

# DBSCAN 算法在电子邮件网络社团发现中的应用

杨芳勋

(四川九洲集团中央研究院 成都 610041)

**摘 要** 针对电子邮件复杂网络中的社团发现问题,将具有良好聚类性能的 DBSCAN 算法引入电子邮件网络社团发现。基于对该算法的分析,研究了电子邮件网络社团发现的系统架构及算法实现流程。最后通过对安然邮件语料集的测试验证了 DBSCAN 算法在社团发现中的可行性。

**关键词** 电子邮件网络,社团发现,DBSCAN 算法

**中图分类号** TP391 **文献标识码** A

## Application of DBSCAN Algorithm in Electronic Mail Network Community Detection

YANG Fang-xun

(Central Research Institute of Sichuan Jiuzhou Group, Chengdu 610041, China)

**Abstract** According to the community detection in the email complex network, the DBSCAN algorithm was introduced into the email network to detect the community in this paper. Based on the analysis of the algorithm, the system architecture and algorithm implementation process of email network community detection were studied. Finally, the feasibility of the DBSCAN algorithm in electronic mail network community detection was verified by the test of the Enron email corpus.

**Keywords** Electronic mail network, Community detection, DBSCAN algorithm

## 1 引言

社会网络、生物系统中的蛋白质网络、科学家协作网络等都可以看作是以复杂网络<sup>[1-2]</sup>的形式存在。复杂网络的一个显著特征就是存在社团结构,社团内部节点之间联系紧密,而社团之间的联系相对稀疏<sup>[3-4]</sup>。近年来,复杂网络中社团结构的挖掘逐步成为一个热门的研究课题<sup>[5]</sup>。互联网时代,电子邮件日益成为人类日常生活中必不可少的通信方式之一,已经成为合作和信息交流的重要途径。私人电子邮件信息能够反映出某个人所涉及的领域以及对该领域的关注程度。如何从复杂的电子邮件链接中准确而迅速地发现社团并获取社团中的重要人物已受到广泛关注。

DBSCAN 算法利用了基于密度的聚类概念,即要求聚类空间中一定区域内所包含对象(点或其他空间对象)的数目不小于某一给定阈值。该算法的显著优点是聚类速度快且能够有效处理噪声点并发现任意形状的空间聚类。本文根据电子邮件网络通信记录,利用 DBSCAN 聚类算法,将邮箱地址划分成不同的社团来实现聚类。

## 2 算法介绍

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) 聚类算法是经典的基于密度的聚类算法,它的目的是寻找数据集中被密度低的区域分隔开的、密度大于给定阈值的高密度区域,并将每个独立的高密度区域组合

成簇,可在噪声的空间数据库中发现任意形状的聚类。

**定义 1(密度)** 空间中任意一点的密度是以该点为圆心、 $Eps$  为半径的圆区域内包含的点的数目。

**定义 2(邻域; Neighborhood)** 空间中任意一点的邻域是以该点为圆心、 $Eps$  为半径的圆区域内包含的点的集合,记作  $NEps(p) = \{q \in D \mid dist(p, q) \leq Eps\}$ ,  $D$  为数据库。

**定义 3(核心点, Core Points)** 如果空间中某一点的密度大于某一给定阈值  $MinPts$ , 则称其为核心点。

**定义 4(边界点, Border Points)** 如果空间中某一点的密度小于某一给定阈值  $MinPts$ , 则称该点为边界点。

**定义 5(直接密度可达到)** 点  $p$  从点  $q$  直接密度可达, 若它们满足:  $p$  处于  $q$  的邻域中, 即  $p \in NEps(q)$ ;  $q$  是核心点, 即  $\beta NEps(q) \geq MinPts$ 。

**定义 6(密度可达到)** 点  $p$  从点  $q$  密度可达, 若  $(p_1, p_2, \dots, p_n)$ , 其中  $p_1 = p, p_n = q$ , 且有  $p_i$  从  $p_{i+1}$  直接密度可达。

**定义 7(密度连接)** 点  $p$  和点  $q$  是密度连接的, 若  $v_o$  使  $p$  和  $q$  都从  $v_o$  密度可达。

**定义 8(类, Cluster)** 数据库  $D$  的非空集合  $C$  是一个类, 当且仅当  $C$  满足以下条件:

对于  $P_{p,q}$ , 若  $p \in C$ , 且从  $p$  密度可达  $q$ , 则  $q \in C$ ;

对于  $P_{p,q}$ , 有  $p \in C$  和  $q \in C$ , 则  $p$  和  $q$  是密度连接的。

**定义 9(噪声, Noise)** 数据库  $D$  中不属于任何类的点为噪声。

考察数据库  $D$  中的某一点  $o$ , 若  $o$  是核心点, 则通过区域查询得到该点的邻域, 若邻域中的点和  $o$  同属于一个类, 这些点将作为下一轮的考察对象 (即种子点), 并通过不断地对种子点进行区域查询来扩展它们所在的类, 直至找到一个完整的类。然后, 依此程序寻找其它的类, 最后剩下的不属于任何类的点即为噪声。

### 3 系统架构

基于电子邮件通信关系的分析, 进行邮件群组关系分析, 具备目标收发邮件中的通信关系进行分析的功能。通过 DBSCAN 聚类算法, 将邮箱地址划分成不同的社团, 提供两层聚类。具体聚类系统架构如图 1 所示。前端采集网络中的电子邮件收发关系, 并将此类数据入库。电子邮件网络社团发现算法访问数据库并提取所需的原始数据进行聚类分析, 得出社团发现结果。界面将分析结果数据进行可视化呈现。

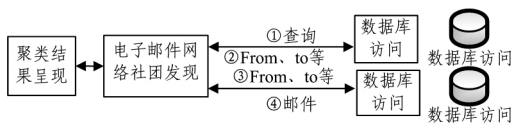


图 1 系统架构

### 4 实现步骤

具体聚类步骤如图 2 所示。

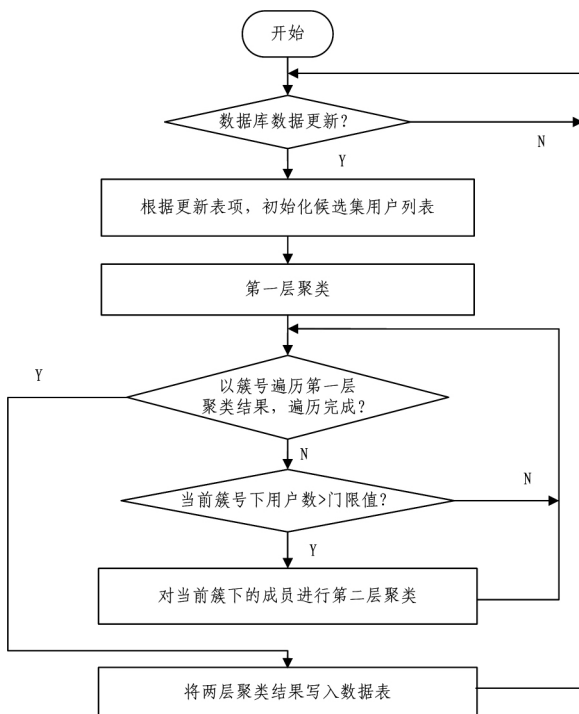


图 2 聚类流程图

首先判断数据库中有无数据更新, 若有, 则初始化候选集用户列表, 否则继续等待数据的更新;

确定候选集用户, 提取数据库表的统计信息, 计算邮件通信的权值 (对通信次数做归一化的处理);

初始化候选集用户列表信息  $setOfPoints$ , 每个用户成员信息包括  $\langle mailIndex, 邻居信息列表, priSetID$  (初始为  $-1$ ),  $subSetId$  (初始为  $-1$ )  $\rangle$ ;

调用 DBSCAN 进行第一层聚类;

对第一层聚类结果进行筛选, 如果聚类用户集合超过门

限值 (可配置) 的用户, 调用 DBSCAN 进行第二层聚类;

将两层聚类结果保存到数据表。

DBSCAN 算法的具体实现过程如下:

步骤 1 逐个扫描邻接表基表中的元素, 判断其是否已被聚类, 如果是则跳过这一元素, 否则开始步骤 2。

步骤 2 判断其是否为核心点, 只有核心点才能发起一次聚类活动。若是则开始步骤 3, 否则跳过这一元素, 留待以后的收尾处理。

步骤 3 开始对这一点进行聚类。赋予该点一个新的簇号并填入变量  $clusterId$  中。然后逐一检索该基表元素后链接的结点链, 从而查找出此点邻域内的所有点, 并对这些点进行判断。设其邻域内的点为  $b$ 。

case1  $b$  已被聚类, 则不对  $b$  进行任何处理;

case2  $b$  未被聚类且是核心点, 则将其作为新种子压栈, 以待以后对其进行递归的聚类处理;

case3  $b$  未被聚类且不是核心点, 则将类号填入  $b$  的  $clusterId$  变量中, 说明  $b$  已经被聚类为  $clusterId$  (不管何种情况下都将  $b$  点标识为已被聚类, 以免以后进行不必要的重复处理)。

步骤 4 从种子栈中取出一个元素, 递归地对其进行聚类。簇号  $clusterId$  不变, 因为这还属于原来的类。如此递归, 直到种子栈为空为止。这时, 标明类号为  $clusterId$  的聚类完成。

步骤 5 再次扫描邻接表基表中的元素, 对步骤 2 中的遗留点进行收尾处理。其中, 判断两点 (两个邮箱地址) 之间是否邻接是通过相似度模块来实现的, 具体有两种计算方法。

(1) 基于非对称二元变量的相似度计算:

$$d(i, j) = \frac{r+s}{q+r+s} \quad (1)$$

其中,  $d(i, j)$  表示评价对象  $i$  和  $j$  之间的相异度,  $q$  表示对于对象  $i$  和  $j$  值都为 1 的变量数目,  $r$  表示对象  $i$  值为 1 而对象  $j$  值为 0 的变量数目,  $s$  是对象  $i$  值为 0 而对象  $j$  值为 1 的变量数目。

(2) 基于欧几里得距离的相似度计算:

$$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2} \quad (2)$$

其中,  $i = (x_{i1}, x_{i2}, \dots, x_{ip})$  和  $j = (x_{j1}, x_{j2}, \dots, x_{jp})$  是两个  $p$  维的邮箱地址。

### 5 实验结果

聚类模型的整体准确率 (Overall Accuracy) 反映了聚类模型正确预测的样本数在预测总数中的比例。精确率 (Precision) 针对每一个应用类别独立考察, 指聚类为一个应用类别的实例当中属于该类别的实例的百分比。

$$OA = \frac{\sum TP \text{ for clusters}}{\text{total number of flows}} \quad (3)$$

$$precision = \frac{TP}{TP + FP} \quad (4)$$

其中,  $TP$  (True Positive) 为实际类型为某一应用的样本中被聚类算法正确预测的样本数。  $FP$  (False Positive) 为实际类型不是某一应用的样本中被聚类算法误判为该应用的样本数。

DBSCAN 算法有 2 个输入参数, 分别为  $minPts$  和  $eps$ 。以安然邮件语料集进行实验, 随机抽取 3000 个样本构成实验

数据子集。整体准确率结果如图 3 所示。

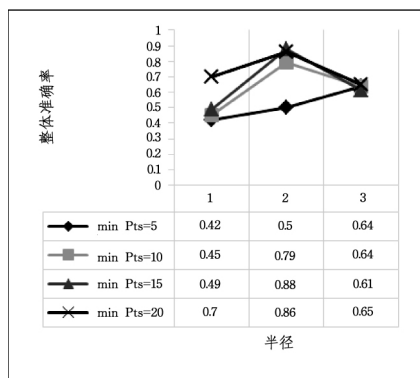


图 3 算法整体准确率

在  $\epsilon$  为 0.02 且  $\min Pts$  为 15 时,实验的整体准确率取得最大值,到达 88%,取得了理想的效果,此时算法的精确率为 0.91,算法的聚类时间平均为 120s。同时,通过实验可知,本文所用算法对输入参数  $\epsilon$ ,  $\min Pts$  较为敏感。

(上接第 574 页)

现有常见工具(如 UPPAAL)所基于的模型(如时间自动机)上没有数据约束;

2)本文的模型中包含的变量既包括连续变量也包括离散变量,现有常见的工具所基于的模型通常只包含离散变量。

如果要在现有工具的基础上实现本文算法,需要把本文所用的模型转换为现有工具所基于的模型,但存在以下两个方面的问题:

1)将本文的模型转换为现有工具所基于的模型会导致转换的模型比较大,例如,一个状态  $0 < x < 3$  就变成  $x=1, x=2$  两个状态;状态  $x > 3$  就变成  $x=4, x=5, \dots$  等多个状态,这样会导致转换的模型状态非常多,使得模型检测的复杂度增加。

2)因为本文的逻辑公式 CTML 与现有工具所基于的模型对应的时序逻辑公式(如 PTCTL)不同,所以也要将模型对应的时序逻辑公式进行转换,同理会导致逻辑公式长度变大,使得模型检测的复杂度增加。

结束语 为了对带有数据约束的概率实时系统进行研究,提出了一种基于连续时间的 CT-PZIA 规范,其既可以描述系统的行为属性,也可以刻画系统数据方面的属性;并给出了基于连续时间 CTML 的模型检测算法,从而实现了带数据约束的概率实时系统的模型检测。本文研究了带数据约束的概率实时系统的原理和方法,下一步的研究工作考虑将这些原理和方法运用到具体的实际概率实时系统中,这是未来研究工作的方向和重点。

## 参考文献

- [1] 李广元,唐稚松.带有时钟变量的线性时序逻辑与概率实时系统验证[J].软件学报,2002,13(1):33-41.
- [2] 戎玫,张广泉.模型检测新技术研究[J].计算机科学,2003,30(5):102-104.
- [3] BÉRARD B, BIDOIT M, FINKEL A, et al. Systems and software verification: model-checking techniques and tools [M]. Springer Publishing Company, Incorporated, 2010.
- [4] 王晶,戎玫,张广泉,等.基于概率模型检测的 Web 服务组合验

结束语 本文运用 DBSCAN 算法对电子邮件网络进行社团发现。在介绍 DBSCAN 算法的基础上,详细设计了算法在邮件网络社团发现中的具体实现流程,最后通过对安然邮件语料集的实验证明了算法的有效性以及在社团发现方面的优越性。

## 参考文献

- [1] 杨博,刘大友,金第,等.复杂网络聚类方法[J].软件学报,2009,20(1):54-66.
- [2] 罗浪,张绍武,陈韬.基于稠密子团和边聚类系数的局部社团挖掘算法[J].电子设计工程,2013,21(18):36-40.
- [3] NEWMAN M E J. Detecting community structure in networks [J]. European Physical Journal, 2004, 38(2): 321-330.
- [4] 林旺群,卢风顺,丁兆云,等.基于带权图的层次化社团并行计算方法[J].软件学报,2012,23(6):1517-1530.
- [5] 马菲,徐汀荣.基于种子边的重叠社团发现算法[J].计算机应用研究,2015,32(9):2563-2596.
- [6] 证[J]. 计算机科学,2012,39(1):120-123.
- [5] CAO Z, WANG H. Extending interface automata with  $\lambda$  notation [M] // Fundamentals of Software Engineering, Springer Berlin Heidelberg, 2012: 359-367.
- [6] CAO Z, WANG H. Hybrid ZIA and its Approximated Refinement Relation [C] // Proceedings of the 6th International Conference on Evaluation of Novel Approaches to Software Engineering, Beijing, China, 2011: 260-265
- [7] 狄杨思.形式规范的自动验证算法的研究[D].南京:南京航空航天大学,2012.
- [8] 狄杨思,曹子宁,王辉.一种基于形式规范 ZIA 的自动验证方法[J].光电与控制,2011,18(增刊):113-120.
- [9] CLARKE E M, GRUMBERG O, PELED D A. Model checking [M]. MIT Press, 1999.
- [10] 周从华,刘志锋,王昌达.概率计算树逻辑的限界模型检测[J].软件学报,2012,23(7):1656-1668.
- [11] MINER A S, JING Y. A formal language toward the unification of model checking and performance evaluation [M] // Analytical and Stochastic Modeling Techniques and Applications, Springer Berlin Heidelberg, 2010: 130-144.
- [12] KOYMANS R. Specifying message passing and time-critical systems with temporal logic [M]. Springer, 1992.
- [13] EMERSON E A, MOR A K, SISTLA A P, et al. Quantitative temporal reasoning [M] // Computer-Aided Verification, Springer Berlin Heidelberg, 1991: 136-145.
- [14] ALUR R, DILL D L. A theory of timed automata [J]. Theoretical computer science, 1994, 126(2): 183-235.
- [15] ALUR R. Timed automata [M] // Computer Aided Verification, Springer Berlin Heidelberg, 1999: 8-22.
- [16] ALUR R. Techniques for automatic verification of real-time systems [D]. Stanford University, 1991.
- [17] ALUR R, COURCOUBETIS C, DILL D. Model-checking for real-time systems [C] // Fifth Annual IEEE Symposium on Logic in Computer Science, 1990 (LICS'90). IEEE, 1990: 414-425.
- [18] 钱俊彦,赵岭忠,古天龙.一种基于时间自动机的时钟等价性优化方法[J].计算机工程,2005,31(18):71-73.