeness degree of texts in one cluster is to be requested similar as possible, on the contrary, the likeness degree of text in different clusters is to be requested distinct as possible, which can discover the whole distribute characteristics of the corpus[2]. Unlike the normal classification clustering analysis doesn't define classification topics in advance[3]. We can make use of text clustering to divide the opinion texts into some clusters, and the users can pay more attention to those related clusters, that should decrease the browsing quantity greatly. Considering the uncertain contents and unstructured of the opinion text an algorithm of opinion text clustering is proposed based on fuzzy set in this paper[4].

Text Clustering algorithm

The attribute words and phrases of the text can be provided according to the concerning area by users. The frequency of these keywords and phrases is the basis of text clustering analysis. When the frequency of a certain keyword or phrase is high, the attribute value that belongs to the text is high also, so make use of the attributes we can depict the characteristic of the text[5]. During the period of the text clustering, keywords can be divided into different grade. We can establish different power parameter for the different grade keyword, and the attribute data should be changed into the decimal value smaller or equal to 1 for the next step of fuzzy clustering[6].

Definition 1. Given a fuzzy set *A* in domain *U*, A can be described by membership function $\mu_{\underset{\sim}{A}}$,

$\mu_{\underset{\sim}{A}}$: $U \to [0,1]$, if $u \in U$, exists $u \to \mu_{\underset{\sim}{A}}(u)$, $\mu_{\underset{\sim}{A}}(u) \in [0,1]$, $\mu_{\underset{\sim}{A}}(u)$ is the membership degree of u to set *A*.

(1) Set the membership function shown in (1), where, *x* is the frequency of attribute word, σ is variance of *x*, *a* is the frequency threshold. By the membership function, we can get the attributes membership of all text, and the membership value is less than 1[7].

© 2015. The authors - Published by Atlantis Press

$$\mu_{\underset{\sim}{A}}(x)=\begin{cases}1-e^{-\left(\frac{x-a}{\sigma}\right)^{2}} & x>a\\ 0 & x\le a\end{cases}\qquad\text{□□□(1)}$$

(2) Set up the fuzzy similarity relation matrix *R*. Element r_{ij} of *R* is calculated by Euclidean distance formula shown in (2), the order of *R* matrix is |*U*|, where, *m* is the number of attributes.

$$r_{ij}=\begin{cases}1 & i=j\\ \sqrt{\frac{1}{m}\sum_{k=1}^{m}(u_{ik}-u_{jk})^{2}} & i\neq j\end{cases}\qquad(2)$$

(3) The graph *G* = (*V*, *E*) can be obtained by R, and the maximum spanning tree *T* = (*V*, *TE*) from *G* can be calculated using Prim algorithm.

(4) According to the real data, set the threshold $\lambda\in[0,1]$, *T*(*e*) is the weight of edge *e*, if $T(e)<\lambda$, edge e should be removed, and the connected components are the classification based on λ.

Example Analysis

We give the original data and analyze the text clustering process by example. According to the keyword frequency the attribution degree to the keywords can be calculated by membership function, and the next clustering is carried out based on the attribution degree. Give an original data table such as Table 1, t_i is text unit, a_i is the keyword.

Table 1 original text data

T	a_1	a_2	a_3	a_4	a_5	a_6
t_1	14	15	10	3	12	10
t_2	20	10	50	51	28	30
t_3	54	27	55	9	18	55
t_4	11	31	5	21	17	4
t_5	46	12	65	31	22	65
t_6	45	29	51	12	24	51
t_7	59	15	58	19	25	58
t_8	19	21	19	20	27	19
t_9	52	25	55	20	15	55
t_{10}	49	19	56	21	19	56
t_{11}	5	7	27	12	26	27
t_{12}	21	15	59	15	47	49
t_{13}	52	15	56	27	17	56
t_{14}	41	17	55	25	31	35
t_{15}	52	29	51	21	16	41

(1) Pretreatment of Original Text

The pretreatment table data can be derived by membership function, such as Table 2.The attribute value has been changed to the value smaller or equal to 1, and the value reflects the dependency to the keyword attribute.

Table 2 Text data after pretreatment

T	a_1	a_2	a_3	a_4	a_5	a_6
t_1	0.233	0.243	0.200	0.019	0.200	0.100
t_2	0.213	0.133	0.600	0.566	0.293	0.333
t_3	0.667	0.267	0.600	0.267	0.400	0.600
t_4	0.533	0.257	0.100	0.467	0.200	0.100
t_5	0.437	0.367	0.067	0.467	0.210	0.200
t_6	0.457	0.417	0.600	0.467	0.500	0.400
t_7	0.467	0.267	0.600	0.467	0.610	0.710
t_8	0.513	0.210	0.133	0.200	0.255	0.267
t_9	0.633	0.231	0.700	0.267	0.337	0.267
t_{10}	0.543	0.220	0.600	0.467	0.680	0.760
t_{11}	0.133	0.150	0.200	0.467	0.232	0.133
t_{12}	0.467	0.133	0.133	0.267	0.168	0.267
t_{13}	0.487	0.163	0.600	0.267	0.681	0.670
t_{14}	0.533	0.200	0.630	0.467	0.390	0.400
t_{15}	0.567	0.320	0.600	0.200	0.167	0.367

(2) Clustering of Opinion Text

The element value in fuzzy similar matrix R can be calculated by Euclidean length formulae shown in (2), and the matrix R is shown in Table 3.

Table 3 Fuzzy similar matrix R

	t_1	t_2	t_4	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{14}	t_{14}	t_{15}
t_1	1														
t_2	0.031	1													
t_4	0.123	0.445	1												
t_4	0.234	0.211	0.250	1											
t_5	0.169	0.140	0.470	0.250	1										
t_6	0.317	0.237	0.126	0.140	0.270	1									
t_7	0.442	0.242	0.189	0.094	0.255	0.189	1								
t_8	0.152	0.079	0.224	0.244	0.144	0.257	0.251	1							
t_9	0.186	0.191	0.074	0.099	0.260	0.226	0.115	0.211	1						
t_{10}	0.235	0.245	0.099	0.094	0.248	0.066	0.042	0.248	0.107	1					
t_{11}	0.327	0.093	0.247	0.184	0.279	0.225	0.269	0.147	0.209	0.215	1				
t_{12}	0.061	0.278	0.225	0.249	0.116	0.247	0.242	0.052	0.214	0.239	0.129	1			
t_{14}	0.204	0.217	0.368	0.144	0.276	0.194	0.158	0.224	0.374	0.099	0.247	0.217	1		
t_{14}	0.235	0.245	0.399	0.094	0.158	0.067	0.042	0.248	0.107	0.033	0.215	0.249	0.099	1	
t_{15}	0.386	0.196	0.074	0.140	0.270	0.146	0.137	0.211	0.042	0.137	0.225	0.234	0.173	0.374	1

The graph $G(V, E)$can be derived from matrix R. The maximum spanning tree $T(V, TE)$ shown in fig.1 can be derived by Prim algorithm, where, $|V|=15$, $|TE|=14$.

According to the actual question, set a proper $\lambda\in[0,1]$. If $T(e)<\lambda$，then delete the edge e. Set $\lambda=0.220$，the connected component show in fig.2 and then the clustering results are obtained.

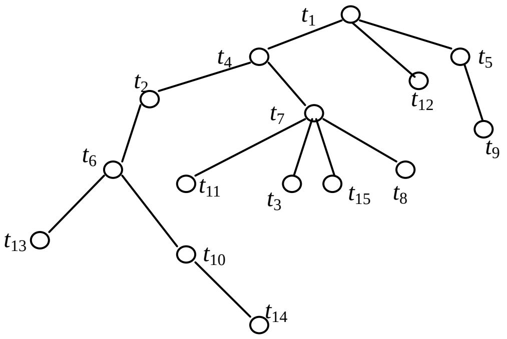

Fig.1 Maximum spanning tree

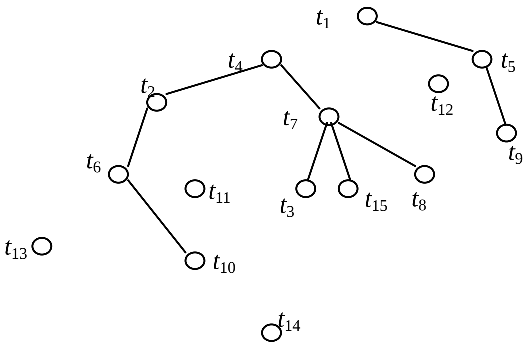

Fig.2 Connected components

$C_1=\{t_2, t_3, t_4, t_6, t_7, t_8, t_{10}, t_{15}\}$，$C_2=\{ t_1, t_5, t_9\}$，$C_2=\{t_{11}\}$，$C_2=\{t_{12}\}$，$C_2=\{t_{13}\}$，$C_2=\{t_{14}\}$

Summary

Recently text mining is an important research area in information technology. In web opinion texts, some data is certain, also exists a lot of uncertain data. Applying fuzzy algorithm can improve the ability to deal with these uncertain data, and also increase the robust of algorithm. Because the amount of opinion text is very big, the algorithm considers the time and space efficiency, and uses the key attributes to decrease the data dimension. The next work we look forward to finding more effective attributes to get more accuracy clustering analysis of opinion texts.

Acknowledgements

This work is supported by Social Science Planning Project of Jiangxi Province (No.13TQ08, 14TQ04). The authors are grateful for the reviewers of initial drafts for their helpful comments and suggestions.

References

[1] Zimmermann H J. Fuzzy set theory-and its applications. Springer Science & Business Media, 2001.

[2] Deschrijver G, Kerre E E. On the relationship between some extensions of fuzzy set theory[J]. Fuzzy sets and systems, 2003, 133(2): 227-235.

[3] Izakian H, Abraham A. Fuzzy C-means and fuzzy swarm for fuzzy clustering problem. Expert Systems with Applications, 2011, 38(3): 1835-1838.

[4] Kriegel H P, Ntoutsi E. Clustering high dimensional data: Examining differences and commonalities between subspace clustering and text clustering-A position paper. ACM SIGKDD Explorations Newsletter, 2014, 15(2): 1-8.

[5] Kriegel H P, Ntoutsi E. Clustering high dimensional data: Examining differences and commonalities between subspace clustering and text clustering-A position paper. ACM SIGKDD Explorations Newsletter, 2014, 15(2): 1-8.

[6] Jun S, Park S S, Jang D S. Document clustering method using dimension reduction and support vector clustering to overcome sparseness. Expert Systems with Applications, 2014, 41(7): 3204-3212.

[7] Otadi M, Mosleh M. Solving fully fuzzy matrix equations. Applied Mathematical Modelling, 2012, 36(12): 6114-6121.