基于词条与语意差异度量的文档聚类算法

魏霖静[1] 练智超[2] 王联国[1] 侯振兴[3]
(甘肃农业大学信息科学技术学院 兰州 730070)[1] (南京理工大学计算机科学与工程学院 南京 210094)[2]
(南京大学信息管理学院 南京 210093)[3]

摘 要 已有的文本聚类算法大多基于一般的相似性度量而忽略了语义内容,对此提出一种基于最大化文本判别信息的文本聚类算法。首先,分别分析词条对其类簇与其他类簇的判别信息,并且将数据集从输入空间转换至差异分数矩阵空间;然后,设计了一个贪婪算法来筛选矩阵每行的低分数词条;最终,采用最大似然估计对文本差别信息进行平滑处理。仿真实验结果表明,所提方法的文档聚类质量优于其他分层与单层聚类算法,并且具有较好的可解释性与收敛性。

关键词 文档聚类,语意分析,贪婪算法,收敛性,可解释性
中图法分类号 TP301.6 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.12.042

Term and Semantic Difference Metric Based Document Clustering Algorithm

WEI Lin-jing[1] LIAN Zhi-chao[2] WANG Lian-guo[1] HOU Zhen-xing[3]
(School of Information Science and Technology, Gansu Agriculture University, Lanzhou 730070, China)[1]
(School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China)[2]
(School of Information Management, Nanjing University, Nanjing 210093, China)[3]

Abstract The existing document clustering algorithms are based on the common similarity measurement, but ignore the semantics. So a document clustering algorithm based on maximizing the sum of the discrimination information provided by documents was proposed. Firstly, the discrimination information of term for the corresponding cluster and for the other clusters was analyzed separately, and the data set was transformed from input space to the difference scores matrix space. Then a greedy algorithm was designed to filter the terms with low score from each row of the matrix. Lastly, maximum likelihood estimation was used to smooth the document difference information. Simulation experiment results show that the proposed method has better cluster quality than the plat and hierarchical clustering algorithms, and has a good quality in interpretability and convergence.

Keywords Document clustering, Semantic analysis, Greedy algorithm, Convergence, Interpretability

1 引言

随着互联网的蓬勃发展,网络已成为工作、生活的一部分,在各种数据库中生成了大量的文档内容。为了高效地获得文档数据集的关键信息,必须先对文档进行聚类处理,文档聚类不同于文档分类,后者对于每个类别已经预先做了类别标注,而文档聚类则无预先标注,并且根据文档内容将文档集合分为若干个类簇,同一类簇内的文档内容应该极为相似,类簇之间则应存在较大差异[1,2]。

许多研究均关注数据集关联信息的分析,而数据之间的判别信息作为一个重要的核心概念并未得到深入的利用。目标数据集判别信息的度量方法主要来自于统计学理论与信息理论,常用的度量方法主要有危险度比、优势比、信息增益、相对熵等,此类度量方法均基于语料库。在生物医药领域的研究中,危险度比与优势比被广泛使用,以进行一致性度量与分析[3,4]。在文本处理领域中,此类度量方法广泛应用于特征选择问题[5],近期的一些文本处理研究中,将危险度比与信息增益用于量化判别信息[6,7],此类研究使用词条判别信息建立学习模型,显示出较好的效果。

文献[8]讨论了词条判别信息的语意,该文提出一个理论框架来估算词条之间的相关性,并分析了该相关性与词条对文档集中各类别支持度的关系。文献[9]测量了词条间关联性的SDC(语意判别能力),并较好地解决了文档聚类问题。在语言心理学领域中发现,人们将词条与相关上下文或主题关联的可能性大于将不同词条进行关联的可能性[10],因此,文本分析中词条之间的语意关系仅能表示出类别上下文的无向信息,而使用词条判别信息指导文档的聚类程序具有较好的语意表达效果。

到稿日期:2016-03-03 返修日期:2016-03-23 本文受国家自然科学基金项目(034031122,61063028),江苏省自然科学基金青年基金(BK20150784),中国博士后面上资助(2015M581800),甘肃省科技支撑计划项目(1604WKCA011),陇原青年创新创业人才项目(2016-47)资助。
魏霖静(1977-),女,博士后,副教授,主要研究方向为智能计算、算法应用研究、生物信息学,E-mail:wlj@gsau.edu.cn(通信作者);**练智超**(1983-),男,博士,副教授,主要研究方向为图像处理、模式识别、医学图像分析、智能算法;**王联国**(1968-),男,博士,教授,硕士生导师,主要研究方向为智能计算、算法应用研究;**侯振兴**(1977-),男,博士生,副教授,主要研究方向为人机交互与用户行为、算法应用研究。

文档聚类问题的难点主要有：1）词条-文档的空间维度较高；2）词条-文档的空间分布稀疏；3）合并文档的词条-文档语意较为困难；4）现实中的文档集合数据量极大，相应的处理算法应具有极高的效率。K-means 算法及其各种改进算法是目前广泛使用的文档聚类算法，K-means 算法效率高并且易于实现，但其类簇的可解释性较差，其类簇的意义向量往往难以表示文档的语意。文献[11-14]提出了分层的文档聚类方法，此类方法以不同的抽象级别将文档集合分组与泛化处理。另外一部分研究采用外部的知识库对文档描述符进行语意丰富化的处理[15,16]，此类方案聚类效果较好，但是从知识库提取信息的过程的计算成本较高，总体效率较低。另一类文档聚类方法的思想是结合聚类方法与降维技术[17]。

综上所述，本方法在低维空间使用迭代程序搜索类簇，在后期的词条选择过程中忽略了部分次要词条，从而自动实现了降维的效果。此外，本文方法是一种单向聚类方法，聚类结果中已经将每个类簇中的词条进行了排序，排序过程基于文档集合中的判别信息自动地生成，因此本文聚类结果具有较好的可理解性。

2 本文文档聚类算法

本文基于判别信息最大化的聚类算法是一个迭代的文档分组框架，将 M 维输入空间转换为 K 维的判别信息空间，然后搜索 K 个文本类别。设计了一个两步骤的程序：文档投影与文档分配，其中对文档的差异分数之和进行最大化处理。本文的类簇结果可通过差异词条所对应的文档上下文/主题描述。

2.1 问题模型

假设 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N]\in\mathcal{R}^{M\times N}$ 表示词条-文档矩阵，其中 $\boldsymbol{x}_i=[x_{1i},x_{2i},\cdots,x_{Mi}]^T$ 表示第 i 个文档（M 维向量），M 是 N 个文档中所有词条的总数量。文档 $\boldsymbol{x}_i$ 中词条 x_j 的权重表示为 x_{ji}，其值等于文档 $\boldsymbol{x}_i$ 中词条 x_j 的总数量。文档聚类的目标是寻找文档的 $K(K\ll\min\{M,N\})$ 个类簇 $C_k(k=1,2,\cdots,K)$，假设一个文档是 $\boldsymbol{x}\in C_k$，则 $\boldsymbol{x}\notin C_j,\forall j\neq k$。

2.2 聚类的目标函数

本方法通过最大化文档差异分数之和来搜索 K 个类簇。如果将文档 $\boldsymbol{x}_i$ 对于类簇 k 的判别信息表示为 d_{ik}，则文档 $\boldsymbol{x}_i$ 对于类簇 k 之外的所有类簇的判别信息表示为 $\bar{d}_{ik}$，将文档 $\boldsymbol{x}_i$ 对类簇 k 的差异分数定义为 $\hat{d}_{ik}=d_{ik}-\bar{d}_{ik}$。可将目标函数写为如下形式：

$$J=\sum_{k=1}^{K}\sum_{\boldsymbol{x}_i\in C_k}\delta_{ik}(d_{ik}-\bar{d}_{ik}) \tag{1}$$

如果文档 $\boldsymbol{x}_i$ 被分类至类簇 k，则式中 $\delta_{ik}=1$，否则为 0。从词条的判别信息可计算出文档的判别信息 $(d_{ik},\bar{d}_{ik})$，否则从当前标记的文档集中估算。

本方法寻找一个目标类簇，该类簇中文档对所属类簇的判别信息度高于对于其他类簇的判别信息。因此，仅仅最大化对其所属类簇的差异度是不够的，因为这些文档对其他类簇也具有较大的差异度。

本文使用两步骤的贪婪算法对目标函数 J 进行最大化处理。第一步：给定一个类簇分配策略，定义为 $\delta_{ik}(\forall i,k)$，估算（采用最大似然估计）标记文档集的 $d_{ik}(\forall i,k)$ 与 $\bar{d}_{ik}(\forall i,k)$ 以实现 J 的最大化。第二步：对于给定的估计差异分数 $\hat{d}_{ik}(\forall i,k)$，将每个文档分配至差异分数最高的类簇 k 以实现 J 的最大化。这两个步骤轮流执行直到两轮连续迭代的 J 值之差低于指定的阈值。

2.3 词条判别信息

通过词条的判别信息计算文档的判别信息。从标记的文档集中估算判别信息，本文采用相对危险度量化判别信息。

2.3.1 相对危险度

相对危险度广泛地应用于生物医学研究中[18]，将类簇 k 中一个词条对其他类簇的相对危险度作为类簇 k 的判别信息。式(2)、式(3)分别计算了词条 x_j 对于类簇 k 的判别信息与词条 x_j 对 k 之外的所有类簇的判别信息：

$$w_{jk}=\begin{cases}\dfrac{p(x_j|C_k)}{p(x_j|\bar{C}_k)}, & \text{if } p(x_j|C_k)-p(x_j|\bar{C}_k)>t\\ 0, & \text{其他情况}\end{cases} \tag{2}$$

$$\bar{w}_{jk}=\begin{cases}\dfrac{p(x_j|\bar{C}_k)}{p(x_j|C_k)}, & \text{if } p(x_j|\bar{C}_k)-p(x_j|C_k)>t\\ 0, & \text{其他情况}\end{cases} \tag{3}$$

其中，$p(x_j|C_k)$ 是词条 x_j 出现在类簇 k 中的条件概率，$\bar{C}_k$ 表示 k 之外的所有类簇，t 是一个词条选择参数，用于控制词条的过滤条件（删除贡献较小的判别信息）。词条判别信息的值为 0 或者大于 1，值越大表示判别能力越大。

2.4 估计与平滑处理

上文使用最大似然估计从标记文档集合中估算判别信息，然而最大似然估计容易过度拟合，并且会产生 0 概率的估计。因此，需要对估计的结果进行适当的平滑处理以提高鲁棒性。因为本方法基于 1-gram 模型，所以本文采用 Good-Turning 估计器[19]。通过因子 $1-E(1)/T$ 降低非零的概率，而零概率为 $E(1)/T$，其中 $E(1)/T$ 是词条出现一次的期望数量，T 是语料库的大小。

2.5 词条对类簇的关联性

通过式(2)、式(3)中的词条选择参数可删除次要的词条。因为 t 值从 0 开始增加，所以仅有极少的词条会具有较高的判别信息。词条的序号集合包含了对类簇 $k(T_k)$ 有意义的词条信息，定义如下：

$$T_k=\{j|w_{jk}>0,\forall j\} \tag{4}$$

此类词条与其判别信息对于类簇 k 中的文档上下文具有较好的可解释性，可表示为：$T_k\cap T_j\neq\emptyset,\forall j\neq k$，可理解为有些词条对于多个类簇可能都具有较好的判别信息。而基于 t 值的变化，有些词条对所有的类簇均不能提供有意义的判别信息。

2.6 文档判别信息

假设一个文档为 $\boldsymbol{x}_i$，则通过该文档的词条描述该文档，文档中的每个词条 x_j 可根据词条判别信息的值 w_{jk} 表征类簇 k 的上下文，可理解为：文档中的每个词条 x_j 对于上下文或者类簇 k 均具有一定的相关度（w_{jk} 值）。可将文档 $\boldsymbol{x}_i$ 对类簇 k 的判别信息表示为类簇 k 的平均词条判别信息量：

$$d_{ik}=\frac{\sum_{j\in T_k}x_{ji}W_{jk}}{\sum_j x_{ji}} \tag{5}$$

其中，$\bar{d}_{ik}$ 的定义方式与式(5)相似。可将文档的判别信息 d_{ik} 理解为文档 $\boldsymbol{x}_i$ 对类簇 k 的关联性。文档差异分数的计算方

法为：$\hat{d}_{ik}=d_{ik}-\bar{d}_{ik}$，分数越高，文档 x_i 属于类簇 k 的可能性越大。如果该词条在文档中出现多次，那么每次的出现对判别信息均有所贡献，因此差异词条出现的频率越高，则文档对于指定类簇的判别信息越大。

2.7 本文算法

假设 $\boldsymbol{W}(\bar{\boldsymbol{W}})$ 是 $M\times K$ 的矩阵，其元素为 w_{jk}（$\bar{w}_{jk}$）（$\forall j,k$），$\boldsymbol{D}$ 设为 $N\times K$ 的矩阵，其元素为 $\hat{d}_{ik}$（$\forall i,k$），$\boldsymbol{R}$ 设为 $N\times K$ 的矩阵，其元素为 δ_{ik}（$\forall i,k$）。首先，基于余弦相似性将所有文档划分至 K 个随机选择的种子，定义为矩阵 $\boldsymbol{R}$。然后，执行一个两步骤的循环体：第一步，从词条-文档矩阵 $\boldsymbol{X}$ 与矩阵 $\boldsymbol{R}$ 中估算词条判别信息矩阵（$\boldsymbol{W}$ 与 $\bar{\boldsymbol{W}}$）；第二步，将所有文档转换至差异分数空间来创建差异分数矩阵 $\boldsymbol{D}$。该转换程序如下所示：

$$\boldsymbol{D}=(\boldsymbol{X}\sum)^{\mathrm{T}}(\boldsymbol{W}-\bar{\boldsymbol{W}}) \tag{6}$$

其中，$\sum$ 是一个 $N\times N$ 的正交矩阵，元素定义为 $\sigma_{ii}=1/\sum_j x_{ji}$。矩阵 $\boldsymbol{D}$ 代表了 K 维差异分数空间的文档。

将文档基于其差异分数重新分配至各类簇。如果 $\hat{d}_{ik}>d_{ij}$（$\forall j\neq k$），则将该文档 x_i 分配至类簇 k，该操作的矩阵形式为：

$$\boldsymbol{R}=\mathrm{maxrow}(\boldsymbol{D}) \tag{7}$$

其中，maxrow 是一个运算：对 $\boldsymbol{D}$ 的每行进行运算，如果是最大值则返回 1，否则返回 0。式(6)、式(7)重复执行直至目标函数的绝对差异低于某个预设值。计算 $\boldsymbol{D}$ 中各行的最大值之和作为目标函数 J。

所提方法是 K-means 方法的一个变种，具体步骤如算法 1 所示，算法输出为最终的文档分配矩阵 $\boldsymbol{R}$ 与词条判别信息矩阵 $\boldsymbol{W}$。

算法 1 判别信息最大化的文档聚类

输入：词条-文档矩阵 **X**，分类数量 K

输出：文档分组矩阵 **R**，词条判别信息矩阵 **W**

1. $\mathbf{R}^{(0)}$ = 文档初始化类簇分配；
2. $\tau=0$；
3. $\mathrm{J}^{(0)}=0$；
4. DO {
5. 词条判别信息估算 $\mathbf{W}^{(\tau)}$，$\bar{\mathbf{W}}^{(\tau)}$ //式(2)、式(3)
6. $\mathbf{D}^{(\tau+1)}=(\mathbf{X}\sum)^{\mathrm{T}}(\mathbf{W}(\tau)-\bar{\mathbf{W}}(\tau))$；
7. $\mathbf{R}^{(\tau+1)}=\mathrm{maxrow}(\mathbf{D}^{(\tau+1)})$；
8. 将 $\mathbf{D}^{(\tau+1)}$ 每行最大区别分数之和赋给 $\mathrm{J}^{(\tau+1)}$；
9. $\tau=\tau+1$；
10. } WHILE ($|\mathrm{J}^{(\tau)}-\mathrm{J}^{(\tau+1)}|<\varepsilon$).

2.8 本算法的优势分析

(1)本方法的时间复杂度是 $O(KMNI)$，其中 I 是总共所需的迭代次数，因此本算法的计算时间基于聚类参数的线性变化。

(2)本方法无需文档之间相似性的度量。将文档提取至 K 维差异分数空间中，将文档在对应坐标轴的值作为文档对类簇的关联性，该值越大表示文档对类簇的关联性越高。

(3)本方法除了输出文档的分类矩阵外，还输出了词条判别信息矩阵（$\boldsymbol{W}$）。$\boldsymbol{W}$ 描述了每个类簇的词条判别信息，该信息对于理解每个类簇中的文档上下文具有重要的作用。

(4)本方法通过词条选择参数 t 对词条进行过滤处理。增加 t 的值，可将判别信息较小的词条（潜在噪声）移除以进一步提高聚类性能。

2.9 本方法的收敛性分析

本文方法是一个两步骤迭代算法，与 K-means 算法类似，在 K-means 中，使用类簇的均值或者质心代表其类簇，而本文方法则使用具有语意的差异词条代表类簇。下面证明本方法每步单调地提高目标函数值，从而证明本方法的收敛性。

定义 1（本算法收敛至局部最大值） 考虑到本方法的目标函数式(1)表示文档对于各类簇区别分数之和，文档 x_i 对类簇 k 的区别分数是 $\hat{d}_{ik}=d_{ik}-\bar{d}_{ik}$。目标函数可转换为 K 个类簇的文档差异分数之和：

$$J=\sum_{x_i\in C_1}\delta_{i1}(\hat{d}_{i1})+\sum_{x_i\in C_2}\delta_{i2}(\hat{d}_{i2})+\cdots+\sum_{x_i\in C_K}\delta_{iK}(\hat{d}_{iK}) \tag{8}$$

假设 $C_1,C_2,\cdots,C_k$ 表示当前的类簇，目标函数值是 J_t。基于当前标记情况，计算所有的差异分数，然后寻找一个需要重新标记的文档 $x_n\in C_i$，即：

$$\hat{d}_{nj}>\hat{d}_{ni},\ \forall i\neq j \tag{9}$$

由此说明文档 n 对类簇 j 的差异分数高于其他的类簇。根据本文的差异分数计算方法，将文档 n 分配至类簇 j。分配结果按 $\hat{d}_{ni}$（类簇 i 的差异分数）递减排序，而按照 $\hat{d}_{nj}$（类簇 j 的差异分数）递增排序。

综合所有类簇的差异分数，生成新的目标函数值 J_{t+1}（下一轮迭代），由式(9)的条件，有：

$$J_{t+1}>J_t \tag{10}$$

文档 n 从类簇 i 到类簇 j 的转移产生了 3 种可行的词条类型。

(1)类簇 j 中权重增加的文档词条 n：假设该词条子集为 S_j；

(2)类簇 j 中权重减少的文档词条 n：假设该词条子集为 S_i；

(3)权重未改变的文档词条 n。

根据式(10)。可得：

$$\sum_{a\in S_j}w_{aj}>\sum_{a\in S_i}w_{ai} \tag{11}$$

因此，可能存在少量权重降低的词条，但是总体而言，大多词条的权重是增加的，因此对于下一次迭代有：$J_{t+2}\geqslant J_{t+1}$。

由于 J 的上界是所有可行聚类解的差异分数之和，因此本方法是收敛的。

3 实验结果与分析

为了评估本文算法的性能，建立了 3 组实验：1)将本方法与多个文档聚类方法对比实验，评估本方法的聚类性能；2)分析本文方法聚类结果的可理解性；3)分析本方法中低价值词条的过滤效果以及本方法的收敛性能。

3.1 实验的文档数据集

本文选择 11 个标准的公共文本数据集作为实验对象，11 个数据集的大小、上下文、复杂度均各不相同，表 1 总结了 11 个数据集的特点。数据集 1 来自于 Cornell University (http://www.cs.cornell.edu/People/pabo/movie-review-data/)，数据集 2～11 来自于 Karypis Lab(http://glaros.dtc.

umn. edu/gkhome/cluto/cluto/download)。

表 1 实验数据集的简要介绍

编号	名字	文档数量	词条数量	类别数量
1	movie	1200	38408	2
2	classic	7094	41681	4
3	reviews	4069	23220	5
4	hitech	2301	13170	6
5	tr31	927	10128	7
6	tr41	878	7454	10
7	ohscal	11162	11465	10
8	re0	1504	2886	13
9	wap	1560	8460	20
10	rel	1657	3758	25
11	reuters	8293	18933	65

3.2 对比文献

本方法是单层的聚类算法，为了充分地评估本算法的性能，将其与单层、多层聚类算法均进行比较实验。单层算法的对比文献选择基于 SC(spectral 聚类)[20]与基于 NMF 方法的 ENMF[21]，而分层算法的对比文献则选择 Rank-2 NMF[22]。文献[20]的方法使用公开的实现程序(http://alumni. cs. ucsb. edu/wychen/sc. html.)，为了接近本文的词条-文档矩阵，文献[20]的方法采用乘法更新规则与欧氏距离。文献[22]的实现程序可从(http://cogsys. imm. dtu. dk/toolbox)获得，设置其样本数量等于语料库的大小，sigma 设为 50。

3.3 聚类有效性评价

3.3.1 F-measure

本文采用 F-measure 评估算法的聚类性能。假设 L_i 与 C_j 分别代表第 i 个类别与第 j 个类簇，则对应的 F-measure 计算方法为：

$$Presision(L_i,C_j)=\frac{n_{ij}}{C_j}$$

$$Recall(L_i,C_j)=\frac{n_{ij}}{L_i} \tag{12}$$

$$F(L_i,C_j)=2\times\frac{Precision(L_i,C_j)\times Recall(L_i,C_j)}{Precision(L_i,C_j)+Recall(L_i,C_j)}$$

其中，n_{ij} 是类簇 C_j 中类别 L_i 的成员数量。F-measure 量化一个聚类解(设为 C)质量的方法可总结为：

$$F(C)=\sum_{L_i\in L}\frac{|L_i|}{|D|}\max_{C_j\in C}F(L_i,C_j) \tag{13}$$

其中，L 表示所有的类别，C 表示类别 L_i 中文档的数量，$|D|$ 表示整个数据集中文档的总数量。$F(C)$的范围是[0,1]，值越大表示分类准确率越高。

3.4 实验结果与分析

3.4.1 聚类质量结果与分析

实验中对每组实验独立运行 30 次，统计 30 次结果的均值与标准偏差。表 2 比较了本方法与其他两个单层聚类算法的聚类质量，从中可看出，本方法对 10 个数据集的聚类性能均优于其他两种算法。表 3 比较了本方法与分层聚类算法的聚类质量，实验中设置了 4 个类簇数量，分别为 3,15,30,60，统计了各方法获得的 F-measure 值，最后一行统计了对应方法所有的平均值。从表 3 可看出，本方法明显优于 Rank-2-NMF 算法。

表 2 3 种单层文档聚类方法的聚类质量统计

数据集	本方法	ENMF	SC
movie	**0.556±0.02**	0.510±0.01	0.519±0.01
classic	0.668±0.08	0.512±0.03	**0.717±0.01**
reviews	**0.675±0.07**	0.552±0.03	0.425±0.00
hitech	**0.442±0.03**	0.399±0.02	0.408±0.01
tr31	**0.613±0.09**	0.362±0.03	0.443±0.01
tr41	**0.578±0.06**	0.361±0.04	0.310±0.00
ohscal	**0.428±0.04**	0.250±0.02	0.300±0.01
re0	**0.411±0.02**	0.345±0.02	0.323±0.01
wap	**0.445±0.02**	0.299±0.02	0.411±0.01
rel	**0.393±0.04**	0.301±0.03	0.299±0.01
reuters	**0.349±0.04**	0.280±0.03	0.219±0.01

表 3 本方法与分层聚类方法的聚类质量比较

数据集	类簇数量	本方法	Rank-2-NMF
classic	3	**0.63**	0.58
	15	**0.66**	0.42
	30	**0.55**	0.39
	60	**0.46**	0.23
	均值	**0.58**	0.41
hitech	3	**0.55**	0.48
	15	**0.52**	0.45
	30	**0.46**	0.33
	60	**0.47**	0.23
	均值	**0.50**	0.37
re0	3	**0.48**	0.40
	15	**0.48**	0.39
	30	**0.47**	0.32
	60	**0.40**	0.32
	均值	**0.46**	0.36
reuters	3	0.49	**0.57**
	15	0.58	**0.60**
	30	0.54	**0.56**
	60	**0.52**	0.45
	均值	0.53	**0.55**
wap	3	**0.39**	0.36
	15	**0.60**	0.25
	30	**0.61**	0.19
	60	**0.53**	0.20
	均值	**0.53**	0.25
总计	均值	**0.52**	0.39

3.4.2 聚类的解释性

数据聚类的一个关键应用是对语料库的理解。在文档聚类的领域，聚类方法输出的信息应当可较好地用于类簇与文档的合并，本方法基于词条判别信息，因此每个类簇具有较好的可解释性。

表 4 ohscal 数据集的各类簇差异词条的前 10 名

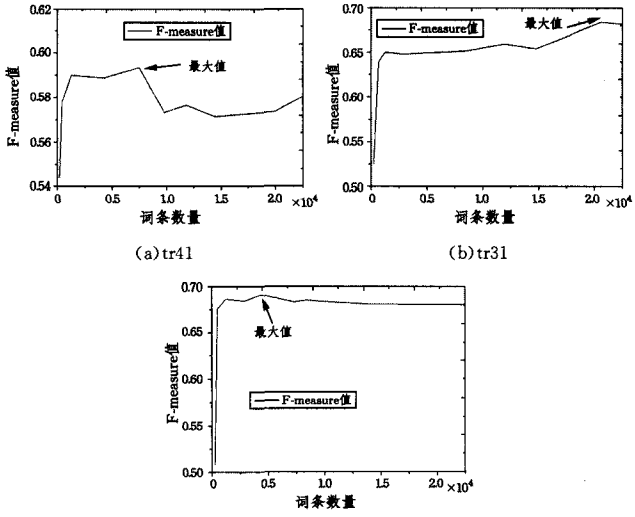
类	第 k 个类簇的差异词条前 10 名
1(Mol.)	platelet, kg, mg, dose, min, plasma, pressur, flow, microgram, antagonist
2(Canc.)	carcinoma, tumor, cancer, surviv, chemotherapi, stage, recurr, malign, resect, therapi
3(Anti.)	antibodies, antigen, viru, anti, infect, hiv, monoclon, ig, immun, sera
4(Prog.)	patient, complic, surgeri, ventricular, infarct, oper, eye, coronari, cardiac, morta
5(Preg.)	pregnanc, fetal, gestat, matern, women, infant, deliveri, birth, labor, pregnant
6(Risk.)	risk, alcohol, age, children, cholesterol, health, factor, women, preval, popul
7(DNA)	gene, sequenc, dna, mutat, protein, chromosom, transcript, rna, amino, structur
8(In-V.)	contract, muscle, relax, microm, calcium, effect, respons, antagonist, releas, action
9(Rece.)	il, receptor, cell, stimul, bind, growth, gamma, alpha, insulin, 0
10(Tomo.)	ct, imag, comput, tomographi, scan, lesion, magnet, reson, cerebr, tomograph

表4列出本方法对ohscal数据集每个类簇判别信息最高的前10名的词条，ohscal数据集包括10个不同的主体(antibodies, carcinoma, DNA, in vitro, molecular sequence data, pregnancy, prognosis, receptors, risk factors, tomography)。观察表中最高的10个词条：类2＝carcinoma，类3＝antibodies，类4＝prognosis，类5＝pregnancy，类6＝risk factors，类7＝DNA，类9＝receptors，类10＝tomography，可以简单地确定大多数类簇的类别。因为本方法在K维的判别信息空间搜索类簇，类簇中文档的分布可通过简单的散点图实现可视化。

3.4.3 词条区别信息的分布与词条选择性能

词条判别信息有一个较长、较窄的轨迹，换句话说，小部分的词条可包含较高的判别能力。因此通过参数t选择词条可降低本方法的空间复杂度，并且对类簇质量的影响较小。

实验中评估了聚类质量随着参数值t增加的变化情况。图1所示是本方法聚类质量的变化情况，可看出聚类质量并未随着词条数量的大幅度下降而衰减，相反，随着选择词条的数量的增加，聚类质量有所提高。实验结果显示本方法对词条选择是可扩展的并且是鲁棒的。



(a)tr41　(b)tr31

(c)classic

图1　词条选择对聚类质量的影响

3.4.4 本文方法的收敛性分析

迭代算法应具有平滑并且快速的收敛性，本方法通过两步骤的贪婪程序最大化其目标函数，并且确保目标函数在两个连续迭代之间是非下降的。

图2所示是本方法对5个数据集的收敛曲线，该图显示了本方法平滑的收敛过程，可以看出本方法在15次迭代之内即可收敛，可见本方法具有较快的收敛速度。

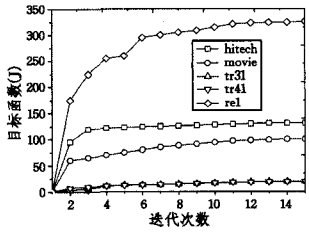


图2　本方法对5个数据集的收敛性

结束语　本文提出一个文档聚类算法，本方法在低维空间使用迭代程序搜索类簇，在后期的词条选择过程中忽略了部分次要词条从而自动地实现了降维的效果。此外，本文方法是一种单向聚类方法，聚类结果中已经将每个类簇中的词条进行了排序，排序过程基于文档集合中的判别信息自动生成，因此本文聚类结果具有较好的可理解性。本文从聚类质量、可解释性、算法的词条过滤以及算法收敛性立体地评估了本方法的性能，结果表明本方法的文档聚类质量优于其他分层与单层聚类算法，并且具有较好的可解释性与收敛性。

参考文献

[1] Zhao Wei-zhong, Ma Hui-fang, Li Zhi-Qing, et al. Efficiently Active Learning for Semi-Supervised Document Clustering[J]. Journal of Software, 2012, 23(6): 1486-1499(in Chinese)
赵卫中，马慧芳，李志清，等. 一种结合主动学习的半监督文档聚类算法[J]. 软件学报，2012，23(6)：1486-1499

[2] Liu Zhen-lu, Wang Da-ling, Feng Shi, et al. An Approach of Latent Semantic Space Partition and Web Document Clustering [J]. Journal of Chinese Information Processing, 2011, 25(1): 60-65(in Chinese)
刘振鹿，王大玲，冯时，等. 一种基于LDA的潜在语义区划分及Web文档聚类算法[J]. 中文信息学报，2011，25(1)：60-65

[3] Hsieh D A, Manski C F, Mcfadden D. Estimation of Response Probabilities From Augmented Retrospective Observations[J]. Journal of the American Statistical Association, 1985, 80(391): 651-662

[4] Junejo K N, Karim A. Robust personalizable spam filtering via local and global discrimination modeling[J]. Knowledge & Information Systems, 2013, 34(2): 299-334

[5] Mee C Y, Yun L J. A Corpus-based Approach to Comparative Evaluation of Statistical Term Association Measures[J]. Journal of the American Society for Information Science & Technology, 2001, 52(4): 283-296

[6] Junejo K N, Karim A. A Robust Discriminative Term Weighting Based Linear Discriminant Method for Text Classification[C]// Eighth IEEE International Conference on Data Mining, 2008 (ICDM'08). IEEE, 2008: 323-332

[7] Malik H H, Fradkin D, Moerchen F. Single pass text classification by direct feature weighting[J]. Knowledge & Information Systems, 2011, 28(1): 79-98

[8] Cai D. An Information-Theoretic Foundation for the Measurement of Discrimination Information[J]. IEEE Transactions on Knowledge & Data Engineering, 2010, 22(9): 1262-1273

[9] Xu Z, Luo X, Mei L, et al. Measuring the semantic discrimination capability of association relations[J]. Concurrency & Computation Practice & Experience, 2014, 26(2): 380-395

[10] Morris J, Hirst G. Non-classical lexical semantic relations[C]// Htl-naacl Workshop on Computational Lexical Semantics. 2004: 46-51

[11] Gil-Garcia R, Pons-Porrata A. Dynamic hierarchical algorithms for document clustering[J]. Pattern Recognition Letters, 2010, 31(6): 469-477

[12] Chen C L, Tseng F S C, Liang T. Mining fuzzy frequent itemsets for hierarchical document clustering[J]. Information Processing & Management, 2010, 46(2): 193-211

(下转第259页)

on Electronic and Mechanical Engineering and Information Technology. 2011:1385-1387

[5] Wang Yan-xia, Qian Long-jun, Guo Zhi, et al. Weapon target assignment problem satisfying expected damage probabilities based on ant colony algorithm[J]. Journal of Systems Engineering and Electronics, 2008, 19(5): 939-944

[6] Aydin M E, Fogarty T C. A distributed evolutionary simulated annealing algorithm for combinatorial optimisation problems [J]. Journal of Heuristics, 2004, 10(3): 269-292

[7] Lee Z J , Lee W L. A Hybrid Search Algorithm of Ant Colony Optimization and Genetic Algorithm Applied to Weapon-Target Assignment Problems[C]// IDEAL 2003. 2003, 2690: 278-285

[8] Ding Zhu, Ma Da-wei, Tang Ming-duan, et al. TSAPSO: A Hybrid Search Algorithm of Tabu Search and Annealing Particle Swarm Optimization for Weapon-Target Assignment[J]. Journal of System Simulation. 2006, 18(9): 2480-2483(in Chinese)
丁铸, 马大为, 汤铭端, 等. 基于禁忌退火粒子群算法的火力分配[J]. 系统仿真学报, 2006, 18(9): 2480-2483

[9] Kennedy J, Eberhart R C. Particle swarm optimization[C]// Proc of IEEE International Conference on Neural Networks. 1995: 1942-1948

[10] Gao Shang, Yang Jing-yu. Solving weapon-target assignment problem by particle swarm optimization algorithm[J]. Systems Engineering and Electronics, 2005, 27(7): 1250-1253(in Chinese)
高尚, 杨静宇. 武器-目标分配问题的粒子群优化算法[J]. 系统工程与电子技术, 2005, 27(7): 1250-1253

[11] Qu Zai-bin, Liu Yan-jun, Xu Xiao-fei. Discrete particle swarm optimization for solving WTA problem[J]. Journal of Harbin Institute of Technology, 2011, 43(3): 67-69, 101(in Chinese)
曲在滨, 刘彦君, 徐晓飞. 用离散粒子群优化算法求解 WTA 问题[J]. 哈尔滨工业大学学报, 2011, 43(3): 67-69, 101

[12] Fan Cheng-li, Xing Qing-hua, Zheng Ming-fa, et al. Weapon-target allocation optimization algorithm based on IDPSO[J]. Systems Engineering and Electronics, 2015, 37(2): 336-342(in Chinese)
范成礼, 刑清华, 郑明发, 等. 基于 IDPSO 的武器目标分配优化算法[J]. 系统工程与电子技术, 2015, 37(2): 336-342

[13] Zadeh L A. Fuzzy sets[J]. Information and Control, 1965, 8(3): 338-356

[14] Wang Yi, Lei Ying-jie. A technique for constructing intuitionistic fuzzy entropy[J]. Control and Decision, 2007, 12(22): 1390-1394(in Chinese)
王毅, 雷英杰. 一种直觉模糊熵的构造方法[J]. 控制与决策, 2007, 12(22): 1390-1394

[15] Wang Yu-zhe, Lei Ying-jie, Zhou Lin, et al. Intuitionistic fuzzy discrete particle swarm algorithm[J]. Control and Decision, 2012, 27(11): 1735-1740(in Chinese)
汪禹喆, 雷英杰, 周林, 等. 直觉模糊离散粒子群算法[J]. 控制与决策, 2012, 27(11): 1735-1740

[16] Ruan Min-zhi, Li Qing-min, Liu Tian-hua, et al. Modeling and Optimization on Fleet Antiaircraft Firepower Allocation[J]. ACTA ArmamentarII, 2010, 31(11): 1525-1529(in Chinese)
阮旻智, 李庆民, 刘天华, 等. 编队防空火力分配建模及其优化方法研究[J]. 兵工学报, 2010, 31(11): 1525-1529

[17] Gao Shang, Yang Jing-yu, et al. Particle swarm optimization based on the ideal of wimulated annealing algorithm[J]. Computer Applications and Software, 2005, 22(1): 103-104, 80(in Chinese)
高尚, 杨靖宇, 等. 基于模拟退火算法思想的粒子群优化算法[J]. 计算机应用与软件, 2005, 22(1): 103-104, 80

[18] Wang Shao-Lei, Chen Wei-yi, Gu Xue-feng, et al. Solving weapon-target assignment problems based on self-adaptive differential evolution algorithm[J]. Systems Engineering and Electronics, 2013, 35(10): 2115-2121(in Chinese)
王少雷, 陈维义, 顾雪峰, 等. 自适应差分进化算法求解多平台多武器-目标分配问题[J]. 系统工程与电子技术, 2013, 35(10): 2115-2121

(上接第 233 页)

[13] Kuang D, Park H. Fast rank-2 nonnegative matrix factorization for hierarchical document clustering[C]// Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2013: 739-747

[14] Jaiswal A, Janwe N J. Fuzzy Association Rule Mining Algorithm to Generate Candidate Cluster: An Approach to Hierarchical Document Clustering[J]. International Journal of Computer Scie-nce Issues, 2012, 9(2)

[15] Kiran K N, Santosh G S K, Varma V. Multilingual Document Clustering Using Wikipedia as External Knowledge[M]// Multidisciplinary Information Retrieval. Springer Berlin Heidelberg, 2011: 108-117

[16] Nasir J A, Varlamis I, Karim A, et al. Semantic smoothing for text clustering[J]. Knowledge-Based Systems, 2013, 54(4): 216-229

[17] Xu Chen-kai, Gao Mao-ting. Improved ART 2 neural network for text clustering based on LSA[J]. Computer Engineering and Applications, 2014, 2(24): 133-138, 177(in Chinese)
徐晨凯, 高茂庭. 使用 LSA 降维的改进 ART2 神经网络文本聚类[J]. 计算机工程与应用, 2015, 2(24): 133-138, 177

[18] Li H, Li J, Wong L, et al. Relative Risk and Odds Ratio: A Data Mining Perspective(Corrected Version)[C]// PODS'05. 2005: 368-377

[19] Gale W A, Sampson G. Good-turing frequency estimation without tears[J]. Journal of Quantitative Linguistics, 1995, 2(3): 217-237

[20] Chen W Y, Song Y, Bai H, et al. Parallel spectral clustering in distributed systems[J]. IEEE Transactions on Software Engineering, 2011, 33(3): 568-586

[21] Kim C W, Sun P. Enhancing Text Document Clustering Using Non-negative Matrix Factorization and WordNet[J]. Journal of Information & Communication Convergence Engineering, 2013, 11(4): 241-246

[22] Kuang D, Park H. Fast rank-2 nonnegative matrix factorization for hierarchical document clustering[C]// Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2013: 739-747

[23] Huang Xian-ying, Liu Ying-tao, Rao Qin-fei. Similarity Algorithm Based on Common Chunks Between English Short Texts [J]. Journal of Chongqing University of Technology(Natural Science), 2015, 29(8): 88-93(in Chinese)
黄贤英, 刘英涛, 饶勤菲. 一种基于公共词块的英文短文本相似度算法[J]. 重庆理工大学学报(自然科学版), 2015, 29(8): 88-93