文章编号: 1003-0077(2008)01-0051-05

改进的 OPTICS 算法及其在文本聚类中的应用

曾依灵[1,2],许洪波[1],白 硕[1]

(1. 中国科学院 计算技术研究所 智能安全中心,北京 100080;2. 中国科学院 研究生院,北京 100080)

摘 要: 基于密度的 OPTICS 聚类算法以可视化的结果输出方式直观呈现语料结构,但由于其结果组织策略在处理稀疏点时的局限性,算法实际性能未能得到充分发挥。本文针对此缺陷提出一种有效的结果重组织策略以辅助稀疏点的重新定位,并针对文本领域的特点改变距离度量方法,形成了 OPTICS-Plus 文本聚类算法。在真实文本分类语料上的实验表明,我们的结果重组织策略能够辅助算法产生更为清晰反映语料结构的可达图,与 K-means 算法的比较则证实了 OPTICS-Plus 具有较为良好的聚类性能。

关键词: 计算机应用;中文信息处理;OPTICS 算法;密度聚类;文本挖掘

中图分类号: TP391 **文献标识码**: A

OPTICS-Plus for Text Clustering

ZENG Yi-ling[1,2], XU Hong-bo[1], BAI Shuo[1]

(1. Research Center of Information Intelligence and Information Security,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China;
2. Graduate University, Chinese Academy of Sciences, Beijing 100080, China)

Abstract: As a density-based clustering algorithm, OPTICS is capable of showing the intrinsic corpus structure within a visual plot. However, due to the improper strategy in organizing the points in sparse space, the algorithm does not reach its best performance. To solve this problem, we proposed an effective result-reorganization strategy for reordering those sparse points. Based on this strategy, a new text clustering algorithm named OPTICS-Plus was proposed according to the characteristic of text mining fields. Experiment on FuDan text classification corpus shows that our result-reorganization strategy is capable of helping the reachability plots generating clearer views of corpus structures. Furthermore, a comparison with K-means proves that the clustering performance of OPTICS-Plus is actually satisfactory.

Key words: computer application; Chinese information processing; OPTICS; density-based clustering; text mining

1 引言

随着网络的飞速发展,越来越多的电子文档触手可及。如何分析和管理大规模的文本数据成为日益急切的需求。聚类分析作为一种重要的数据分析方法,能够很好地满足这方面的需求。它能挖掘语料的潜在结构,将文档划分成有意义的子簇,协助人们更好地对大规模文本进行理解,同时也能作为一种有效的预处理步骤,为进一步的文本分析提供初步的语料结构。随着信息检索的发展,它已被成功地应用到加速检索过程、文档检索结果聚类呈现、话题的自动发现、文本摘要等方面[1~3],在文本挖掘领域扮演着日益重要的角色。正因如此,聚类算法的相关研究一直是深受关注的热点。

一直以来,研究者对如何提高聚类算法的性能

收稿日期: 2007-05-02 **定稿日期**: 2007-12-03

基金项目: 国家 973 资助项目(2004CB318109)

作者简介: 曾依灵(1980—),男,研究实习员,博士生,主要研究方向为大规模文本处理、文本表示、文本聚类等;许洪波(1975—),男,博士,副研,主要研究方向为大规模文本处理、互联网搜索,文本过滤等;白硕(1956—),男,研究员,主要研究方向为计算语言学、数据挖掘、网络安全等。

费尽心思,因为性能的好坏通常是衡量聚类算法优劣的重要标准。然而,很多时候更为本质的问题在于语料本身,大多数聚类算法对语料非常敏感,在一个语料上取得优良性能的算法常常在另一个语料上效果不尽如人意。与经典聚类算法重点关注性能不同,Ankerst M 等于 1999 年提出的 OPTICS (Ordering Points To Identify the Clustering Structure)算法更为关注如何直观地反映语料自身的潜在结构[4]。

OPTICS 是一种基于密度的聚类算法,它从一个随机选定的对象出发,朝着数据最为密集的区域扩张,最终将所有对象组织成一个能够反映语料结构的可视化有序序列。然而,由于 OPTICS 算法自身策略的局限,低密度区域的对象往往被累积在结果序列的末尾,使算法的性能未能充分体现。针对此不足,我们在本文中提出一种有效的结果重组织策略,使稀疏区域的对象能够更合理地与离自己最近的高密度区域的对象相邻。在该策略的基础上,我们针对文本领域的特点对算法进行局部调整,形成改进的 OPTICS-Plus 文本聚类算法。

本文后面的内容组织如下:第二节首先介绍聚类方面的相关工作,包括聚类算法的简单综述以及 OPTICS 算法的详细介绍,第三节阐述我们对 OPTICS 算法的相关改进工作,在第四节中,我们将 OPTICS-Plus 和 OPTICS 算法在复旦语料上进行聚类性能对比实验,并最终在第五节进行全文总结。

2 相关工作

2.1 聚类算法

作为一种无导的学习方法,聚类在无类别标记信息下将事物自动分组,使每个分组能进行自我识别并且区分于其他分组。该过程的数学描述为:对集合 $X=\{x_1, x_2, \cdots, x_n\}$ 进行划分,分成为 $C=\{C_i \mid C_i \subset X, i=1, \cdots, m, C_1 \cup \cdots \cup C_m = X\}$,$C_i$ 称为一个簇,有时也称为类。

常用的聚类算法[5]一般可以分为层次式聚类(如自底向上进行层次组织的凝聚式聚类,自顶向下进行层次组织的分裂式层次聚类)、划分式聚类(如以类内点均值作为中心的 K-Means 以及以实际点作为中心的 K-Medoids)、基于网格的聚类(如利用存储在网格中的统计信息进行操作的 STING,利用小波变换聚类对象的 WaveCluster,以及 CLIQUE)、基于密度的聚类(如 DBScan 和 OPTICS)等。在文本挖掘领域中,目前较为常用的是划分式聚类和凝聚式聚类。由于基于密度的 OPTICS 聚类能以可视化的方式直观反映语料结构,有利于进一步的文本分析,我们针对其局限性进行相应改进后将其引入文本领域。下一小节中,我们首先介绍 OPTICS 算法。

2.2 OPTICS 算法

OPTICS 是由 DBScan 发展而来的一种密度聚类算法。密度聚类的核心思想是用一个点的 ε 邻域内的邻居点数衡量该点所在空间的密度。如果 ε 邻域邻居数超过某个指定阈值 $MinPts$,则认为该点处于某个簇内,称为核心点(*core-object*),否则认为该点处于某个簇的边界上,称为边界点(*bounder-object*)。下面给出一些定义[6]:

定义 1 直接密度可达(*directly density-reachable*)

如果 p 是核心点,q 在 p 的 ε 邻域内,则 p 直接密度可达 q。

定义 2 密度可达(*density-reachable*)

如果存在序列 $p_1, \cdots, p_n$,其中 $p_1 = p, p_n = q$,并且对于任意 $1 \leq i < n$,p_i 直接密度可达 p_{i+1},那么,p 密度可达 q。

定义 3 密度相连(*density-connected*)

如果 o 密度可达 p,且 o 密度可达 q,则 p 和 q 密度相连。

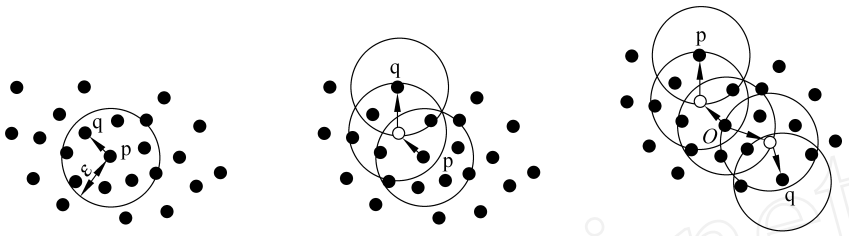
三者关系如图 1 所示:密度可达是直接密度可达的传递;密度相连则是从同一点密度可达的任意两点的对称关系。由此,如果从某个选定的核心点出发,不断向密度可达的区域扩张,将得到一个包括核心点和边界点的最大化区域,区域中任意两点密度相连,这即为一个聚类簇。DBScan 算法就是通过上述过程搜索和提取尺度为 ε 的所有簇[6]。为了具备更为精细的刻画能力,OPTICS 引入了核心距离和可达距离的概念[4]:

定义 4 核心距离(*core-distance*)

假定点 p 包含 $MinPts$ 个邻居的最小半径为 $MinPts\text{-}distance(p)$,那么 p 的核心距离定义为:

$$core\text{-}distance_{\varepsilon, MinPts}(p) = \begin{cases} Undefined, if\ p\ is\ not\ a\ core\text{-}object \\ MinPts\text{-}distance(p), otherwise \end{cases}$$

也就是说,核心距离是一个点成为核心点的最小邻域半径。



(a) 直接密度可达　　(b) 密度可达　　(c) 密度相连

图 1　直接密度可达、密度可达和密度相连

定义 5　可达距离(*reachability-distance*)

假定 p 是某点 o 的 ε 邻域中的点,那么 p 与 o 相关的可达距离定义为:

$$reachability\text{-}distance_{\varepsilon,MinPts}(p,o)=\begin{cases}Undefined, & if\ p\ is\ not\ a\ core\text{-}object\\ \max(core\text{-}distance(o), distance(o,p)), & otherwise\end{cases}$$

可见,p 与 o 相关的可达距离即是从 o 直接密度可达 p 的最小距离。该距离与空间密度直接相关,如果该点的所在空间密度大,它从相邻点直接密度可达的距离就小,反之亦然。如果我们想要朝着数据尽量稠密的空间进行扩张,那么可达距离最小的点是最佳的选择。为此,OPTICS 算法用一个可达距离升序排列的有序种子队列(*OrderSeeds*)存贮待扩张的点,以迅速定位稠密空间的数据对象。算法具体过程如下:

Step 1: 有序种子队列初始为空,结果队列初始为空;
Step 2: 如果所有点处理完毕,算法结束;否则选择一个未处理对象放入有序种子队列:
Step 3: 如果有序种子队列为空,返回 Step 2,否则选择第一个对象 p 进行扩张:
　Step 3.1: 如果 p 不是核心点,转 Step 4;否则,对 p 的 ε 邻域内任一未扩张的邻居 q:
　　Step 3.1.1: 如果 q 已在有序种子队列中且从 p 到 q 的可达距离小于旧值,则更新 q 的可达距离,并调整 q 到相应位置以保证队列的有序性;
　　Step 3.1.2: 如果 q 不在有序种子队列中,则根据 p 到 q 的可达距离将其插入有序队列;
Step 4: 从有序种子队列中删除 p,并将 p 写入结果队列中,返回 Step 3。

图 2　OPTICS 算法

算法的时间复杂度为 $O(n^2)$。以结果队列中的点序列作为横坐标,可达距离作为纵坐标,将得到图 3 所示的可达图(*reachability plots*)。图中可达距离小的波谷区域代表数据稠密的簇内的点;波峰区域则代表数据稀疏的边界点。因此,可以在遍历可达图的过程中,通过陡峭下降区域(*steep down region*)和陡峭上升区域(*steep up region*)来辨别和提取聚类簇[4]。

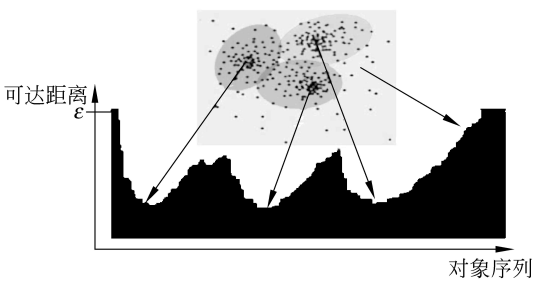


图 3　可达图

3　OPTICS-Plus 算法

在 OPTICS 算法的有序队列中,待扩张的数据点按可达距离升序排列。执行时,算法总是选择可达距离最小的点进行处理,也就是说,总是沿着尽可能稠密的区域进行扩张。直至处理完当前稠密区域,它才会探索稀疏的边界,进入下一个稠密区域。而一旦进入新的稠密区域,未处理的稀疏点将再次搁置,直至没有比它们更为稀疏的数据点为止。因此那些无法连向下一个稠密区域的稀疏点,常常被搁置在结果队列的末尾。如图 3 所示的语料空间中,圆圈外浅灰色区域的稀疏数据点,就堆积在了可达图的末尾,形成一个缓慢上扬的尾部。

我们认为,这种处理稀疏点的方式不够合理,上述稀疏点处理策略会导致可达图失真,不能完全反映数据空间的真实结构。为进一步提高 OPTICS 算法的性能,我们设计了新的结果重组织策略,该策略基于如下假定:

基本假定: 两个数据点的距离越近,其属于同一簇的概率越大。

本假定的合理性显而易见,OPTICS 算法本身

也是遵循该假定选择可达距离最小的点进行扩张,但由于贪心式的扩张策略使得稀疏点与稠密点的相邻关系断裂,从而导致稀疏点只能堆积在可达图末尾。为弥补这一缺陷,我们为每个对象增加一个 *Referrer* 域,指向最近的相邻数据点,并利用它最终将稀疏点放入与之相邻的稠密区域所在的簇中。

由图 2 的 OPTICS 算法描述可知,有序种子队列中每个数据点都保存着一个动态的可达距离,它随着算法的运行而改变(Step 3.1.1)。这个可达距离记录的是所有已扩张的数据点到它的最小可达距离。如果可达距离发生更新,则预示着与当前点最近的相关点发生改变,为记录这种改变,新增的 *Referrer* 域也需要实时更新。

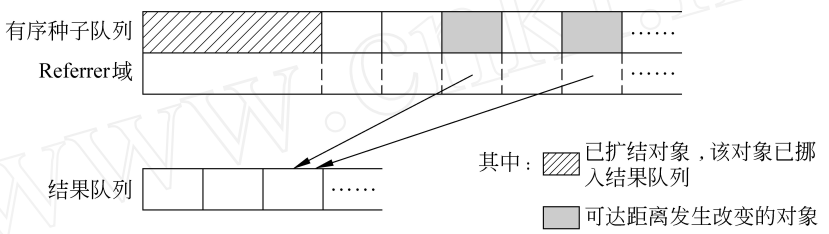


图 4 Referrer 域的更新

如图 4 所示,如果在当前状态下,有序结果序列中的某些点的可达距离发生改变,导致此改变的一定是当前刚完成扩张的数据点,它被放置在结果队列的末尾,因此,只需要将可达距离改变的点的 *Referrer* 域指向结果队列的末尾即可。同理,新加入有序结果序列中的点,也需要将 *Referrer* 域指向结果队列末尾。因此,只需在 OPTICS 算法中 Step 3.1.1 和 Step 3.1.2 后同时加入将 q 指向队尾的操作,该操作无需引入额外的计算代价(修改后的算法不再给出)。

处理完有序种子序列后,所有点的 *Referrer* 域随数据点一并存入结果序列中。我们将在 *Referrer* 域的帮助下将稀疏点移入与其最相邻的点所属的簇中。为提取可达图中的潜在簇,OPTICS 算法运行完后通常需要进行 *ExtractClusters* 操作[4],我们可以在 *ExtractClusters* 执行的同时重新组织结果队列。包含结果重组织的新的 *ExtractClusters* 子算法如下:

Step 1: 将待处理指针指向结果队列头部;
Step 2: 如果待处理指针指向队尾,转 Step 3;否则后移指针:
  Step 2.1: 如果发现陡峭下降区域,开始新簇,将从陡峭下降区域的点存入新簇中;
  Step 2.2: 如果发现陡峭上升区域,寻找对应的陡峭下降区域,该簇提取结束;
  Step 2.3: 对于当前待处理的 p 对象,假定 p 的 *Referrer* 域指向 q:
    Step 2.3.1: 如果 q 属于当前簇,则将 p 加入当前簇;
    Step 2.3.2: 如果 q 不在当前簇中,则将 p 插入 q 所在簇的末尾;
Step 3: 将所有簇中的点首尾拼接,形成新的可达图,算法结束。

图 5 结果重组织和聚类簇提取子算法

如果在算法运行的同时实时记录已处理点到对应簇的映射,寻找 q 所在的簇并将 p 插入簇尾将在常数时间内完成。因此,结果的重组织同样无需引入额外的时间代价。

至此 OPTICS 算法已具备了更为合理的结果重组织策略,为了将其应用到文本领域中,我们还需要进行一些相应修改。与 OPTICS 使用的欧式距离相异,文本领域更偏好 *cos* 相似度,因为它具有以下优点: 度量合理,*cos* 相似度更关注向量方向上的一致性,即特征词的共现的一致性;计算简单,计算归一化向量的 *cos* 相似度只需计算二者的点积;结果规范,*cos* 相似度取值为规范的 $[0,1]$ 区间。为了将 *cos* 相似度应用到 OPTICS 中,需要对算法进行相应的修改: 1) ε 邻域改为 *cos* 相似度大于 ε 的区域;2) 有序种子序列的排序方式改为降序排列;3) 可达距离在大于旧值时更新;4) 类的提取方式以陡峭上升区域开始,陡峭下降区域结束。

我们将经过结果重组织策略和距离度量方法改进的 OPTICS 算法命名为 OPTICS-Plus 文本聚类算法。它适应文本挖掘场景,能够产生更为合理的

结果序列,并拥有与 OPTICS 算法同样的时间复杂度。

4　实验结果及分析

本节中,我们通过实验检验 OPTICS-Plus 文本聚类算法的实际性能。为保证实验结果的可信度,要求实验语料规模不能太小,并能反映实际语料的不均衡特性。实验语料从复旦大学文本分类语料库的测试集中构造,为保证语料规模,我们随机选出 5 个较大的类,并从每个类中随机抽取出数百篇文档,构成文档总数为1 000 篇的语料;为保证不均衡性,语料中每个簇的文档数从 100 篇到 400 篇不等,具体情况如表 1 所示。

表 1　实验语料组成

簇编号	1	2	3	4	5
来源	C3-Art	C11-Space	C19-Computer	C34-Economy	C38-Politics
文档数	200	150	400	150	100

语料经过分词,停用词过滤以及简单的特征选择后转化成 VSM 向量以供聚类实验使用。

为比较 OPTICS-Plus 和 OPTICS 的性能,OPTICS 也需要进行面向文本语料的相关修改。两算法选取同样的参数: ε取较为较小值 0.03 以保证可达图的描述能力, $MinPts$ 取经验值 15。将 OPTICS-Plus算法和 OPTICS 算法应用到实验语料上,得到如图 6 所示的可达图。对比可知,OPTICS-Plus 的可达图具有更清晰的类边界和更短的缓慢下降尾部。这表明,在结果重组织策略的作用下,原本堆积在尾部的稀疏点已被归入更为恰当的簇中。

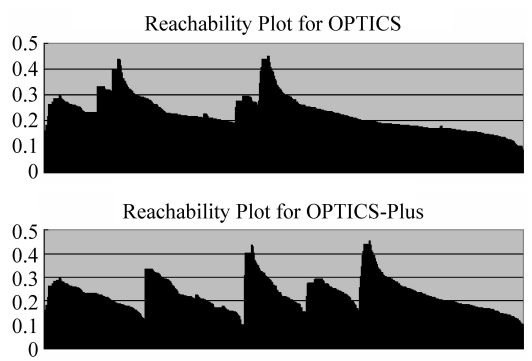


图 6　OPTICS 与 OPTICS-Plus 的可达图

为进一步量化分析实验结果,我们对可达图中提取出的簇进行聚类结果评价。聚类评价指标目前尚不统一,我们参照周昭涛等的分析结果[7],采用较为有效的 Class F、Cluster F 和 Entropy 作为聚类效果评价标准。同时,为使实验结果能与常见文本聚类算法性能相比,我们还加入了 K-means 算法在该实验语料上的运行结果。K-means 算法中,聚类数目 $ClusterNo$ 指定为 5,控制收敛的 $Threshold$ 设置为较小值 0.2。实验结果如表 2 所示。表中数据显示: 与 OPTICS 相比,OPTICS-Plus 在各项指标上都有较为明显的提高,这再次证实了我们的结果重组织策略的有效性;与人为设定正确簇数的 K-means 算法相比,OPTICS-Plus 也能达到与之大体上相等的聚类效果,这表明 OPTICS-Plus 算法的实际聚类性能也较为令人满意。

表 2　不同算法的聚类评价指标比较

聚　类	实验参数	Class F	Cluster F	Entropy
K-means	ClusterNo = 5, Threshold = 0.2	0.611 982 04	0.621 559 89	**0.382 158 87**
OPTICS	MinPts = 15, ε= 0.03	0.565 148 51	0.571 237 70	0.524 920 92
OPTICS-Plus	MinPts = 15, ε= 0.03	**0.633 591 29**	**0.653 111 52**	0.422 920 69

5　结论及下一步研究

OPTICS 作为一种基于密度的聚类方法,能够以可达图的形式清晰反应语料的自身结构,但其结果组织策略在处理稀疏点时具有一定的局限性。本文针对此问题提出了一种改进的结果重组织策略,使稀疏点能够纳入离自己最近的类中。我们进一步调整算法以应用到文本领域,形成了 OPTICS-Plus 文本聚类算法。在真实文本分类语料集上的实验表明,采用结果重组织策略的 OPTICS-Plus 算法的可

(下转第 60 页)

参考文献:

[1] Claudine Santos Badue, Ricardo A. Baeza-Yates, Berthier A. Ribeiro-Neto, Nivio Ziviani. Distributed Query Processing Using Partitioned Inverted Files [A]. String Processing and Information Retrieval [C]. 2001. 10-20.

[2] Byeong-Soo Jeong, Edward Omiecinski. Inverted File Partitioning Schemes in Multiple Disk Systems [J]. IEEE Transactions on Parallel and Distributed Systems, 1995(6), 142-153.

[3] Jinxi Xu, W. Bruce Croft. Cluster-Based Language Models for Distributed Retrieval [A]. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval [C]. 1999. 254-261.

[4] C.J. van Rijsbergen. Information Retrieval [M]. Butterworths, second edition, 1979.

[5] 搜狗搜索引擎(2007) http://www.sogou.com/labs/[DB/OL].

[6] MacQueen, J.B. Some Methods for Classification and Analysis of Multivariate Observations [A]. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability [C]. 1, 1967. 281-297.

[7] Bloom, B. Space/time Trade-offs in Hash Coding with Allowable Errors [J]. Communications of the ACM, 13(7). 422-426.

(上接第 55 页)

达图具有更清晰的簇结构,能够更好地反应语料的自身结构,聚类效果评价指标则从数据上证实了 OPTICS-Plus 算法能够提升聚类性能。提升后的聚类效果可与指定正确簇数的 K-means 算法相比,这表明 OPTICS-Plus 算法的实际性能也较为令人满意。在下一步的工作中,我们准备将 OPTICS-Plus 算法应用到不同类型、更大规模的语料场景中,以深入探索不同语料结构对文本挖掘任务的影响。

参考文献:

[1] Zeng H-J, He Q-C, Chen Z, et al. Learning To Cluster Web Search Results[A]. In: Proceedings of the 27th Int. Conf. on Research and Development in Information Retrieval (SIGIR'04)[C]. July 2004. 210-217.

[2] 宋擒豹, 沈钧毅. 基于关联规则的 Web 文档聚类算法[J]. 软件学报, 2002, 13(3): 417-423.

[3] 李保利, 俞士汶. 话题识别与跟踪研究[J]. 计算机工程与应用, 2003, 39(17).

[4] Ankerst M, Breunig M, Kriegel H-P, et al. OPTICS: Ordering Points to Identify the Clustering Structure [A]. In: Proc 1999 ACM-SIGMOD Int. Conf. on Management of Data [C]. Philadelphia, PA, June 1999. 49-60.

[5] Dubes R C and Jain A K. Algorithms for Clustering Data[M]. Prentice Hall, 1988.

[6] Ester M, Kriegel H-P, Sander J, et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise [A]. In: Proc 1996 Int. Conf. on Knowledge Discovery and Data Mining (KDD'96)[C]. Portland, Oregon: AAAI Press, 1996. 226-231.

[7] 周昭涛. 文本聚类分析效果评价及文本表示研究[D]. 北京: 中科院计算技术研究所. 2005.