符保龙,张爱科.中心聚类和语义特征融合的网页信息文本挖掘方法[J].辽宁工程技术大学学报(自然科学版),2016,35(1):85-88. doi:10.11956/j.issn.1008-0562.2016.01.017
FU Baolong,ZHANG Aike.Text mining method of web information based on fusing center clustering and semantic features[J].Journal of Liaoning Technical University(Natural Science),2016,35(1):85-88. doi:10.11956/j.issn.1008-0562.2016.01.017

中心聚类和语义特征融合的网页信息文本挖掘方法

符保龙，张爱科
（柳州职业技术学院 电子信息工程系，广西 柳州 545006）

摘　要：针对网页信息内容丰富且结构复杂，难以准确挖掘的问题，采用中心聚类和语义特征相互融合的方法.利用中心聚类算法确定样本最终的聚类中心，根据每个词在网页中出现的频率和词的上下文语义，构造一个网页-词语的权重映射矩阵，并将语义特征作为中心聚类相似性的判断依据，完成网页文本信息的挖掘.实验结果表明：利用该方法对网页文本进行挖掘，在时间增加不多的情况下，可以获得更高的召回率和准确率.

关键词：中心聚类；语义特征；矩阵；网页信息；文本挖掘

中图分类号：TP 311.13　　**文献标志码**：A　　**文章编号**：1008-0562(2016)01-0085-05

Text mining method of web information based on fusing center clustering and semantic features

FU Baolong, ZHANG Aike
（Electronic Information Engineering Department,Liuzhou Vocational Technological College, Liuzhou 545006, China）

Abstract: It is difficult to get accurate data mining due to the rich information contents and the complex structure of webpage. The mining issues can be solved by using the method of mutual integration on central clustering and semantic features. First determining the final cluster centers samples using center clustering algorithm, then according to the each word frequency and semantic in the webpage to construct a web page-right words remapping matrix, and finally using semantic features to judge the similar of center cluster and completing text mining information of the pages. The experimental results show that this method of web text mining can obtain higher rate of recall and precision with the same amount of time.

Key words: center clustering, semantic feature, matrix ,web information, text mining

0　引　言

21 世纪以来，人类社会信息化的进程进一步加快.难以计数的信息呈现在人们的面前，网络已经成为最大的信息储存库，仅中国的网页数量就超过数百亿[1].如此激增的信息量，给人们从中筛选并提取对自己有价值的信息提出了挑战，从而也推动数据挖掘技术飞速发展[2].在各种数据挖掘技术中，网页上的数据挖掘技术成为最重要的分支之一[3].网页上的数据文本信息居多，并且符合一定的结构化特征，因此早期的网页文本信息挖掘技术主要针对结构化数据提出[4].但是，随着网页信息形式的丰富化，半结构化数据开始大量出现，使得传统的网页文本信息挖掘方法难以实用[5].在这种情况下，根据网页中原始的的数据形式，挖掘其中的潜在知识表达，成为实现准确网页文本信息挖掘的前提[6].以此为切入点，出现了网页文本信息聚类技术.从现有的网页文本聚类方法来看，形成“倒排序”聚类方法和“后缀树”聚类方法两大类型[7-8]，这两类方法都是根据文档到词语的关系，构建有针对性的挖掘模型.但是只从词语的文字形式上进行挖掘，往往存在很大的出错概率[9].这是因为，无论是中文文本还是英文文本，都存在一词多义、同义、近义等现象，不考虑文本语义特征的挖掘显然会导致信息挖掘的失效[10-11].

因此，对于网页文本信息的挖掘技术来讲，考察文本信息的语义特征至关重要.在这种思路的指

收稿日期：2014-11-05
基金项目：广西教育厅科研项目基金项目（201106LX745，201204LX593）
作者简介：符保龙（1978-），男，广西 龙州人，硕士，副教授，主要从事数据挖掘、演化计算等方面的研究. **本文编校**：焦丽

引下，本文将语义特征提取和一种中心聚类算法融合在一起，提出一种新的网页信息文本挖掘方法，以期得到更好的网页文本信息挖掘性能.

1 中心聚类融合语义特征的挖掘方法

在网页信息的挖掘过程中，挖掘算法面对的主要信息就是文本，文本挖掘中比较实用的方法是聚类算法.所谓文本聚类，实际上就是把一个文本信息构成的信息集合，执行内容上的分组处理.经过这个分组应该达到的情况是，组内的文本信息从内容上高度相似，而组间文本信息的关联性则应该尽可能的低[12].所以，文本聚类方法的实现实际上包含两个最重要的环节，一个就是选取用于判断相似性的特征，另一个就是如何利用这些特征去判断相似性.本文提出的中心聚类和语义特征融合的、面向网页信息的文本挖掘方法，就是要从这两个环节设计出更加有效的手段.

1.1 中心聚类划分

如果对现有的文本聚类方法进行一个分类，大概可以分成 5 大类方法：第一种是基于划分的文本聚类方法，第二种是基于层次的文本聚类方法，第三种是基于密度的文本聚类方法，第四种是基于网格的文本聚类方法，第五种是基于模型的文本聚类方法.在这 5 种方法中，基于划分的方法原理简单并且易于实现，算法的复杂度也比较低，因此应用也最为广泛.

K 均值聚类方法就是一种典型的基于划分的文本聚类方法，这种方法的思路是：在原始文本信息集合中，随机选取 k 个文本信息作为这个集合的中心，之后计算其它文本信息和这 k 个中心的距离并比较这些距离的远近，和哪个中心的距离最近，就将这个文本信息归入到以最近中心为标志的分类之中.因为 K 均值聚类方法和它的变形方法，都是依托分类中心展开聚类的，因此也称之为基于中心的聚类.

为了便于说明基于中心聚类的文本挖掘方法，给出如下的实施过程.

步骤一 在给定的原始文本信息构成的集合中，随机选取 k 个文本信息作为几个分类的中心，这样得到的 k 个中心可以表示为 $(C_1,\cdots,C_i,\cdots,C_k)$.

步骤二 将其余的文本信息逐一和这 k 个中心进行相似性的判断，这种判断最常用的方法就是距离相似性，其数学描述式为

$$d_i = \left|T_i - C_i\right|, \qquad (1)$$

式中，d_i 为第 i 个文本和第 i 个分类中心的距离，T_i 为第 i 个文本.

步骤三 各个分类形成后，检验第一步设定的 k 个分类中心是否是最优的.即在各个分类中用其它文本信息取代原有的信息中心 C_i，并判断新的中心 C_i' 是否是最优的.这个判断方法，一般是考察新中心形成的类内距离总和是否小于原中心形成的类内距离总和，数学描述为

$$\varepsilon = \sum_i (d_i' - d_i), \qquad (2)$$

公式中，判断 ε 和事先设定的阈值的大小，决定是否停止更新.

步骤四 不断重复步骤二和步骤三的过程，直到确定出最终的聚类中心和分类.

1.2 语义特征提取

对于网页文本信息而言，各个词语之间存在着密切的上下文联系，每个词语出现的频度也反映出文档关注内容的区别.因此，如果仅仅用词语的文字形式去判断如何进行分类，显然很难反映实际情况.所以，对于网页文本信息来讲，关注文本的语义特征，对于更加准确的信息挖掘有着更重要的意义.

假设一个用户在一定时期内对某个网站进行连续访问，那么就可以通过这些访问的历史纪录，来构建起网页访问的原始集合 $H=\{h_1,h_2,\cdots,h_m\}$ 和词语的集合 $W=\{w_1,w_2,\cdots,w_n\}$.进而，每个网页和这些词语的映射关系又可以表征出一个新的集合 $h_i=\{\beta_{i1},\beta_{i2},\cdots,\beta_{in}\}$，这个集合中每一个变量表示了对应词语对于网页的权重系数.这样，这个用户访问的所有页面和词语之间就会形成一个矩阵的形式，见式（3）.

$$\boldsymbol{HW}=\begin{bmatrix}\beta_{11} & \beta_{12} & \cdots & \beta_{1n}\\ \beta_{21} & \beta_{22} & \cdots & \beta_{2n}\\ \vdots & \vdots & & \vdots\\ \beta_{m1} & \beta_{m2} & \cdots & \beta_{mn}\end{bmatrix} \qquad (3)$$

实际上，这样一个矩阵就在很大程度上反映了网页和词语之间的关系，进而揭示了用户的兴趣所在.要揭示出这个矩阵中潜在的语义特征，可以按照下面的流程对其进一步分解，实现数学参数化，易于从数学形式上加以刻画.

语义特征提取的数学描述式为.

$$\begin{aligned} h_i' &= h_i VS = (h_{i1},h_{i2},\cdots,h_{ik})\\ h_{ij} &= \sum_{k=1}^{n}\beta_{ik}\,p(w_j)\end{aligned} \qquad (4)$$

式中，h_i' 为最终分解出来的语义特征，$p(w_j)$ 为词语 w_j 出现的概率.

1.3　网页文本挖掘

经过前面两小节的工作，已经基本明确了本文文本挖掘方法中两个关键环节的操作.下面，进一步将这些工作整合化，构建一个完整的网页信息文本挖掘方法，此方法的具体操作步骤如下：

步骤一　构建网页-词语矩阵 $\boldsymbol{HW}$，设定后续进行相似性判断的阈值 ε，估计潜在的语义特征数目 k.

步骤二　对网页页面信息资源执行预处理工作，并根据第一步构造的网页-词语矩阵 $\boldsymbol{HW}$，计算矩阵中每个对应位置的权重.

步骤三　根据式（4）对网页-词语矩阵 $\boldsymbol{HW}$ 进行处理，求取出所有的语义特征向量 h_i'.

步骤四　在每一个语义特征向量 h_i' 中设定一个出示聚类中心 C_i，那么全部 n 个语义特征向量形成的聚类中心集合为 $C=\{C_i\}$.

步骤五　对集合 $C=\{C_i\}$ 中的所有的聚类中心执行两两之间的相似性估算.

步骤六　将聚类中心两两之间的相似性数据 $\{S_i\}$ 执行阈值判断，如果相近两个聚类中心之间的相似性 $S_i<\varepsilon$，则将这两个聚类中心执行合并处理.

步骤七　重新整理合并后的聚类中心集合，形成新的聚类中心集合 C'.

2　实验与分析

为了验证本文方法的性能，从新浪、网易等门户网站中随机选取 8 种类型的网页（每类 200 张），作为网页信息文本聚类的实验数据.这 8 种类型的网页分归可以纳入下列的主题，体育主题类网页、娱乐主题类网页、教育主题类网页、旅游主题类网页、军事主题类网页、政治主题类网页、经济主题类网页、科技主题类网页.为了确保验证过程的可信度，进一步将这 1600 张网页随机划分成 4 个数据集，每类数据集中包含的网页类型和数量见表 1.

表 1　4 类数据集的划分

Tab.1　classification of 4 kinds of data sets

数据集序号	包含网页种类	详细分类信息
第一个	5 类	体育(80)、娱乐(100)、军事(50)、政治(50)、科技(50)
第二个	6 类	教育(80)、旅游(125)、军事(50)、政治(50)、经济(120)、科技(50)
第三个	7 类	体育(60)、娱乐(50)、教育(80)、军事(50)、政治(50)、经济(40)、科技(50)
第四个	8 类	体育(60)、娱乐(50)、教育(40)、旅游(75)、军事(50)、政治(50)、经济(40)、科技(50)

为了便于形成和其它方法的比较，选取基于划分方法中的 K 均值聚类方法、基于层次方法中的平衡迭代聚类方法，同本文构建的中心聚类和语义特征融合的网页信息文本挖掘方法一起进行实验.对于本文的方法，相似性阈值 ε 设定为 3，语义特征数目 k 设定为 8.实验过程中，表 1 中的 5 个数据集合全部被执行聚类实验.先以第二个数据集合为例，给出其详细的实验数据，见表 2.

从表 2 中的实验数据可以看出：K 均值聚类方法的迭代次数最少，但召回率、准确率和 F 评价值都比较低；本文提出的方法迭代次数最多，但召回率、准确率和 F 评价值都是三种方法中最高.

表 2　第二数据集合的实验数据

Tab.2　experimental data of second data set

方法	迭代次数	召回率						方法	F 评价值	准确率					
		教育	旅游	军事	政治	经济	科技			教育	旅游	军事	政治	经济	科技
K 均值	10	0.75	0.52	0.81	0.85	0.67	0.90	K 均值	0.65	0.64	0.58	0.72	0.81	0.75	0.83
平衡迭代	14	0.81	0.74	0.73	0.66	0.85	0.84	平衡迭代	0.69	0.65	0.60	0.71	0.74	0.77	0.80
本文方法	18	0.82	0.80	0.83	0.87	0.91	0.86	本文方法	0.85	0.81	0.79	0.84	0.85	0.87	0.83

分别针对表 1 中的 4 类数据集合都执行 K 均值聚类方法、平衡迭代聚类方法和本文提出的方法，各执行 10 次，取 10 次的平均 F 评价值和迭代次数，对比情况列在表 3 中.

表 3　3 种方法在 5 类数据集合上执行的效果对比

Tab.3　comparison of 3 methods implemented on 5types of data sets

实验数据	K 均值聚类方法		平衡迭代聚类方法		本文方法	
	迭代次数	F 评价值	迭代次数	F 评价值	迭代次数	F 评价值
第一集合	6	0.63	11	0.68	13	0.88
第二集合	10	0.65	14	0.69	18	0.85
第三集合	7	0.71	12	0.70	15	0.87
第四集合	9	0.68	13	0.72	16	0.83

从全部 4 类测试数据集合来看，本文提出的方法都获得了网页信息文本挖掘更高的 F 评价值，当然迭代次数也比 K 均值聚类和平衡迭代聚类方法要高些.

进一步比较一下随着参与挖掘的网页数量增多，三种方法的用时情况，见图 1.

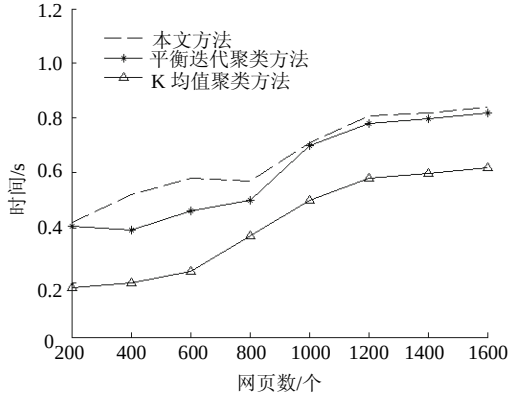

图 1 三种方法的时间比对结果

Fig.1 comparison results of using time in three methods

从图 1 中的曲线对比情况可以看出，本文提出的方法在用时上是最多的，但和平衡迭代方法的用时差距并不明显.考虑到本文方法在数据挖掘过程中召回率和准确率上的优势，所多花费的时间是可以接受的.分析本文方法用时最多的原因，在于本文方法在进行文本挖掘的过程中，构造了网页-词语矩阵及相应的预处理工作，加之在聚类过程中对于语义特征的运用.

3 结论

本文针对网页信息的文本挖掘方法展开了研究.为了提高挖掘方法的准确性，在一般聚类方法的基础上考虑引入语义特征.在求取语义特征时，考察了词语对于网页的重要性权重，进而得到全部网页和关键词语的映射矩阵.获得语义特征矩阵之后，在基于中心聚类方法的框架之下，运用语义特征作为相似性判断的依据，完成对网页信息的文本挖掘.在整个方法之中，网页-词语矩阵的构建及其在之后预处理中的应用，体现了对网页信息文本挖掘的针对性；而第四步中基于语义特征构建聚类中心，则使得中心聚类方法和语义特征的运用紧密地结合在一起.通过同其它两种方法的比对实验发现，本文提出的文本挖掘方法具有更高的召回率、更高的准确率.同时，其挖掘过程所增加的时间消耗也在可以接受的范围之内.

参考文献：

[1] BEGEMAN G,KELLER P,SMADJIA F.Automated tag clustering: improving search and exploration in the tag space[C].In:Collaborative Web Tagging Workshop,15th International World Wide Web Conference, Edinburgh,UK,2006(5):22-26.

[2] 郭景峰,赵玉艳,边伟峰,等.基于改进的凝聚性和分离性的层次聚类算法[J].计算机研究与发展,2008, 45 (S1):202-206.
GUO Jingfeng,ZHAO Yuyan,BIAN W F,et al.A hierarchical clustering algorithm based on improved cluster cohesionand separation [J].Journal of Computer Research and Development,2008,45(S1):202-206.

[3] 刘一鸣,张化祥.引入信息增益的层次聚类算法[J].计算机工程与应用,2012,48(1):142-144.
LIU Yiming,ZHANG Huaxiang.New hierarchical clustering method using information gain[J].Computer Engineering and Application, 2012,48(1):142-144.

[4] CHOW TOMMY W S,ZHANG Haijun,Rahman M K M.A new document representation using term frequency and vectorized graph connectionists with application to document retrieval[J].Expert Systems with Applications,2009,36(10):12 023-12 035.

[5] 张玉芳,朱俊,熊忠阳.改进的概率潜在语义分析下的文本聚类算法[J].计算机应用,2011,3(31):674-676.
ZHANG Yufang,ZHU Jun,XIONG Zhongyang.Improved text clustering algorithm of probabilistic latent with semantic analysis[J].Journal of Computer Applications,2011,3(31):674-676.

[6] 熊忠阳,暴自强,李智星.结合 LSA 的中文谱聚类算法研究[J].计算机应用研究,2010,27(3):917-918.
XIONG Zhongyang,BAO Ziqiang,LI Zhixing.Research of chinese spectral clustering with LSA[J].Application Research of Computers, 2010,27(3):917-918.

[7] 毛婷,杨敬辉,杨晶东.基于模糊聚类的自然语言语义特征[J].辽宁工程技术大学学报(自然科学版),2013,32(1):81-84.
MAO Ting,YANG Jinghui,YANG Jingdong.Semantic Feature of natural langurage based on fuzzy cluster[J].Journal of Liaoning Technical University(Natural Science),2013,32(1):81-84.

[8] 王秀慧,王丽珍,麻淑芳.结合语义的改进 FTC 文本聚类算法[J].计算机工程与设计,2014,35(2):515-519.
WANG Xiuhui,WANG Lizhen,MA S F.Improvement On FTC text clustering algorithm combined with semantics[J].Computer Engineering and Design,2014,35(2):515-519.

[9] 何祥,骆祥峰.基于关联语义链网络的文本聚类方法[J].上海大学学报(自然科学版),2013,20(2):11-19.
HE Xiang,LUO Xiangfeng.Document clustering method based on association link network[J].Journal of Shanghai University(Natural Science Edition,) 2013,20(2):11-19.

[10] 王永贵,林琳,刘宪国.结合双粒子群和 K-means 的混合文本聚类算法[J].计算机应用研究,2014,31(2):364-368.
WANG Yonggui,LIN Lin,LIU Xianhuo.Hybrid text clustering algorithm based on dualparticle swarm optimization and K-means algorithm[J].Applicaion Resrarch of Computers,2014,31(2):364-368.

[11] 马素琴,施化吉.阈值优化的文本密度聚类算法[J].计算机工程与应用,2011,10(37):134-136.
MA Suqin,SHI Huaji.Text denity clustering algorithm with optimized threshold values[J].Computer Engineering and Application,2011,10(37): 134-136.

[12] 魏桂英,高学东,武森.基于领域本体的个性化文本信息检索[J]. 辽宁工程技术大学学报(自然科学版),2011,32(1):316-320.
WEI Guiying,GAO Xuedong,WU Sen.Individualized text information retrieval based on domain ontology[J].Journal of Liaoning Technical University(Natural Science),2011,32(1):316-320.