2008 International Symposium on Computer Science and Computational Technology

Volume 1

ISCSCT 2008

20-22 December 2008

Shanghai, China

An Algorithm of Web Text Clustering Analysis Based on Fuzzy Set

Yun PENG
College of Computer Information and Engineering
Jiangxi Normal University
Nanchang ,China
pengyunmail@126.com

Shu-liang DING
College of Computer Information and Engineering
Jiangxi Normal University
Nanchang ,China
ding06026@163.com

Abstract: **There are a large quantity of non-certain and non-structure contents in the web text at the present time. It is difficult to cluster the text by some normal classification methods. An algorithm of web text clustering analysis based on fuzzy set is proposed in this paper, and the algorithm has been described in detail by example. The technique can improve the algorithm complexity of time and space, increase the robustness of the algorithm. To check the accuracy and efficiency of the algorithm, the comparative analysis of the sample and test data is provided in the end.**

Keywords: web text; clustering analysis; fuzzy set; membership function

I. *Introduction*

Most web information had been stored as text, so the text mining has a higher commercial value and business potential than data mining in web. When the mining data is almost composed of text, this mining process is called text mining. The main function of text mining is to withdraw an unknown knowledge from the original text. Text mining is also a very difficult work, because those texts to handle are fuzzy and non-structure[1-6].

The text clustering analysis should divide the text files into some clusters. The likeness degree of text 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. It can discover the whole distribute characteristics of the text set. Unlike the normal classification, clustering doesn't define a classification topic in advance. We can make use of text clustering to divide the web texts into some clusters. The users can pay more attention to those related clusters, that should decrease the result quantity to be browsed greatly[7-12]. In considering the fuzzy contents and structures of the web text, it is put forward a text clustering analysis technique based on fuzzy set in this paper[13-14].

II. *Pretreatment of Original Text*

The attribute dictionary of the text can be provided according to the search keywords or the professional phrase of the related field[15]. 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 attribute dictionary we can depict the characteristic of the text. 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[16-18].

Definition 1. Given a fuzzy set A in domain U, A can be described by membership function $\mu_{\underline{A}}, \mu_{\underline{A}}: U \to [0,1]$, if $u \in U$, exists $u \to \mu_{\underline{A}}(u)$, $\mu_{\underline{A}}(u) \in [0,1]$, $\mu_{\underline{A}}(u)$ is the attribute degree of u to set A.

Suppose U is text domain, Ui is the No. i text element in U, $i \in 1,2,3\cdots n$, K_j is the No. j keyword attribute in U, $j \in 1,2,3\cdots m$, S_{ij} is the attribute value

978-0-7695-3498-5/08 $25.00 © 2008 IEEE
DOI 10.1109/ISCSCT.2008.139

of the No. i element and the No. j attribute in U(the frequency of the keyword), $P(K_j)$ is the No.j keyword grade power value. The membership function of the keywords is: $\underset{\sim}{\mu}_A(S_{ij})=F(S_{ij})/(S_{1j}+S_{2j}+S_{3j}+\ldots\ldots+S_{nj})$ $*P(K_j)$

III. *Clustering Analysis Algorithm of Text*

The fuzzy clustering algorithm can be designed to the web text after pretreatment. The algorithm is described as following:

Set up the fuzzy similar matrix R of text element on domain U. The dimension of matrix R is |U|. The count of attribute is m. The element rij in matrix R can be calculated by Euclidean length formulae:

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

The graph G（V，E）can be derived from matrix R. The maximum spanning tree T（V，TE） can be derived by Prim algorithm.

According to the actual question, set a proper $\lambda\in[0,1]$, T（e）is the power value of edge e, if T（e）$<\lambda$，then delete the edge e, the connected component is the classification based on λ.

The analysis of singular class. The clustering class can be sifted after clustering analysis. If the elements number is lower than a given limit value, the class can be regarded as a singular class. Some singular class may lurk special clustering property.

IV. *An Example Analysis*

We can analyze the text clustering process of test data by example. According to the keyword frequency of the two-dimension table, the attribution degree of the keywords can be calculated by membership function. The next clustering is carried out based on the attribution degree. Give an original data table such as table Ⅰ , U is text unit, KW is the keyword frequency.

TABLE Ⅰ. ORIGINAL TEXT DATA

U	KW_1	KW_2	KW_3	KW_4	KW_5
u_1	30	35	10	3	12
u_2	20	10	40	43	28
u_3	45	27	55	9	18
u_4	11	33	5	21	17
u_5	46	32	65	13	22
u_6	45	29	53	12	24
u_7	49	34	48	19	25
u_8	39	21	39	20	27
u_9	42	25	45	20	14
u_{10}	49	19	56	21	19
u_{11}	4	7	27	12	26
u_{12}	23	34	49	15	27
u_{13}	42	15	46	27	17
u_{14}	41	17	55	24	21
u_{15}	52	29	51	23	16

A. *Pretreatment of Original Text*

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

TABLE Ⅱ. TEXT DATA AFTER PRETREATMENT

U	KW_1	KW_2	KW_3	KW_4	KW_5
u_1	0.533	0.133	0.200	0.200	0.200
u_2	0.533	0.133	0.200	0.267	0.133
u_3	0.467	0.267	0.600	0.267	0.400
u_4	0.533	0.267	0.600	0.467	0.200
u_5	0.467	0.267	0.067	0.467	0.200
u_6	0.467	0.067	0.600	0.467	0.400
u_7	0.467	0.267	0.600	0.467	0.400
u_8	0.533	0.200	0.133	0.200	0.267
u_9	0.533	0.200	0.600	0.267	0.267
u_{10}	0.533	0.200	0.600	0.467	0.400
u_{11}	0.533	0.200	0.200	0.467	0.133
u_{12}	0.467	0.133	0.133	0.267	0.267
u_{13}	0.467	0.133	0.600	0.267	0.400
u_{14}	0.533	0.200	0.600	0.467	0.400
u_{15}	0.467	0.200	0.600	0.200	0.267

B. *Clustering of Web Text*

The element value in fuzzy similar matrix R can be calculated by Euclidean length formulae, the R such as table Ⅲ.

The graph G（V,E）can be derived from matrix R. The maximum spanning tree T（V,TE）can be derived by Prim algorithm. In the tree, |v|=15,|TE|=10,such as figure 1.

According to the actual question, set a proper $\lambda\in[0,1]$. If T（e）$<\lambda$，then delete the edge e. Set $\lambda=0.250$，the connected component and the clustering results are obtained from figure 2.

TABLE III. FUZZY SIMILAR MATRIX R

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	u_{11}	u_{12}	u_{13}	u_{14}	u_{15}
u_1	1														
u_2	0.042	1													
u_3	0.213	0.225	1												
u_4	0.223	0.211	0.130	1											
u_5	0.149	0.130	0.270	0.240	1										
u_6	0.237	0.237	0.126	0.130	0.270	1									
u_7	0.242	0.242	0.089	0.094	0.255	0.089	1								
u_8	0.052	0.079	0.223	0.244	0.133	0.257	0.251	1							
u_9	0.186	0.191	0.073	0.099	0.260	0.126	0.115	0.211	1						
u_{10}	0.235	0.235	0.099	0.094	0.258	0.066	0.042	0.248	0.107	1					
u_{11}	0.127	0.094	0.237	0.184	0.079	0.225	0.219	0.137	0.209	0.215	1				
u_{12}	0.060	0.073	0.225	0.239	0.116	0.237	0.242	0.052	0.213	0.239	0.119	1			
u_{13}	0.204	0.217	0.060	0.143	0.276	0.094	0.108	0.223	0.073	0.099	0.237	0.217	1		
u_{14}	0.235	0.235	0.099	0.094	0.258	0.066	0.042	0.248	0.107	0.000	0.215	0.239	0.099	1	
u_{15}	0.186	0.196	0.073	0.130	0.270	0.146	0.137	0.211	0.042	0.137	0.225	0.213	0.073	0.173	1

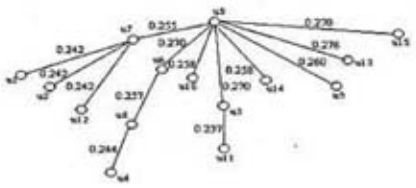

Figure 1. MAXIMUM SPANNING TREE

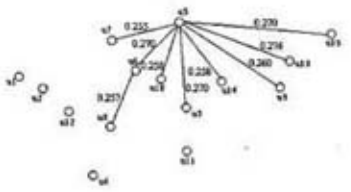

Figure 2. CONNECTED COMPONENT

C1={u3, u5, u6, u7, u8, u9, u10, u13, u14, u15}, C2={u1}, C3={u2}, C4={u4}, C5={u11}, C6={u12}

Set singular limit value z0=0.10, the element count proportion of class that lower 0.10 can be considered as singular class.Because P(C1)=10/15=0.667, P(C2)=1/15=0.067 , P(C3)=1/15=0.067 , P(C4)=1/15=0.067 , P(C5)=1/15=0.067, P(C6)=1/15=0.067. so C1 is the interested class.

V. *Comparative Analysis*

The test data includes 5 types of total 1000 articles, selected from http://www.cnki.net. The 5 types are basic science, engineering science, agricultural science, medical and sanitary science and information science. In the test, 550 articles are taken as training sample, and 450 articles are taken as test sample. These articles have been handled by participle processing and characteristic selection, then 560 characteristic phrases can be withdrawn.

We adopt Kohonen network of non-supervision to cluster the test sample. Because the dimension number is huge, so the network training is too slowly, and clustering accuracy is not high. Test result as table IV.

We adopt pattern clustering and Kohonen network with supervision to cluster the test sample. Because the dimension number is decreased to 210, so the network training time is less than above-mentioned method, and clustering accuracy has improved a lot. Test result as table V.

We adopt text clustering analysis based on fuzzy set to cluster the test sample. The training time is more less than above-mentioned methods, and clustering accuracy has improved significantly. Test result as table VI.

TABLE IV. TEXT CLUATERING RESULT BASED ON KOHONEN NETWORK

	Basic science	Engineering science	Agricultural science	Medical and sanitary science	Information science
Wrong articles	30	37	37	35	31
Correct articles	65	57	51	51	56
Articles after clustering	97	93	89	87	84
Correction ratio	67%	61.3%	57.3%	58.6%	66.7%

TABLE V. THE CLUSTERING RESULT BASED ON PATTERN CLUSTERING AND KOHONEN NETWORK

	Basic science	Engineering science	Agricultural science	Medical and sanitary science	Information science
Wrong articles	28	30	37	33	31
Correct articles	65	57	60	55	54
Articles after clustering	95	90	94	86	85
Correction ratio	68.4%	63.3%	63.8%	64%	63.5%

TABLE VI. TEXT CLUATERING RESULT BASED ON FUZZY SET

	Basic science	Engineering science	Agricultural science	Medical and sanitary science	Information science
Wrong articles	23	27	35	28	30
Correct articles	63	60	61	59	64
Articles after clustering	88	90	92	89	91
Correction ratio	71.5%	66.7%	66.3%	66.3%	70.3%

We can draw the conclusion through the experimental result. The pattern clustering can decrease the dimension to a certain extent, so the clustering speed and accuracy should be increased. But the dimension count is still large, so the implement of the algorithm is difficult still. The clustering analysis based on fuzzy set can decrease the complexity of time and space greatly, and reduce the dimension count. It can improve the speed and accuracy of text clustering.

VI. *Conclusions and Future Work*

Recently text mining is an important research area in information technology. Applying fuzzy technique to text mining has a lot of content to research. In text data, some data is certain, also exists much non-certain data. So the text mining based on fuzzy technique has great theoretical significance and practical value. Using several methods included of genetic algorithm, fuzzy set and rough set to solve text classification has an immense research prospect.

VII. *Acknowledgement*

This work is supported by Natural Science Foundation of China (No.60263005), Natural Science Foundation of Jiangxi Province (No.0011013), Science Research Foundation for Young Teachers of Jiangxi Normal University (No.200509). The authors are grateful for the reviewers of initial drafts for their helpful comments and suggestions.

REFERENCES

[1] P.Li and YYang. A Loss function analysis for classification methods in text categorization. In ICML 2003: 472–479.

[2] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, et al. Web usage mining: discovery and applications of usage patterns from Web data. SIGKDD Explorations, 2000, 1(2): 12–23.

[3] Haesun Park, Lower dimensional representation of text data based on centroids and least squares, BIT Numerical mathematics 43, 2003, 427–448.

[4] Yu Wang, Zheng-ou Wang, Text Categorization Rule Extraction Based on Fuzzy Decision Tree, IEEE Proc. of 2005 International Conference on Machine Learning and Cybernetics. 2004.8.26–29, Guangzhou, CHINA.

[5] Konald N Kostoff. High Quality Information Retrieval for Improving the Conduct and Management of Research and Development. Lecture Notes in Computer Science. 2003,1932/2000:86–95.

[6] Ji He, Ah-Hwee Tan, Chew-Lim Tan. A Comparative Study on Chinese Text Categorization Methods.PRICAI 2000 Workshop on Text and Web Mining, Melbourne, 2000, (8): 24–35.

[7] Anton Leuski., James Allan. Interactive Information Retrieval Using Clustering and Spatial Proximity. User Modeling and User-Adapted Interaction. 2004,14:259–288.

[8] DW vander Merwe, AP Engelbrecht, Data Clustering using Particle Swarm Optimization, Machine Learning, Vol.19, 2003, 139–174.

[9] Xiang Xiao, Ernst R.Dow, Russell Eberhart, Zina Ben Miled, Rober J, Gene Clustering Using Self-Organizing Maps and Particle Swarm Optimization, Nat.Genetics,Vol.32, 2003, 276–285.

[10] Ajith Abraham, Vitorino Ramos. Web Usage Mining Using Artificial Ant Colony Clustering and Genetic Programming. CEC03-Congress on Evolutionary Computation, IEEE Press: Canberra, 2003: 8–12.

[11] H. SHAH, J. UNDERCOFFER, A. JOSHI. Fuzzy Clustering for Intrusion Detection.Proceedings of The IEEE International Conference on Fuzzy Systems, 2003: 1274–1278.

[12] Guandong Xu, Hohl L, Merialdo B, et al. Using probabilistic latent semantic analysis for Web page grouping. Research Issues in Data Engineering Stream data Mining and Applications,2005: 29–36.

[13] Gui SN,Feng BQ.A fuzzy integral method to merge search engine results on web.HAO Y. Computational Intelligence and Security, LNCS 3802.Xi'an China,2005,Part II , 731–736.

[14] M.N.Amhed, S.M.Yamany, N.Mohamed, etc. A modified fuzzy c-means algorithm for bias field estimation and segmentation of MRI data. IEEE trans.on Medical Imaging,2002, 21(3): 193–199.

[15] HongLan Jin, Kam-Fai Wong. A Chinese Dictionary Construction Algorithm for Information Retrieval. ACM Transactions on Asian Language Information Processing; 2002, 1(4): 281–296.

[16] G. Deschrijver, E. E. Kerre. On the relationship between some extentions of fuzzy set theory. Fuzzy Sets and Systems, 2003, 133: 227–235.

[17] Yo-Ping Huang, Hong-Jin Chen, Hung-chi Chu. Identifying a Fuzzy Model by Using the Bipartite Membership Functions. Fuzzy Sets and Systems, 2001, 118(2): 199–214.

[18] H. Bustince, P. Burillo. Mathematical analysis of interval-valued fuzzy relation: Application to approximate reasoning. Fuzzy Sets and Systems, 2000, 113: 205–219.