情报学报 ISSN 1000-0135
第28卷第2期220-224，2009年4月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol.28 No.2，220-224 April 2009

doi:10.3772/j.issn.1000-0135.2009.02.010

基于关键词和摘要相关度的文献聚类研究[1)]

魏建香[1,2] 苏新宁[1]

（1. 南京大学信息管理系，南京 210093；2. 南京人口管理干部学院信息科学系，南京 210042）

摘要 现有的文献聚类方法都是通过文献关键词来进行的。本文在研究大量文献聚类方法的基础上，提出了一种通过文献关键词和摘要进行加权的新的文献聚类算法。首先，改进了传统相似度计算的方法，设计出基于关键词和摘要词加权的相似度公式，使文献相似度计算更加精确。其次，基于“文献距离越大，聚为一类的概率越小”的思想，提出了一种“最大距离聚类法”，并给出了算法的详细步骤。最后，实现算法并进行了大量的实验仿真。通过改进相似度计算公式，调整关键词和摘要词的权重，提高了聚类的质量。结果表明，本文提出的文献聚类算法是一种行之有效的方法。

关键词 文献聚类 相似度 关键词 摘要 最大距离

A New Document Clustering Algorithm Based on Keywords and Abstract Correlation

Wei Jianxiang[1,2] and Su Xinning[1]

(1. *Department of Information Management, Nanjing University, Nanjing* 210096;
2. *Department of Information Science, Nanjing College for Population Programme Management, Nanjing* 210042)

Abstract All document clustering methods are based on keywords now. By researching a lot of methods for document clustering, a new dynamic method is presented, which is based on the idea that the longer is the distance between two documents, the lesser probability that they can be classified in same class. Documents' similar matrix is computed accurately by a new formula based on keywords and abstract correlation. The steps of the algorithm are given in detail. Experimental results show that this is an effective method and the quality of clustering is improved by combining keywords with abstract's weight.

Keywords document clustering, similar measurement, keywords, abstract, maximum distance

1 引 言

随着信息社会的发展，人们追求知识的渴望变得越来越强烈，从文献数据库检索所需的文献成为人们学习的一个重要的途径和手段。目前，许多文献数据库已经建成并投入使用，其中包含了大量丰富的文献数据。面对庞大的文献数据库，用户需要花费大量的时间寻找适合的文献，其心情可以用“淹没于信息，却饥饿于知识”来概括人们对信息爆炸的恐惧。如何帮助用户快速而准确地定位到所需资源，更好地为用户提供优质服务已经成为图书情报学界迫切需要解决的问题。将检索到的文献进行自动分类是解决此问题的一种行之有效的方法，而文献聚类是文献分类的前提。本文正是在这种背景下，提出了一种基于文献关键词和摘要相关度的新

收稿日期：2008年8月20日

作者简介：**魏建香**，南京大学信息管理系博士生，南京人口管理干部学院信息科学系副教授，副主任，从事人工智能、数据挖掘研究。E-mail:jxwei@foxmail.com。**苏新宁**，南京大学信息管理系博士生导师，教授，从事信息处理与检索、知识管理、引文分析等。

1） 国家自然科学基金(40771163)、江苏省高校青蓝工程“优秀青年骨干教师”基金(2004～2008)资助项目。

的文献聚类算法。

在文献聚类算法中,高维数据的处理是影响算法效率的一个很重要的方面。我们通过对2005年CSSCI文献数据库中图情档三类学科文献的统计发现,在3 932篇文献中有14 202个关键词,平均每篇文献有3.61个关键词,互异的关键词达到6 708个。利用普遍认可的VSM(向量空间模型)来计算文献相似度矩阵时,特征空间的维度很大,增加了计算的复杂性。因此,许多文本聚类方法的研究者都在考虑特征空间的降维问题。文献[1]利用基于关系的方法引入合适的相似空间代替原始空间;文献[2]提出了一种可任意分配的低维度子空间的方法;文献[3]将关键词映射到概念级来降低维数;文献[4]则通过二次特征提取来降低维度;文献[5]提出了一种基于语义相似度的文本聚类方法,采用概念列表表示文档来降低维度;文献[6]提出了一种VHDR方法减少维度。由于某些关键词在所有文献中出现的频次很低,它们对文献的特征值的影响几乎为零,因此本文采用消除频次较低的关键词的方法来实现降维。

在相似度计算方面,文献[7]通过文献标题、关键词和摘要合并形成特征向量空间来提高文献表示的精度,但大大增加了计算的维度。同时,传统的文本聚类方法都是将文档表示成关键词特征空间中的一个向量,其取值非0即1,没有考虑关键词部分的相似性,也没有考虑文献的摘要对文献相似度的贡献。本文正是基于以上两点,在不增加特征向量空间维数的情况下,提出了基于关键词和摘要词加权的文献相似度计算方法,并且考虑了关键词之间的部分相似性,使文献相似度矩阵不再为稀疏矩阵,提高了相似度计算的精度。

在目前的文档聚类算法中,主要是基于"距离越小,聚为一类的可能性越大"的思想,本文从相反的角度出发,提出了一种基于"距离越大,聚为一类的概率就越小"的"最大距离聚类"方法。

2 相关的定义

文献聚类的首要步骤是建立向量空间模型,通过将文献表示成特征空间向量来构造文献相似矩阵。传统的相似度计算的方法主要分两个步骤:① 用所有文献的关键词构成多维向量特征空间,通过每一篇文献在特征空间的取值来构建文献-关键词矩阵,该矩阵为稀疏矩阵;② 通过距离公式,如余弦、欧氏距离等公式来计算两篇文献之间的距离,构建文献相似矩阵,该矩阵为对称矩阵。在高维空间中,由于关键词的个数非常有限(通常为3~5个),因此文献-关键词矩阵的大部分数值为0,其他为1。这种方法忽视了关键词之间的相似性以及摘要词对关键词相似度的贡献。为了改善这一状况,本文在原来的方法中增加了一个步骤,即在计算文献-关键词矩阵之前先计算关键词之间的相似度并通过摘要词进行加权。本文中的定义建立在以下假设的基础上:设待聚类的文献总数为 n,所有文献互异关键词的总数为 m,关键词集合 $W=\{kw_1,kw_2,\cdots,kw_m\}$。

定义1:假设关键词 kw_i 和 kw_j,字符长度分别为 l_i 和 l_j,最大连续相同的字符串长度为 l,则关键词相似度可以表示为 $Q(kw_i,kw_j)$:

$$Q(kw_i,kw_j)=\begin{cases}\dfrac{l}{l_i+l_j-l} & \text{当 } l>=4\\ 0 & \text{当 } l<4\end{cases} \tag{1}$$

显然有 $Q\in[0,1]$。

定义2:n 篇文献在 m 维空间的特征向量构成了文献-关键词矩阵($n\times m$),定义为:

$$\begin{bmatrix} Q'_{11} & Q'_{12} & \cdots & Q'_{1m}\\ Q'_{21} & Q'_{22} & \cdots & Q'_{2m}\\ \cdots & \cdots & \cdots & \cdots\\ Q'_{n1} & Q'_{n2} & \cdots & Q'_{nm}\end{bmatrix} \tag{2}$$

其中,Q'_{ij} 定义如下:

$$Q'_{ij}=\lambda\times\frac{\sum_{s=1}^{\alpha}Q(kw_{is},kw_j)}{\alpha}+(1-\lambda)\times\frac{\sum_{t=1}^{\beta}(Q(abs_{it},kw_j)\times T(abs_{it})\times W(abs_{it}))}{\sum_{t=1}^{\beta}(T(abs_{it})\times W(abs_{it}))} \tag{3}$$

其中,λ 为平衡系数,取值在0~1之间,用来调节关键词和摘要词的权重。$Q(kw_{is},kw_j)$ 表示文献 i 的第 s 个关键词 kw_{is} 与关键词集合 W 中第 j 个关键词的相似度(计算公式见定义1);$Q(abs_{it},kw_j)$ 表示文献 i 的第 t 个摘要词与 W 中第 j 个关键词的相似度;$T(abs_{it})$ 是文献 i 的第 t 个摘要词在文献摘要中出现的频次;$W(abs_{it})$ 是文献 i 的第 t 个摘要词的权重,它的值取该摘要词与该文献所有关键词相似度的最大值。由于 Q 的取值在[0,1]之间,显然 $Q'_{ij}\in[0,1]$。Q'_{ij} 的定义考虑了将关键词和摘要词综合加权来表示文献在特征空间中的取值。

通过定义1和定义2得到的文献-关键词矩阵将不再是大部分数值为0的稀疏矩阵，这种方法将进一步增加文献相似度计算的准确度。

定义3：本文用欧氏距离表示文献相似度。文献相似度矩阵 $S(n \times n)$ 定义为：

$$\begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1n} \\ S_{21} & S_{22} & \cdots & S_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ S_{n1} & S_{n2} & \cdots & S_{nn} \end{bmatrix} \quad (4)$$

此矩阵为对称矩阵，对角线上的数值为0，其中：

$$S_{ij} = \sqrt{\sum_{k=1}^{m} (Q'_{ik} - Q'_{jk})^2}。 \quad (5)$$

定义4：在聚类操作中，调整类中心点的方法通常采用重心法。类重心公式定义如下：

$$x^* = \sum_{x_i \in c_i} | x_i - x_i^* | / n_i \quad (6)$$

其中，x^* 表示类 c_i 的新的类中心，x_i^* 表示类 c_i 的原中心，n_i 表示类 c_i 中文献的个数。

定义5：聚类评价函数采用分类正确率表示，其定义如下：

$$E = \frac{1}{k} \sum_{i=1}^{k} (n'_i / n_i) \quad (7)$$

其中，n_i 表示类 c_i 中实际的文献个数，n'_i 表示被划分在类 c_i 中正确的文献个数。k 为总的类别数。显然 E 的值越大，聚类的效果越好。

3 聚类算法的设计

聚类算法不依赖人工预先定义和标注的训练文本集进行样本学习，而是通过“自我观察”进行自我学习。聚类算法主要分为层次聚类和划分聚类两大类。层次聚类将对象组织成为一个树，每个叶子节点是一个簇，该树称为一个聚类谱系图。层次聚类算法复杂度高，但被认为效果比较好，在聚类过程中只需扫描所有聚类样本数据一次。该类算法中一旦一个合并或者分裂被执行就不能再修改，运行速度比较慢，不适合于大样本数据。该算法的设计是基于“最小距离”、“从小到大”进行合并的思想。基于划分的聚类算法不需要建立类别之间的层次结构，只需将数据集进行划分，形成一个平面的类结构。这种聚类都有一个初始划分假设，任何样本都可以作为种子(Seed)，即初始化的聚类中心点，然后在此基础上不断地更新聚类划分。这种聚类的算法很多，相对于层次聚类而言，大多具有简单、复杂度低的特点。但该类算法的结果依赖于初始中心点的选择，具有贪心性。综合上述分析，在目前的算法中，主要是基于“距离越小，聚为一类的可能性越大”的思想。本文从相反的角度提出的算法，称为“最大距离聚类法”，此算法主要是基于“两篇文献之间的距离越大，聚为一类的概率就越小”的思想。由于本文算法中初始中心点的选择是确定的，因此能够避免划分聚类算法贪心性的不足。具体的算法设计如下：

假设有 n 篇文献，通过公式(1)～(5)，计算出任意两篇文献的相似度，将此称为文献距离。

(1)指定适当的阈值 t，计算文献相似度矩阵 S。

(2)取 $S_{max} = \max\{S_{ij} \mid S_{ij}$ 是相似矩阵 S 中的任意元素$\}$。如果 S_{max} 小于阈值 t，则所有文献聚为一类，算法终止，否则继续。

(3)记录距离为 S_{max} 的两篇文献分别为 i, j。分别以 i、j 为类中心，对于小于阈值 t 的所有距离，采用最小值优先聚类的原则进行聚类，形成两个新类别 C_1 与 C_2。在聚类时，在类别中每次增加一篇文献时，就用公式(6)调整类中心点。

(4)两次聚类完成后，将已经聚类的文献与其他文献之间的距离从集合 S 中删除。如果集合 S 为空，则算法终止，否则继续。

(5)将 S_{max} 从集合 S 中删除，$k = 3$，转步骤(2)。

该算法的步骤示意图如图1所示。

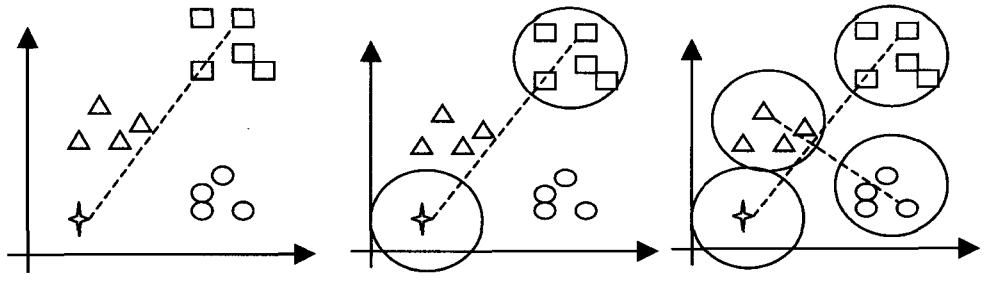

图1 算法的示意图

算法分析：由于本文提出的聚类算法是基于最大距离来设计的，因此，初始中心点有两个且为边缘点。算法具有以下特点：

(1)在样本空间确定的情况下，初始中心点是确定的，因此该算法不具有贪心性。

(2)为了避免将边缘点作为最终的类中心点而影响聚类质量，算法中采用了重心法对中心点进行调整，因此该算法能发现非球状的聚类点。

(3)因为初始中心点为边缘点，因此本算法可以快速地发现孤立点，如图中的星型点。

(4)与层次聚类相似，算法开始时，需要全局计算。

4 实验仿真

为了测试算法的可行性和有效性，从2005年图情档文献数据中选择207篇文献作为测试样本，包括文献学70篇、情报学51篇、图书馆学86篇。

4.1 数据预处理

数据预处理主要分为以下步骤：

(1)数据降维。为了降低计算的复杂性，去掉样本数据中词频为1的关键词。样本数据中，207篇文献共有关键词989个，互异关键词629个，通过消除关键词频次为1的降维方法，剩余关键词112个，大大降低了数据维度。

(2)文献摘要分词。将文献摘要进行分词，去掉虚泛词。

(3)选择摘要词。计算摘要词与该篇文献关键词的相似度，并取最大值作为公式(3)中的 $W(abs_{ii})$。处理完所有摘要词后，去掉相似度为0的摘要词，并计算剩余摘要词的频次作为公式(3)中的 $T(abs_{ii})$。

(4)计算文献-关键词矩阵。利用公式(1)计算文献关键词与特征关键词的相似度，再利用公式(3)计算出文献-关键词矩阵的所有分量。

4.2 仿真及分析

通过数据预处理后，利用本文给出的最大距离聚类算法进行实验仿真。

首先，在不考虑摘要词加权的情况下，即平衡系数 $\lambda = 1$ 时，仿真结果如图2所示。

从图2可以看出，由于初始阈值很小，也就是划分区域半径小，这时划分的类别数目较多，导致分类的正确率较低。当阈值不断增大时，分类的正确率明显提高，但当阈值超过一定值($t = 0.28$)后，由于划分类别的区域半径增加，划分的类别数目变小，因而也会导致分类正确率降低。当阈值达到0.45时，所有样本都划分在一个类别中。

在没有摘要词参与加权的仿真结果中，最大距离算法获得了较高的正确率。最大的正确率是98.37%，但离100%的目标还有一定的差距。下面针对平衡系数的调整对算法进行仿真。根据图2的仿真结果，阈值的最佳取值在[0.1, 0.28]之间。

从表1中可以看出，阴影部分的正确率达到95%以上。在一定的范围内，随着平衡系数的减小，

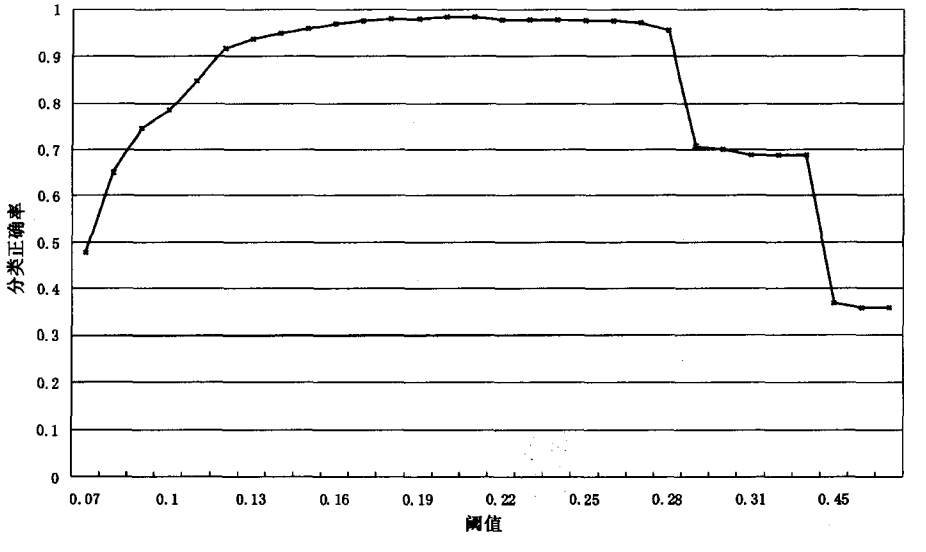

图2 平衡系数为1的仿真结果

表1 仿真结果

平衡系数 \ 阈值 \ 正确率	0.10	0.12	0.14	0.16	0.18	0.20	0.22	0.24	0.26	0.28
1.0	0.7841	0.9166	0.9481	0.9683	0.9789	0.9837	0.9772	0.9772	0.9754	0.9558
0.9	0.7999	0.9167	0.9280	0.9545	0.9584	0.9584	0.9584	0.9518	0.9548	0.9527
0.8	0.8112	0.9102	0.9462	0.9584	0.9584	0.9697	0.9649	0.9631	0.9548	0.9495
0.7	0.8760	0.9188	0.9545	0.9584	0.9697	0.9697	0.9649	0.9631	0.9518	0.6667
0.6	0.9280	0.9545	0.9584	0.9697	0.9744	0.9932	0.9935	0.9869	0.6797	0.6667
0.5	0.9545	0.9631	0.9792	0.9905	1.0000	1.0000	0.7059	0.6797	0.6667	0.6667
0.4	0.9679	0.9905	1.0000	1.0000	0.7451	0.6928	0.6732	0.6667	0.6667	0.6667
0.3	1.0000	1.0000	0.9869	0.6928	0.6824	0.6628	0.6550	0.6318	0.6008	0.5698
0.2	1.0000	0.7412	0.6955	0.6928	0.6824	0.6628	0.6550	0.6318	0.6008	0.5698

即摘要词权重的加大，从平衡系数等于0.5开始，最大正确率达到100%，这说明摘要词对文献聚类效果具有一定的贡献作用。当选择一定的平衡系数后，文献之间的相似度更加精确，此时当相应的阈值较小时，聚类的精度更高。但另一方面，如果摘要词权重太大且阈值太小时，会导致划分类别数的增加而影响分类的正确率，因此，必须选择适当的阈值和平衡系数。从观察得到，平衡系数的最佳取值范围为[0.3,0.5]，阈值的最佳范围为[0.1,0.2]。我们取平衡系数为0.4，阈值为0.15进行了仿真，分类的正确率为100%。

5 结 论

为了解决文献自动聚类的问题，本文提出了一种最大距离文献聚类方法。与传统的聚类算法不同，该算法是从逆向思维的角度，基于“文献之间的距离越大，聚为一类的概率就越小”的思想进行设计的。实验仿真表明，该算法是一种实际有效的聚类算法。

本文的创新点在于：① 提出了一种新的有效的文献聚类算法；② 在相似度计算时，改变传统的相似度取值为0或1的做法，运用关键词部分相似性来增加文献之间的相似程度；③ 改变了传统文献聚类方法中仅利用关键词进行聚类的方法，通过摘要词和关键词加权并利用平衡系数进行调节的方法计算文献在特征空间的取值。

参 考 文 献

[1] Alexander Strehl, Joydeep Ghosh. Relationship-based clustering and visualization for high-dimensional data mining [J]. INFORMS Journal on Computing, 2003,15 (2): 208-230.

[2] Aggarwal C C, Yu P S. Redefining clustering for high-dimensional applications [J]. IEEE Transactions on Knowledge and Data Engineering, 2002,14 (2): 210-225.

[3] 杨彩莲，谢福鼎. 基于主题概念聚类的中文文本聚类[J].现代电子技术，2007,22:161-163.

[4] 孙学刚，陈群秀，马亮. 基于主题的Web文档聚类研究[J].中文信息学报，2003,17(3):21-26.

[5] 孙爽，章勇.一种基于语义相似度的文本聚类算法[J].南京航空航天大学学报，2006,38(6):712-716.

[6] Yang J, Ward M O, Rundensteiner E A, et al. Visual hierarchical dimension reduction for exploration of high dimensional datasets[C]// Proceedings of the Symposium on Data Visualisation. Grenoble, France, May 26-28,2003:19-28.

[7] 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.

(责任编辑 许增棋)