第31卷 第1期 82-94, 2012年1月
AND TECHNICAL INFORMATION Vol. 31 No. 1, 82-94 ISSN 1000-0135 January 2012

doi:10.3772/j.issn.1000-0135.2012.01.011

基于学术文献同被引分析的K-means算法改进研究[1)]

吴夙慧[1] 成 颖[1] 郑彦宁[2] 潘云涛[2]

(1.南京大学信息管理系,南京 210093; 2.中国科学技术信息研究所,北京 100038)

摘要 K-means算法是一种应用广泛的聚类算法,但是存在初始聚类中心和K值选取的难题。本文提出了一种基于学术文献同被引分析的初始聚类中心和K值选取的K-means改进算法。该算法属于两步聚类算法,首先对学术文献进行同被引分析,得到同被引矩阵,然后基于同被引矩阵进行层次聚类。算法记录每次迭代过程中被聚为一类的学术文献间的距离以及两次迭代间的距离差,当两次迭代的距离差取得最大值时取其聚类数作为第二步K-means算法的K值,并且将此时的类中心作为第二步K-means算法的初始聚类中心。第二步聚类则依据文献内容实现K-means算法。实验通过与经典K-means算法和基于凝聚层次聚类算法的改进K-means算法的对比,证明了本文提出的改进的K-means算法具备更优的聚类效果。

关键词 K-means算法 K值 初始聚类中心 同被引 文献聚类

Improvement of K-means Algorithm Based on Co-citation Analysis

Wu Suhui[1], Cheng Ying[1], Zheng Yanning[2] and Pan Yuntao[2]

(1. *Department of Information Management, Nanjing University, Nanjing* 210093;
2. *Institute of Scientific and Technical Information of China, Beijing* 100038)

Abstract K-means algorithm is a widely-used clustering algorithm. The main problem of the algorithm is the determination of the optimal number of clusters and the selection of initial cluster centers. In this paper, a novel algorithm based on co-citation analysis is proposed. This algorithm is divided into two steps. The first step is to do co-citation analysis in the academic literature set, and get the matrix of co-citation, and run hierarchical clustering algorithm based on the matrix. In each iteration, distance of academic literature in a cluster and the difference of the distance between two iterations are recorded. In the end of first step, the value of K and the centers of every cluster are selected for the second step when the maximum of the difference is achieved. The second part of the research is to execute the K-means algorithm based on the content of academic literature. Experimental results show that the clustering quality is improved.

Keywords K-means algorithm, number of clusters, initial clustering centers, co-citation, papers clustering

1 引 言

随着CNKI、Elsevier等国内外学术全文数据库的广泛应用,科研工作者已经可以轻松地获取到大量的学术信息。但是由于学术信息专业性强,彼此间相似度高,用户在检索学术信息时要从成百上千条检索结果中找到符合自身需求的信息,需要付出

收稿日期:2011年1月5日

作者简介:**吴夙慧**,男,1988年生,硕士生,主要研究方向:自然语言处理。E-mail: wush13@126.com。**成颖**,男,1971年生,副教授,主要研究方向:信息检索。**郑彦宁**,男,1965年生,研究员,主要研究方向:情报技术与方法、竞争情报。**潘云涛**,女,1967年生,研究员,主要研究方向:科学计量学。

1) 本文得到国家社科基金项目"中文学术信息检索系统相关性集成研究"(项目批准号:10CTQ027),教育部人文社会科学研究规划基金项目"面向用户的相关性标准及其应用研究"(项目批准号:07JA870006),中国科学技术信息研究所合作研究项目的资助。

大量的精力。面对这样的窘境,迫切需要引入一种方法把用户从重复的逐条点击中解放出来,而文本聚类技术无疑是一个不错的选择。运用文本聚类技术对检索结果自动聚类,这样返回给用户的将不再是一个简单的线性检索结果列表,而是多个具有层次结构的检索结果类别,通过赋予每个类别足以反映其内容的聚类标签,这样用户就可以快速地了解到检索结果的全貌,大大缩短信息获取的时间。目前,在 Elsevier 等一些西文学术数据库中,已经提供了检索结果的聚类分析。在中文学术数据库中,CNKI 也已经提供了基于文献关键词的检索结果聚类。章成志等[1]基于特征组合方法实现了主题提取,然后采用 K-means 算法完成主题聚类,其研究成果已经在 http://topic.cnki.net 系统中得到了实际应用。Elsevier 等检索系统提供的检索结果聚类功能缩短了用户获取相关文献的时间,然而依然存在聚类粒度偏大等有待改进之处。

本文提出了一种改进的 K-means 聚类算法,该算法基于学术文献同被引分析选择 K 值和初始聚类中心。算法首先在对学术文献进行同被引分析的基础上,利用同被引矩阵进行层次聚类;然后通过本文设计的算法取得较优的 K 值和初始聚类中心;最后再依据文献内容采用 K-means 算法进行文献聚类。实验通过与经典 K-means 算法和基于凝聚层次聚类算法的改进 K-means 算法的对比,证明了本文提出的改进 K-means 算法具备更优的聚类效果。

2 相关研究

文本聚类是应用于文本数据上的聚类分析,它将文本集分成若干簇,使得同一簇内的文本相似度尽可能大,不同簇间的文本相似度尽可能小。目前,学界已经提出了多种文本聚类算法,其中应用最广泛的是 K-means 算法。K-means 算法由 Macqueen[2]在 1967 年提出,由于其简单的算法思想、优秀的聚类效果,很快得到了广泛的应用。不过随着对 K-means 算法研究的深入,算法的许多不足逐渐暴露出来,其中学界研究最多的是初始聚类中心和 K 值的选取问题。

自 1984 年 Selim 和 Ismail[3]发现 K-means 算法会受到初始聚类中心的影响而陷入局部最优解以来,学界提出了多种初始聚类中心和 K 值选取的方案,主要有以下三类:

(1) 基于密度的解决方案。基本思想是首先通过对数据点的密度分布情况进行先期了解,进而选取初始聚类中心和 K 值,从而有效地避免初始聚类中心过于密集的情况。其中较具代表性的有 Katsavounidis[4]、Khan 等[5]和 Redmond 等[6]的研究。

(2) 基于优化算法的解决方案。将模拟退火算法和遗传算法等优化算法应用于初始聚类中心和 K 值的选取,这些优化算法通过多次的优化使得算法逐步逼近最优解。其中较具代表性的研究有 Krishna 和 Murty[7]提出的 GKA 算法、Bandyopadhyay 和 Maulik[8]提出的 GCUK 算法等。

(3) 基于凝聚层次聚类算法的解决方案。其基本思想是通过凝聚层次聚类算法对数据集进行初始聚类,得到数据集的分布以获得第二步 K-means 算法的初始聚类中心。这种解决方案需要预先确定 K 值,即确定第一步层次聚类算法的最佳聚类数。这也是学界的另一个研究重点。Halkidi 等[9]提出 SD 有效性指数,即通过判断聚类平均散布性和聚类间的总体分离性来评价聚类效果,从而得到层次聚类算法的最优划分。Jung[10]采用聚类熵作为指标确定最佳聚类数,当聚类熵取得最小值时认为达到最佳聚类划分。胡晓庆等[11]则采用类间相似度与类内相似度之差作为评价指标,当该值取得最大时,即认为聚类的类间离散性和类内致密性最令人满意,从而取得最佳聚类数。

上述三类方法适用于所有的数据集。对于学术文献的聚类而言,学术界则另辟蹊径,提出了利用学术文献的引文信息进行文献聚类的思想。根据文献计量学的观点,学术文献的引文信息可以较好地反映文献的内容,因而引文信息的相关度可以认为是文献内容相关度的重要度量。近些年研究与应用较多的同被引聚类方法就是该思想的典型应用。同被引的概念最早由美国情报学者 Small[12]提出。基于同被引聚类的前提假设是:如果两篇或多篇文献同时被另一篇文献引用,则它们之间在内容上具备一定程度的相关性,因而可以将同被引次数较多的两篇或多篇论文聚成一类。目前,该方法被广泛地应用于揭示某一学科或主题的结构和研究热点等领域,代表性的研究有 Sullivan[13]、Chen [14]、崔雷等[15~17]。不过,作者的引用行为是多样的,除了文献之间内容的相关性之外,还包括"对开拓者表示尊重"、"为阿谀某人而引证"、"迫于权威压力的引证"等行为[18],因此仅依据引文信息的聚类结果也没有获得令人满意的效果。

基于此，近年来有研究提出了结合引文信息与文献内容信息的聚类方案，具代表性的有 Aljaber 等[19]、章成志等[20]的研究。Aljaber 等的研究利用施引文献与被引文献的内容相关性，利用引文中的特征项对施引文献的特征向量进行加权。实验证明，这种方法可以明显地改善聚类效果。章成志等的研究基于 PageRank 算法思想，利用被引频次信息计算论文的 PageRank 值，并依据该值对论文加权，然后用 K-means 算法完成聚类。实验证实该算法得到了较好的聚类效果。

3 基于同被引的 K-means 聚类算法

本文提出的面向学术文献聚类的 K-means 改进算法是一种两步聚类算法。第一步聚类使用层次聚类算法，利用学术文献间的同被引信息对文献进行聚类。第二步聚类则依据学术文献的内容使用 K-means算法进行聚类。第一步聚类的目的是为第二步聚类的 K-means 算法提供更优的聚类数 K 和初始聚类中心。

3.1 算法思想

在第一步聚类中，首先将原始同被引矩阵转化为相关矩阵。原始同被引矩阵见公式(1)，其中 C_{ij} 表示学术文献 i 和 j 的同被引频次，当 $i=j$ 时表示学术文献 i 的被引频次，可简写为 C_i。

$$\begin{pmatrix} C_{00} & C_{01} & \cdots & C_{0(m-1)} \\ C_{10} & C_{11} & \cdots & C_{1(m-1)} \\ \vdots & \vdots & & \vdots \\ C_{(m-1)0} & C_{(m-1)1} & \cdots & C_{(m-1)(m-1)} \end{pmatrix} \quad (1)$$

采用 Jaccard 系数[公式(2)]将同被引矩阵进行转换，得相关矩阵。其中，$J_{ij} \leqslant 1$。

$$J_{ij} = C_{ij}/(C_i + C_j - C_{ij}) \quad (2)$$

可见，相关矩阵是一个主对角线为 1 的对称矩阵，如公式(3)所示。

$$\begin{pmatrix} 1 & J_{01} & \cdots & J_{0(m-1)} \\ J_{10} & 1 & \cdots & J_{1(m-1)} \\ \vdots & \vdots & & \vdots \\ J_{(m-1)0} & J_{(m-1)1} & \cdots & 1 \end{pmatrix} \quad (3)$$

对相关矩阵中的列向量执行凝聚层次聚类算法，即不断将距离最近的两个数据点聚为一类，并重新计算聚类中心。对一个具有 m 个元素的文本集，在迭代 $m-1$ 次之后，即可完成聚类。凝聚层次聚类算法的一个核心问题就是最优聚类数的确定，本文针对该问题提出了如下的解决方案。首先给出如下定义：

定义 1（聚簇集 $C(i)$）：给定具有 n 元素（或称数据点）的初始文本集，记为集合 $P=\{p_1,p_2,\cdots,p_n\}$。第 i 次迭代后，文本集被聚为 $n-i$ 个聚簇，记为集合 $C(i)=\{c_1,c_2,\cdots,c_{n-i}\}$（$C(i)$ 的初始值为 P）。$C(i)$ 即为第 i 次迭代后的聚簇集。

定义 2（聚类距离 D_i）：D_i 为第 i 次迭代后被聚为一类的两个数据点间的距离，对于第 i 次迭代而言，即有

$$D_i = \min_{c_j,c_k \in C(i-1)} (Sim(c_j,c_k)) \quad (c_j \neq c_k) \quad (4)$$

也就是说，D_i 为 $C(i-1)$ 中聚簇间距离的最小值，其中 $Sim(c_j,c_k)$ 表示聚簇 c_j 和 c_k 之间的距离。

由于凝聚层次聚类算法每次迭代总是先将距离最小的两个数据点聚为一类，因此本算法经过 $n-1$ 次迭代后将得到一个逐步增大的聚类距离序列集合 $D=\{D_1,D_2,\cdots,D_{n-1}\}$。

根据聚类算法的基本思想，即聚类算法应使类内相似度尽可能大，而类间相似度尽可能小的基本原则，因此仅仅依据 D_i 的绝对值显然难以同时衡量聚簇内以及聚簇间的离散度，相应地也就难以确定最优聚类数。不过可以通过考察 $\Delta D_i = D_{i+1} - D_i$ 进行尝试。

下面考察第 i 次和第 $i+1$ 次迭代。第 i 次迭代是在 $C(i-1)=\{c_1,c_2,\cdots,c_j,\cdots,c_k,\cdots,c_{n-i+1}\}$ 上进行的，第 i 次迭代将 $C(i-1)$ 集合中距离最近的两个聚簇 c_j、c_k 聚为一类，形成一个新的聚簇 c'_j，c_j 与 c_k 的距离即为 D_i，第 i 次迭代后得到一个新的 $C(i)=\{c_1,c_2,\cdots,c'_j,\cdots,c_{n-i}\}$。对于集合 $C(i)$ 而言，D_i 是最后被聚为一类的两个聚簇的距离，可以看作集合 $C(i)$ 类内距离的最大值。同理，随后的第 $i+1$ 次迭代继续将集合 $C(i)$ 中最近的两个聚簇聚为一类，这个最近的距离即为 D_{i+1}，D_{i+1} 即是集合 $C(i)$ 类间距离的最小值。综上所述，$\Delta D_i = D_{i+1} - D_i$ 反映了聚簇集合 $C(i)$ 类间距离和类内距离的差。

设具有最大值的 ΔD_i 记为 ΔD，则有，

$$\Delta D = \max(\Delta D_i)，\text{其中 } \Delta D_i = D_{i+1} - D_i，\quad i = 1,2,\cdots,n-2 \quad (5)$$

因为 $\Delta D = \max(\Delta D_i)$，故有 $D_{i+1} >> D_i$，根据 D 的生成算法有，

$$D_1 < D_2 < \cdots < D_{i-1} < D_i << D_{i+1} < \cdots < D_{n-1} \quad (6)$$

据此,本算法有如下性质。

性质:满足 $\Delta S = \max(\Delta S_i)$ 的 S_i 对应的聚类数(表示为 $|C(i)|$)是最优聚类数。

证明:

(1) 当 $k < i$ 时,根据公式(6)有 $D_k < D_{k+1} < \cdots < D_i$,其类间距离和类内距离之间并没有极大的差异,因而尚有进一步聚类的空间。一般地,当 $k \to 1$ 时,$|C(k)| \to n$,这显然不是所期望的结果。

(2) 当 $k > i+1$ 时,由于 $D_i << D_{i+1}$,根据式(6)和不等式的传递规则有 $D_i << D_k$,因而此后的聚类使类内相似度迅速减小;且当 $k \to n-1$ 时,$|C(k)| \to 1$,即聚簇数愈来愈小直至为 1,这显然也不是所期望的结果。

(3) 当 $k = i$ 时,$\Delta D_i = D_{i+1} - D_i$ 取得最大值,由(2)可知,此时类内距离已经达到比较满意的状态,后继的聚类已经没有意义。根据聚类算法的基本思想,即聚类算法应使类内距离尽可能小、而类间距离尽可能大的基本原则,应该选取在(2)基础上的最大类间距离。由(1)可知,当 $k = i$ 时的类间距离 D_{i+1} 达到最大值,即聚簇之间已经足够离散。

由此可得出结论,当 $\Delta D = \max(\Delta D_i)$ 时,类间距离与类内距离达到最优的平衡,从而此时对应的 $C(i)$ 为最优的聚类结果。

3.2 算法描述

实验表明,本算法需要注意孤立点,即孤立聚簇问题。所谓孤立点,即该点与其他所有聚簇的距离都很大,是孤悬在外的点。孤立点的出现会导致第一步聚类所生成的 K 值过小,从而导致整个聚类过程变得意义非常有限。本实验的解决方案是当 ΔS_i 取得最大值时,如果 K 值很小或者是将一个文本数量很少的聚簇和其他聚簇进行聚类,则放弃最大值,选取次大值。

第一步凝聚层次聚类算法描述如下:

Input:同被引矩阵($n \times n$)。

Output:K 值、初始聚类中心集 $C\{C_1, C_2, \cdots, C_k\}$。

Step:

Step1:根据公式(1)计算得到相关矩阵。令 $i=0$。

Repeat

Step2:将相关矩阵中的列向量(或行向量)中距离最近的聚为一类,记录此距离值为 distance[i]。

Step3:采用算术平均方法对聚为一类的点重新计算中心。

Step4:$i = i+1$。

Until $i = n-1$。

Step5:计算 disdif[j] = distance[$j+1$] - distance[j]值,$j \in [1, n-2]$,对数组 disdif 排序。

Step6:取出最大值 disdif[k],如果第 k 次聚类时有孤立点,则转向次大值。

Step7:得到聚类数 k,以各类均值作为初始聚类中心集 $C\{C_1, C_2, \cdots, C_k\}$。算法结束。

在第二步聚类中,使用学术文献的题名和文摘作为聚类数据,运用向量空间模型和 TFIDF 加权方法将学术文献表示为一个向量。由于题名相比文摘在语义上更为重要[21],因而对题名中的特征项赋予更高的权重。在先前的研究中,学者们也都意识到了题名、文摘以及全文中的特征项应该赋予不同的权重,但是具体加权方案并没有统一的规范,不同研究之间差异较大。赵云志[22]在研究中文自动标引时,对来自题名和文摘的特征项赋以 3:2.5 的权值。孟海涛和陈笑蓉[23]在研究科技文献聚类时,对题名和文摘中的特征项赋以 4:1 的权值。刘海峰等[24]在研究文本特征提取时将题名和文摘等同看待,赋以相同的权值。赵纪元和罗霄[25]则按照 4:1.5进行加权。

目前,题名、文摘以及全文中特征项的加权方案更多依靠学者的研究经验,并没有太多的理论依据。比如韩客松和王永成[26]认为,"据我们的经验,标题中的关键词比摘要中的大概重要 3 ~ 5 倍"。其他研究的依据也都类似。本文中结合前面几项研究,采取折中的方案,即相比于文摘,对题名中的特征项赋予两倍的权值。由于 K 值和初始聚类中心已经在第一步聚类中得到,就可以运用 K-means 算法完成最终的聚类。K-means 算法的描述请参阅文献[2],本文不再赘述。在实验中,本文使用 Salton 主持研发的 SMART 系统的停用词表[27]进行停用词处理,使用最常用的余弦距离[公式(7)]作为相似度计算方法。

$$Sim(d_i, d_j) = \frac{d_i \cdot d_j}{|d_i| \times |d_j|} \tag{7}$$

其中,$\cdot$ 表示两文本向量的点积,$|d_i|$ 表示向量的长度。

与经典的 K-means 算法一样,本算法依然采用距离平方和 W_n 作为算法的收敛条件[2]:

$$W_n = \sum_{i=1}^{n} \min_{1 \le j \le k} |x_i - a_j|^2 \tag{8}$$

其中 $x_1, x_2, \cdots, x_n$ 表示全部 n 个数据点，$a_1, a_2, \cdots, a_k$ 表示 k 个聚类中心。W_n 即为所有的数据点到该点所属聚类中心的距离的平方和。当前后两次的 W_n 值不再发生变化时，算法结束。

完成聚类后，需要从每个聚簇的学术文献中抽取 TFIDF 值较大的多个词作为聚类标签。

3.3 聚类标签的生成

在 Elsevier 数据库中，一般采用一个单词或词组作为聚类标签，这样的标签不足以让用户深入了解类别的内容。因而本文在生成聚类标签时，选择多个单词作为聚类标签，以便更好地反映类内文献的内容。

具体的聚类标签生成算法如下：从各聚簇内抽取出所有的特征项，计算它们的 *TFIDF* 值并加以排序，选择排在前面的几个词作为聚类标签。在本文中，选择前五个词作为聚类标签。本文对 *TFIDF* 值的计算公式进行一些修改，以确保选取的聚类标签既具备较高的重要性，又能很好地反映聚簇内所有文档的内容。计算公式如下：

$$TFIDF(\omega_i) = \frac{TF_j(\omega_i) \times DF_j(\omega_i)}{DF(\omega_i)} \quad (9)$$

其中，$TF_j(\omega_i)$ 表示特征项 ω_i 在聚簇 j 内所有学术文献中的词频值，$DF(\omega_i)$ 表示特征项 ω_i 在整个文献集中的文档频率，$DF_j(\omega_i)$ 表示特征项 ω_i 在聚簇 j 中的文档频率。加入 $DF_j(\omega_i)$ 项的目的是使抽取的聚类标签尽可能全面地反映聚簇内所有文档的内容。

3.4 算法性能分析

本文提出的算法是 K-means 算法在学术文献聚类上的扩展，主要区别在于通过前期对学术文献的同被引分析得到 K-means 算法的 K 值和初始聚类中心，再利用 K-means 算法完成聚类。两步聚类方法无疑增加了算法的时间复杂度。本算法采用的凝聚层次聚类法的时间复杂度为 $O(n^2 \log n)$，K-means 算法的时间复杂度为 $O(nkt)$。因而本算法的时间复杂度为 $O(n^2 \log n + nkt)$，其中 n 代表所有学术文献的数目，k 为聚类数，t 为 K-means 算法的迭代次数。

考察基于优化算法的 K-means 初始中心选择算法的时间复杂度。以 Krishna 和 Murty[28] 提出的 GKA 算法为例，GKA 算法经过种群初始化、逐代的交叉遗传和变异等步骤完成聚类，其算法的时间复杂度为 $O((nkl + nkt) * g)$。其中 n 代表遗传算法中的种群数目，即所有学术文献的数目，k 为聚类数，l 为染色体的长度，t 为 K-means 算法的迭代次数，g 为遗传算法的遗传代数，也即整个算法的迭代次数。由于大部分学术文献数据库已经提供了学术文献的被引频次以及施引学术文献链接等信息，因此同被引矩阵是易于获得的。另外，学术文献数据库检索结果的记录数 n 通常不大，则 GKA 算法与本文提出的两步聚类算法相比较，本文提出的算法时间复杂度明显降低。因此，将本算法应用于学术文献检索结果的聚类，其时间复杂度是可接受的。

4 实验与结果分析

4.1 实验步骤与数据来源

为简化实验，本文采用了英文论文作为数据源。本文的实验分为两次进行。第一次实验选择情报学领域的五个研究热点：信息检索、数据挖掘、知识管理、引文分析、知识产权。各选取 50 篇论文，共 250 篇论文，以它们的题名和文摘作为数据集。第二次实验在 ISI of Knowledge 数据库中以“主题 = K-means”作为检索词，得到了 1572 篇文献，选取其中被引频次最高的前 100 篇论文，套录每篇论文的题名和文摘以及全部施引文献。由于 ISI of Knowledge 并没有提供同被引次数的统计，因而本文通过自编程序对两篇论文的全部施引文献题名的重复次数进行统计作为同被引次数。

为了检验算法效果，实验采用经典的 K-means 算法（以下简称算法 1）、基于凝聚层次聚类算法（agglomerative hierarchical clustering）的 K-means 算法（以下简称算法 2）与本文提出的算法进行对比。其中基于凝聚层次聚类算法的 K-means 算法已在第二节叙及，其思想与本文类似，区别在于第一步层次聚类是应用于文本的内容向量，在确定最佳聚类数时本实验采用第二节提及的胡晓庆等[11]的方法，即利用类间相似度与类内相似度之差作为聚类划分的评价指标。

本文提出的改进算法，整个实验步骤按照图 1 所示算法流程进行。算法在第一步同被引层次聚类之前需要生成原始同被引矩阵，记录论文两两之间的同被引次数，然后转换为相关矩阵。在进行第二步 K-means 算法聚类前需要依次进行题名与文摘提取、停用词去除、词根处理和特征项权重计算等文本预处理工作，所选用的停用词表、权重计算方法已经在 3.2 节完成。

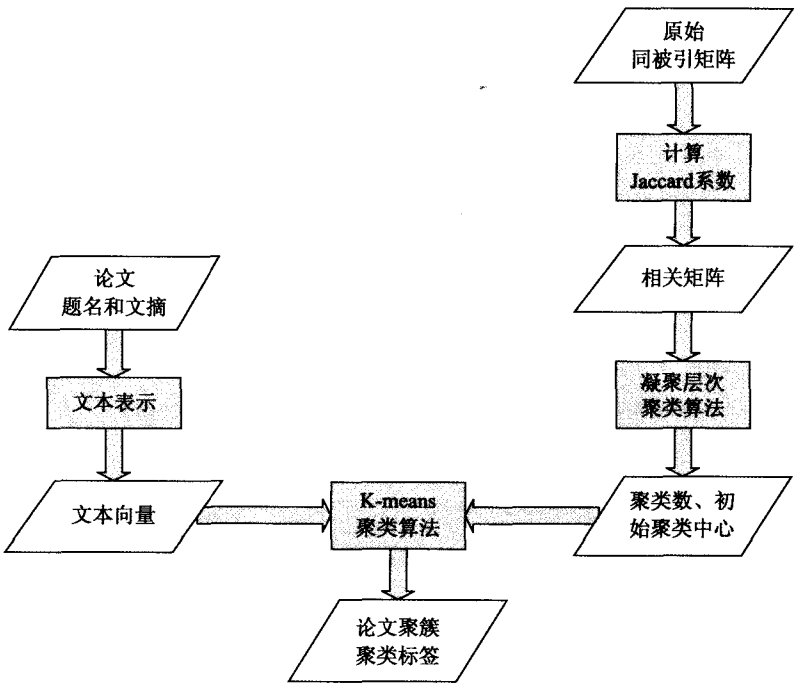

图 1　聚类算法基本流程

4.2　文本聚类评价指标

学界对于文本聚类的评价指标可以分为外部评价标准和内部评价标准两大类。

4.2.1　外部评价标准

外部评价标准主要是通过算法生成的聚类和实际的人工分类的对应程度来评价聚类的效果。主要包括熵、平均准确率等指标。

1）熵(entropy)[20]

熵是一种应用广泛的指标，计算方法为：设聚类算法生成的聚簇集为 $C\{C_1,C_2,\cdots,C_k\}$，人工的文本分类集为 $P\{P_1,P_2,\cdots,P_n\}$，首先计算每一个聚簇 C_i 的熵，计算公式为：

$$E_{C_i} = -\sum_{j=1}^{n}(p_{ij}\log_{10}p_{ij}) \tag{10}$$

其中，p_{ij} 为聚簇 i 中的文献属于分类 j 的概率。p_{ij} 的计算公式为：

$$p_{ij} = m_{ij}/m_i \tag{11}$$

其中，m_{ij} 是聚簇 i 中属于分类 j 的文献数量，而 m_i 为聚簇 i 中所有文献的数量。

在得到所有聚簇的熵值 $\{E_{C1},E_{C2},\cdots,E_{Ck}\}$ 之后，就可以计算总熵值[公式(12)]。总熵值越低，则认为聚类效果越好。

$$E = \sum_{i=1}^{k}\left(\frac{m_i}{m}E_{C_i}\right) \tag{12}$$

其中，m 为聚类文献集的总数目。

2）平均准确率(average accuracy)[29]

平均准确率是通过考察文献两两之间是否属于相同的聚簇和相同的分类来评价聚类效果的一种指标。对于文献集中的任意两篇文献，存在四种不同的关系：

(1) 两篇文献属于同一个分类，也被算法分入同一个聚簇中。使用 TP(True Positives，积极正确数)来表示该文献对的数目。

(2) 两篇文献属于同一个分类，被算法分入不同的聚簇中。使用 FN(False Negatives，消极错误数)来表示该文献对的数目。

(3) 两篇文献不属于同一个分类，被算法分入同一个聚簇中。使用 FP(False Positives，积极错误数)来表示该文献对的数目。

(4) 两篇文献不属于同一个分类，被算法分入不同的聚簇中。使用 TN(True Negatives，消极正确数)来表示该文献对的数目。

得到 TP、FN、FP 和 TN 四个指标后，可以利用式(13)、式(14)来计算积极准确率 PA 和消极准确率 NA。

$$PA = \frac{TP}{TP + FN} \tag{13}$$

$$NA = \frac{TN}{FP + TN} \tag{14}$$

PA 表示在所有的同类文献对中，被正确聚类

的文献对所占的比例。*NA* 表示在所有的不同类文献对中，被正确聚类的文献对所占的比例。利用公式(15)就可以得到平均准确率 *AA* 的值

$$AA = \frac{PA + NA}{2} \tag{15}$$

AA 值越大，则认为聚类效果越好。

4.2.2 内部评价指标

在外部信息缺乏的情况下，如难以对文献进行人工分类时，就需要利用内部评价指标来评价聚类效果。经常采用的内部评价指标有凝聚度。凝聚度是指所有文献向量到其聚类中心的距离和。实际应用中一般采用最大凝聚度(*SSE*)，计算公式如下：

$$SSE = \sum_{i=1}^{K} \sum_{j=1}^{m_i} (1 - d(v_j, c_i)) \tag{16}$$

其中，v_j 表示文本向量，c_i 表示聚簇 i 的聚簇中心，m_i 表示聚簇 i 中的文献数目。$d(v_j, c_i)$ 表示两向量间的距离。*SSE* 值越小，则认为聚类效果越好。

实验分别选取上述三种评价指标来评价聚类效果。

4.3 实验结果及分析

本文设计的两次实验是在两种不同类型的数据集上进行的。第一次实验所选用的论文集分为五类，不同类之间的论文内容差别较大，结构较为松散。第二次实验所选用的论文集都是 K-means 算法研究领域的论文，因而结构较为紧密。本文通过两次实验，以期全面地测试算法的效果。

4.3.1 第一次实验结果及分析

第一次实验分为三组，第一组采用算法1，设置 $K=5$，进行三次实验，得到的聚类评价指标和聚类标签如表1、表2所示。可以看到聚类效果不稳定，效果的好坏取决于初始聚类中心的选择，随机性比较大。

表1 算法1聚类指标($K=5$)

实验序号	熵值(entropy)	准确率(accuracy)			凝聚度(*SSE*)
		积极准确率(*PA*)	消极准确率(*NA*)	平均准确率(*AA*)	
1	0.253285	0.712327	0.920120	0.816224	214.22303
2	0.448946	0.620735	0.77056	0.695647	217.518533
3	0.281461	0.744490	0.85304	0.798765	215.231911
平均值	0.327897333	0.692517	0.847907	0.770212	215.6578

表2 算法1聚类标签

实验序号	聚簇1	聚簇2	聚簇3	聚簇4	聚簇5
1	intellectual property right information-retrieval law	knowledge management organization relevance ecological	data mining gene knowledge measure	information retrieval effect memory visual	citation analysis journal science impact
2	management knowledge information-retrieval journal citation	co-citation science citation market analysis	property mining intellectual right patent	information retrieval visual storage relevance	biology molecule accomplishment bank geometry
3	management knowledge citation technology analysis	right property law intellectual protection	privacy mining data gene pattern	cognitive information retrieval theory memory	information-retrieval online software user co-occurrence

第二组采用算法 2，第一步得到聚类数 $K=7$，计算得到 7 个初始聚类中心，代入 K-means 算法得到的聚类评价指标和聚类标签如表 3、表 4 所示。

第三组采用本文提出的算法，第一步层次聚类迭代得到 $K=7$，并计算得到相应的 7 个聚类中心。第二步 K-means 算法聚类，得到的聚类评价指标和聚类标签如表 5、表 6 所示。

从表中可以看到，即便是在经典 K-means 算法事先知道聚类数的情况下，本文提出的算法依然可以取得更好、更稳定的聚类效果。但是相比算法 2，则聚类效果略为逊色。算法所生成的聚类标签可以很好地反映聚簇的内容，但是标签同时也存在一个问题：由于在文本预处理阶段将所有的单词分开，因而生成的聚类标签将许多诸如 data mining 等短语拆分开来，造成了语义上的损失。

4.3.2 第二次实验结果及分析

第一次实验是将五个不同类别的论文合并在一起构成聚类文本集，而第二次实验则采用真实的论文检索结果作为文本集，文本集中的论文内容更为接近，这样的实验结果应该更能真实地反映算法的聚类效果。

由于论文没有事先进行分类，这样就需要使用外部评价指标进行评价。由人工考察 100 篇论文的内容，通过人工阅读的方式将论文分为十大类。

表 3 算法 2 聚类指标

熵值(entropy)	准确率(accuracy)			凝聚度(*SSE*)
	积极准确率(*PA*)	消极准确率(*NA*)	平均准确率(*AA*)	
0.240462	0.684840	0.941457	0.813149	204.821322

表 4 算法 2 聚类标签

聚簇 1	聚簇 2	聚簇 3	聚簇 4	聚簇 5	聚簇 6	聚簇 7
information-retrieval	right	science	mining	assessment	retrieval	impact
information	property	citation	data	perspective	study	citation
retrieval	intellectual	analysis	model	knowledge	theory	research
system	organization	journal	database	management	storage	patent
study	law	literature	privacy	strategy	functional	co-citation

表 5 本文算法聚类指标

熵值(entropy)	准确率(accuracy)			凝聚度(*SSE*)
	积极准确率(*PA*)	消极准确率(*NA*)	平均准确率(*AA*)	
0.255182	0.653061	0.95144	0.802251	208.530631

表 6 本文算法聚类标签

聚簇 1	聚簇 2	聚簇 3	聚簇 4	聚簇 5	聚簇 6	聚簇 7
information-retrieval	mining	citation	property	knowledge	information	literature
theory	data	analysis	law	management	retrieval	knowledge
retrieval	sequence	journal	intellectual	organizational	effect	discovery
storage	gene	impact	right	ecological	visual	biological
cognitive	privacy	research	drug	assessment	image	database

由于算法1需要事先确定聚类数和初始聚类中心，而在算法1中，这些信息是不能获得的。因而在算法1实验中，选取 $K=5$、$K=10$ 两种情况各进行两次实验，随机选取初始聚类中心，以期获得更加稳定的实验数据。得到的熵值、准确率和凝聚度三项指标如表7、表8所示。

对于算法2，第一步层次聚类取得 $K=15$，第二步聚类得到的熵值、准确率和凝聚度三项指标如表9所示。

而对于本文提出算法，第一步层次聚类取得 $K=14$，第二步聚类得到的熵值、准确率和凝聚度三项指标如表10所示。

从表中可以看到，相比算法1和算法2，本文提出的改进算法的三项评价指标都有了提升。

对于实验中算法所生成的聚类标签，囿于篇幅，这里不全部列出。对于算法1所进行的六次试验仅就 $K=5$ 和 $K=10$ 各列出标签一次，如表11、表12所示。另外列出算法2和本文提出的改进算法实验所生成的聚类标签，如表13、表14所示。

表7 算法1聚类结果($K=5$)

实验序号	熵值(entropy)	准确率(accuracy)			凝聚度(*SSE*)
		积极准确率(*PA*)	消极准确率(*NA*)	平均准确率(*AA*)	
1	0.573662	0.549719	0.0.688766	0.619243	79.77803
2	0.548542	0.554579	0.761284	0.657932	79.74362
3	0.634128	0.491284	0.657960	0.574622	80.13129
平均值	0.585444	0.531861	0.70267	0.617265	79.88431

表8 算法1聚类结果($K=10$)

实验序号	熵值(entropy)	准确率(accuracy)			凝聚度(*SSE*)
		积极准确率(*PA*)	消极准确率(*NA*)	平均准确率(*AA*)	
1	0.380482	0.642289	0.700480	0.671386	70.36573
2	0.481574	0.541278	0.651629	0.596454	71.96000
3	0.364127	0.624874	0.775234	0.700054	71.34834
平均值	0.408728	0.602814	0.709114	0.655964	71.22469

表9 算法2聚类结果

熵值(entropy)	准确率(accuracy)			凝聚度(*SSE*)
	积极准确率(*PA*)	消极准确率(*NA*)	平均准确率(*AA*)	
0.412135	0.586214	0.752482	0.669348	65.462454

表10 本文算法聚类结果

熵值(entropy)	准确率(accuracy)			凝聚度(*SSE*)
	积极准确率(*PA*)	消极准确率(*NA*)	平均准确率(*AA*)	
0.396744	0.601472	0.784567	0.693020	65.254610

表 11　算法 1 聚类标签($K=5$)

聚簇序号	聚簇 1	聚簇 2	聚簇 3	聚簇 4	聚簇 5
标签	K-means Cluster image analysis algorithm	large K-means cluster strong value	K-means segmentation neural market integration	local cluster K-means clustering optima	K-means fuzzy algorithm cluster method

表 12　算法 1 聚类标签($K=10$)

聚簇序号	聚簇 1	聚簇 2	聚簇 3	聚簇 4	聚簇 5
标签	K-means data cluster large algorithm	adaptive K-means algorithm cluster imaging	local algorithm cluster K-means characterization	neural network partition optimal segmentation	algorithm K-means method quantization vector
聚簇序号	聚簇 6	聚簇 7	聚簇 8	聚簇 9	聚簇 10
标签	analysis classification data cluster fuzzy	cluster genetic K-means segmentation initialization	base cluster conceptual distance K-means	K-means technique cluster algorithm clustering	design quantizer soil vector fuzzy

表 13　算法 2 聚类标签

聚簇序号	聚簇 1	聚簇 2	聚簇 3	聚簇 4	聚簇 5
聚类标签	hierarchical K-means algorithm optimal partition	algorithm cluster data vector method	categorical k-modes object K-means order	genetic gka K-means cluster partition	K-means segmental estimation dataset cluster
聚簇序号	聚簇 6	聚簇 7	聚簇 8	聚簇 9	聚簇 10
聚类标签	image K-means algorithm pattern recognition	convergence K-means local algorithm group	initialization centers K-means random model	K-means kaufman dataset gene result	optimal cluster K-means algorithm classification
聚簇序号	聚簇 11	聚簇 12	聚簇 13	聚簇 14	聚簇 15
聚类标签	soil K-means cluster model group	fuzzy K-means cluster algorithm group	vector K-means space model codevector	K-means Isolate method recognition algorithm	K-means fuzzy data mining analysis

表14 本文算法聚类标签

聚簇序号	聚簇1	聚簇2	聚簇3	聚簇4	聚簇5
标签	weight K-means cluster variable optimal	analysis cluster algorithm K-means application	cluster method modify segmentation vector	estimate hide markov-models parameter recognition	K-means algorithm cluster convergence generalize
聚簇序号	聚簇6	聚簇7	聚簇8	聚簇9	聚簇10
标签	data K-means fuzzy map classification	strong K-means large law fuzzy	K-means neural gas grow hierarchical	gene cluster heuristic K-means variable-selection	markov method restoration segmental fuzzy
聚簇序号	聚簇11	聚簇12	聚簇13	聚簇14	
标签	genetic algorithm K-means integration self-organizing	fit K-means line straight algorithm	fuzzy K-means cluster trim classification	an-image-hiding base K-means cluster algorithm	

从表中可以看到,算法所生成的聚类标签良莠不齐,有些聚类标签可以在很大程度上揭示聚簇内论文的内容,用户可以从聚类标签上了解到聚簇内容,而有些聚类标签所选用的单词分辨力不够,不足以揭示文本内容。总体来说,K 值较大时所生成的聚类标签更优。另外,标签仍然存在第一次实验中拆分固定短语的问题。

4.3.3 聚类结果的假设检验

从表中可以看到,改进算法的三项聚类评价指标都优于经典算法,不过聚类效果的改进是否达到了统计学上的显著性则需要通过假设检验来证实。本文采用 t 检验对三项聚类指标进行检验。本文将经典算法得到的熵值、平均准确率、凝聚度各9个指标值与本文改进算法得到的对应指标值进行配对 t 检验,得到的检验结果如表15所示。

从表中可以看到,在95%的置信区间下,P 值分别等于0.026、0.023、0.000,都通过了检验。故可以认为,相比经典K-means算法,本文提出的改进算法在聚类效果上的改进具有统计学意义。

表15 经典算法与本文改进算法聚类指标 t 检验结果

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	entropy1 - entropy2	0.0911	0.1016	0.033893	0.012975	0.169290	2.689	8	0.028
Pair 2	accuracy1 - accuracy2	-0.0482	0.05138	0.017129	-0.08778	-0.00878	-2.819	8	0.023
Pair 3	sse1 - sse2	9.2423	4.1788	1.39295	6.03016	12.4544	6.635	8	0.000

比较算法 2 与本文算法的聚类效果,可以发现,在第一次实验中算法 2 聚类指标优于本文提出的算法,而在第二次实验中则本文算法聚类效果更优。因而可以认为本文算法更适用于数据点较为紧密的数据集,而在相对松散的数据集上,算法 2 则更为优秀。由于文献检索结果集一般是一个较为紧密的数据集,因而本文提出的算法更适合应用于检索结果聚类。

5 结 语

本文以学术文献作为聚类对象,提出了一种改进的 K-means 算法。该算法利用学术文献的同被引信息,较好地解决了 K-means 算法的聚类数和初始聚类中心选择时存在的随意性问题。本文在聚类中采用学术文献间同被引信息与文本信息作为聚类的数据源,有利于多角度、更加充分地揭示学术文献的内容信息。实验结果表明,本文的改进思路应用于学术论文检索结果聚类,较之经典的 K-means 算法和基于 AHC 算法的 K-means 算法,都取得了更优的聚类效果。

由于本文提出的算法仅仅考虑了文献之间的同被引关系,因而本算法对于发表时间较短、没有被大量引用的文献则依然不能取得理想的聚类效果,因此下一步的研究工作需要综合考虑文献的引用和被引情况,形成引文网络,然后利用社会网络分析等方法在引文网络上进行挖掘,为文本聚类算法提供更多的辅助信息,更好地改进聚类算法。

另外,本文采用的聚类标签生成方法使用多个词作为聚类标签,但是生成的聚类标签依然存在一些不足,如在选词过程中将一些固定的短语割裂开来,丧失了原文中的语义含义。2009 年,骆雄武等[30]提出了一种基于后缀树模型的聚类标签生成方法,通过在检索结果集中建立一棵后缀树,计算后缀树中短语的得分,选择得分最高的若干短语作为聚类标签,这样的解决方法较好地解决了语义割裂的问题。总而言之,如何生成粒度更细、更精确的聚类标签,仍将是文本聚类研究的重要方向。

参 考 文 献

[1] 章成志,张庆国,师庆辉. 基于主题聚类的主题数字图书馆构建[J]. 中国图书馆学报,2008,34 (6):64-69.

[2] Macqueen J. Some methods for classification and analysis of multivariate observations [C]//Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967,1:281-297.

[3] Selim S Z, Ismail M A. K-means type algorithms: A generalized convergence theorem and characterization of local optimality [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1984,6(1):81-87.

[4] Katsavounidis I, Kuo J, Zhang Z. A new initialization technique for generalized Lloyd iteration [J]. Signal Processing Letters,1994,1(10):144-146.

[5] Khan S S,Ahmad A. Cluster center initialization algorithm for K-means clustering[J]. Pattern Recognition Letters, 2004,25:1293-1302.

[6] Redmond S J,Heneghan C. A method for initialising the K-means clustering algorithm using kd-trees[J]. Pattern Recognition Letters,2007,28(8):965-973.

[7] Krishna K,Murty M N. Genetic K-means algorithm[J]. Systems,Man and Cybernetics,1999,29(3):433-439.

[8] Maulik U, Bandyopadhyay S. Genetic algorithm-based clustering technique[J]. Pattern Recognition, 2000, 33 (9):1455-1465.

[9] Halkidi M,Vazirgiannis M,Batistakis Y. Quality scheme assessment in the clustering process[C]//PKDD, Berlin Heidelberg:Springer,2000:265-276.

[10] Jung Y. Design and Evaluation of Clustering Criterion for Optimal Hierarchical Agglomerative Clustering [D]. University of Minnesota,2001.

[11] 胡晓庆,马儒宁,钟宝江. 层次聚类算法的有效性研究[J]. 山东大学学报:工学版,2010,40(5):146-153.

[12] Small H. Co-citation in the scientific literature: A new measure of the relationship between two documents[J]. Journal of the American Society for Information Science, 1973,24(4):265-269.

[13] Sullivan D,White D H,Barboni E J. Co-citation analyses of science:An evaluation[J]. Social Studies of Science, 1977,7(2):223-240.

[14] Chen C M,Ibekwe-SanJuan F,Hou J H. The structure and dynamics of co-citation clusters: A multiple perspective co-citation analysis [J]. Journal of the American Society for Information Science and Technology,2010,61(7):1386-1409.

[15] 崔雷,胡海荣,李纪宾. 文献计量学共引分析系统设计与开发[J]. 情报学报,2000,19(4):308-312.

[16] 赵悦阳,崔雷. 专题文献的同被引聚类分析在表现学科专业发展历史的可靠性评价[J]. 情报学报,2005, 24(4):414-421.

[17] 王孝宁,崔雷. 2001-2006 年国际情报学研究的引文分析[J]. 情报学报,2007,26(3):399-407.

[18] 邱均平. 信息计量学. 武汉:武汉大学出版社,2007:318-319.

[19] Aljaber B, Stokes N, Bailey J, et al. Document clustering of scientific texts using citation contexts[J]. Information Retrieval, 2010, 13(2): 101-131.

[20] 章成志,师庆辉,薛德军. 基于样本加权的文本聚类算法研究[J]. 情报学报,2008,27(1):42-48.

[21] 赵妍,侯汉清,耿金玉,等. 中文期刊论文自动标引加权设计研究[J]. 新世纪图书馆,2004(1):40-43.

[22] 赵云志. 统计分析法自动标引的改进[J]. 情报学报,2000,19(4):333-337.

[23] 孟海涛,陈笑蓉. 基于模糊相似度的科技文献软聚类算法[J]. 贵州大学学报:自然科学版,2007,24(2):175-178.

[24] 刘海峰,王元元,张学仁,等. 文本分类中基于位置和类别信息的一种特征降维方法[J]. 计算机应用研究,2008,25(8):2292-2294.

[25] 赵纪元,罗霄. 面向中图法的学术文献自动分类研究[C]//中国中文信息学会. 中国计算机语言学研究前沿进展(2007-2009),2009:507-513.

[26] 韩客松,王永成. 中文全文标引的主题词标引和主题概念标引方法[J]. 情报学报,2001,20(2):212-216.

[27] All-smart-stop-list[EB/OL]. http://jmlr.csail.mit.edu/papers/volume5/lewis04a/all-smart-stop-list/. 2005-12-01/2005-01-08.

[28] Krishna K, Murty M N. Genetic K-means algorithm[J]. Systems, Man and Cybernetics, 1999, 29(3): 433-439.

[29] 孙爱香,杨鑫华. 关于文本聚类有效性评价的研究[J]. 山东理工大学学报:自然科学版, 2007,21(5):65-68.

[30] 骆雄武,万小军,杨建武,等. 基于后缀树的Web检索结果聚类标签生成方法[J]. 中文信息学报,2009,23(2):83-88.

(责任编辑 许增棋)