第21卷第10期
2003 年 10月

情 报 科 学

Vol. 21, No. 10
October, 2003

网络计量学的一种 Internet 分布式聚类分析软件

齐艳红

(中山大学生命科学学院，广州 510275)

摘 要 聚类分析是人类认识世界和归纳知识最基本的技术之一。对网络计量学也不例外，如对网站及文献资源进行分类分析等。模糊聚类分析通过计算模糊等价关系矩阵，保证了样品间相似关系的自反性、对称性，以及传递性，是一种反映客观世界真实相似关系的聚类技术。在本研究中，研制了可选择多种相似性测度的模糊聚类分析的 Internet 分布式计算软件 FuzzyCluster©。该软件由 6 个 Java 类和一个 HTML 文件组成，具有友好的图型用户界面，可在多种网络浏览器上运行，可用于网络计量分析研究。

关键词 网络计量学 聚类分析 Internet 软件

中图分类号 G202 **文献标识码** A **文章编号** 1007-7634 (2003) 10-0-0

A Web Based Computation Software of Fuzzy Clustering Analysis for Webmetrics

Qi Yanhong

(School of Life Science, Zhongshan University, Guangzhou 510275)

Abstract In this paper, an Internet based computation software, FuzzyCluster©, in which the algorithm of fuzzy clustering analysis is included, is dveloped and introduced. In this software, Manhattan distance, Euclidean distance, cosine coefficient, and Jaccard distance can be chosen as the similarity measure, and weighting of indices can also be made. FuzzyCluster© is made of 6 Java classes and 1 HTML file, which can be run on various operational systems, platforms and web browsers. This software can be used to the studies in webmetrics.

Keywords Webmetrics Fuzzy Clustering analysis Internet Software

网络计量学(Webmetrics)，是基于互联网的一种信息计量分析方法和工具。网络计量学是近年来随着互联网和网络信息资源的急速发展而出现的一门新型学科。因此，对网络计量学的定义和研究内容，目前尚有不同的表述。例如，网络计量学主要对网络文献信息进行统计分析，或认为它是基于Web 的统计分析工具，或网络计量学采用统计学和信息学方法，利用计算机技术和网络技术对网络信息资源进行统计分析。不管其定义如何，统计学方法和网络技术及其结合是网络计量学的核心内容。聚类分析是人类认识世界和归纳知识最基本的技术之一，其思想从远古时代就已形成。对网络计量学也不例外，如对网站及文献资源进行分类分析等等。

聚类分析是一类多元统计分析技术。常用的许多聚类方法尽管满足样品间相似关系的自反性和对称性，但不能保证传递性。即样品A 和B 相似，样品B 和C 相似，但样品A 和C 不相似。显然，这并不符合客观世界的真实状况。模糊聚类分析通过计算模糊等价关系矩阵，保证了样品间相似关系的自反性、对称性，以及传递性。是一种反映客观世界真实相似关系的聚类技术。为此，本文研制了这种聚类方法的 Internet 计算软件 FuzzyCluster©，为网络计量学提供一种在线计算工具。Internet 上运行的计算软件具有分布式计算、平台无关、可随时更新等优点，已在有关领域得到了成功地应用。随着图书馆数字化和网络化的快速发展，相信这类软件的开发与应用将会更加普遍。

1 算法概要

为了能正确地选择算法过程和使用计算软件，需要对本计算软件采用的方法作一概略介绍。

假设有 n 个方案 (如网站，文献等等)，m 个指标 (如信息时效性，入选链接数等等)，已知 $n\times m$ 取值矩阵 (x_{ij})。构造规范化取值矩阵 (y_{ij})，即取

$$y_{ij}= (x_{ij}- x_j^{min}) / (x_j^{max}- x_j^{min})$$

其中 $x_j^{max}= \max_i x_{ij}$，$x_j^{min}= \min_i x_{ij}$，$i= 1, 2, \cdots\cdots, n$；$j= 1, 2, \cdots\cdots, m$。

经过规范化后，就有 $0\le y_{ij}\le 1$，$i= 1, 2, \cdots\cdots, n$；$j= 1, 2, \cdots\cdots, m$。

若需对指标给定权重，各指标的权重为$W_1, W_2, \cdots\cdots, W_m$，且有

$$\sum_{i=1}^{m} CW_i= 1$$

收稿日期：2002-11-05

则加权的规范化取值矩阵为 $y_{ij}= W_j \times y_{ij}$，同样有 $0 \le y_{ij} \le 1$，i= 1，2，…，n；j= 1，2，…，m。

样品间的相似性基于4种距离或系数，即*Euclidean*距离、*Manhattan*距离、夹角余弦，及*Jaccard*距离：

$$Z_{ij}= \sqrt{\left(\sum_{k=1}^{m} (y_{ik}- y_{jk})^2/m\right)} \quad i, j= 1, 2, \cdots\cdots, n$$

$$Z_{ij}= \sum_{k=1}^{m} |y_{ik}- y_{jk}|/m \quad i, j= 1, 2, \cdots\cdots, n$$

$$Z_{ij}= \sum_{k=1}^{m} (y_{ik} \times y_{jk}) \Big/ \sqrt{\left(\sum_{k=1}^{m} y_{ik}^2 \times \sum_{k=1}^{m} y_{jk}^2\right)} \quad i, j= 1, 2, \cdots\cdots, n$$

$$Z_{ij}= (b_i+ b_j) / (c_i+ c_j- e) \quad i, j= 1, 2, \cdots\cdots, n$$

其中，b_i 为在样品i中出现而不在样品j中出现的非零值的个数，b_j 为在样品j中出现而不在样品i中出现的非零值的个数，c_i 和 c_j 分别为在样品i中出现和在样品j中出现的非零值的个数，e同时在样品i中和样品j中出现的非零值的个数。

显然，对规范化取值后的*Manhattan*距离、*Euclidean*距离，及*Jaccard*距离，均有 $0 \le z_{ij} \le 1$。

对*Manhattan*距离、*Euclidean*距离，及*Jaccard*距离，令相似性系数 $r_{ij}= 1- z_{ij}$，而对夹角余弦，令 $r_{ij}= z_{ij}$。对矩阵 (r_{ij})，进行模糊关系运算，最后可得模糊等价关系矩阵 $R= (r_{ij})$。给定临界值p，若 $r_{ij}< p$，则取 $r_{ij}= 0$，否则，$r_{ij}= 1$。将 $r_{ij}= 1$ 的样品划为同一类，其它样品则各自成一类。

2 软件结构与功能

FuzzyCluster©以Java程序设计工具包JDK 1.1.8开发，由6个类和一个HTML文件组成(http://www.brow-so.com/sites/reference/mathsoft/pattern_recogni-tion/，浏览器设置方法见http://www.browso.com/sites/reference/mathsoft/pattern_recognition/demo.htm)。

①FuzzyCluste类。该类执行聚类方法的计算，并调用其它类协同完成有关任务。其Applet被载入浏览器后，显示输入窗口。内容包括：选择何种相似性测度，选择是否给定指标权重、指标数、样品数，打开取值矩阵文件（图1)。

②ResultShow类（图2)、GraphicsFrame类、Hint类，以及WarningShow类见文献所述。

③ClusterGraphics类。该类用来输出样品的聚类树状图。通过ClusterGraphics类窗口上的“Rotate”按钮，可对聚类树状图进行四个方向的旋转（图3)。

④FuzzyCluster.html文件。该文件将FuzzyCluster类载入Web浏览器，并传入窗口大小及组件大小参数。

在取值矩阵文件中，第一行为各指标的编号，若选择给定指标权重，则第一行为各指标的权重值。以后每行第一个值为样品编号，该行中其余值为该样品下各指标的取值。

取值矩阵文件为普通的MS-DOS文本文件(.txt)。可在MS-DOS中的文本编辑器中编辑，或在Windows中选开始→程序→附件→记事本，在记事本中编辑文件。

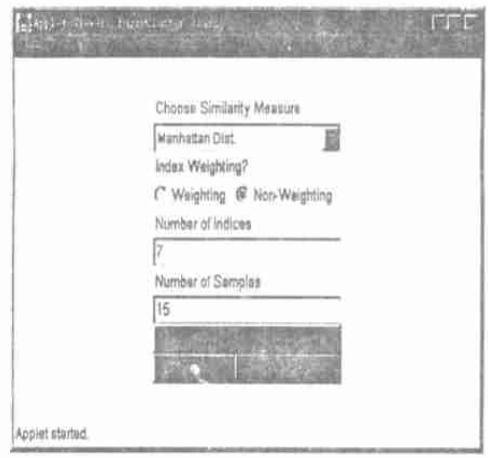

图1 主窗口（FuzzyCluster类）

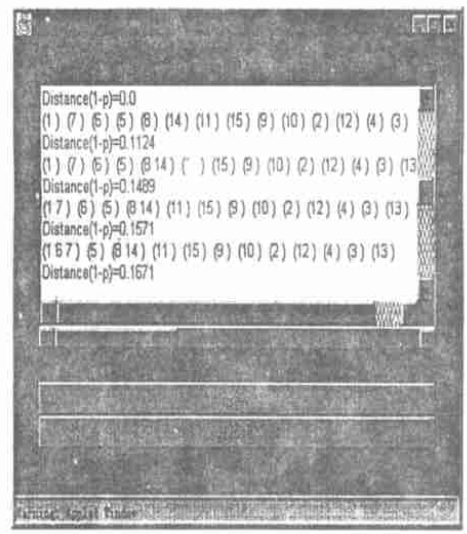

图2 结果输出窗口（ResultShow类）

软件运行后输出规范化取值矩阵，样品的模糊聚类结果等。

FuzzyCluster©具有若干特点，包括：Internet分布式计算，平台无关性，可即时更新，稳定性和安全性好等。

3 计算及分析

网络计量学中的一个问题，是对网站资源进行分类分析。对网站进行分类分析时，可选定这些指标：(1) 网站设计与美观程度；(2) 主办机构权威性；(3) 信息时效性；(4) 信息来源的权威性；(5) 易用性；(6) 链接有效性；(7) 入选链接数。设需对15个网站进行分类，若选择指标不加权，则取值矩阵如表1所示。

其中，第一行为各指标的编号。以后每行第一个值为网站编号。

选择Manhattan 距离，指标不加权，取值矩阵如上所示。FuzzyCluster©的聚类结果见图 3 所示。

3.1 指标加权的效应

若对各指标取定权重分别为 0.2195，0.1707，0.0731，0.2195，0.0731，0.1219，0.2195（或 5，7，3，9，3，5，9，程序会自动将权重进行归一处理）。在上列取值矩阵中将第一行的指标编号代之以各指标权重值。以此为取值矩阵，并选择Manhattan 距离，指标加权，则各网站的聚类结果见图 4 所示。

表 1

	1	2	3	4	5	6	7
1	7	5	2	13	5	9	2
2	18	3	6	17	6	2	1
3	9	13	6	11	5	14	5
4	19	8	5	4	15	5	7
5	4	5	4	7	4	6	1
6	6	7	1	6	1	15	5
7	11	5	1	8	2	10	3
8	7	4	9	9	1	3	2
9	15	1	5	6	6	1	8
10	1	10	7	12	14	5	9
11	2	1	6	3	2	3	7
12	5	18	6	6	2	1	9
13	7	9	10	3	9	14	15
14	5	4	9	8	1	4	10
15	2	6	3	13	6	2	9

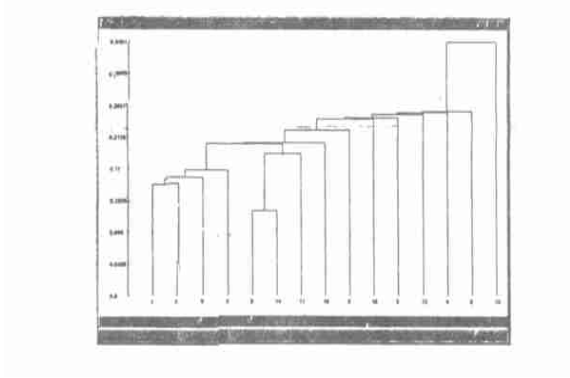

图 3 聚类树状图输出窗口（ClusterGraphics 类）

显然，指标加权与否的聚类结果是不相同的。指标加权重，在权重值大的指标上取值较接近的样品更易划归为同一类。从而，这种选择为在实际应用中根据指标的重要性取定不同的权重，并进行聚类分析提供了便利。

3.2 选择不同相似性测度的效应

选择不同的相似性测度，其聚类结果也将不同。例如，选夹角余弦，指标不加权，取值矩阵如前，则 FuzzyCluster©的聚类结果如图 5 所示。显而易见，图 5 与图 3 中Manhattan 距离的聚类结果有显著差异。

从原理上看，夹角余弦只考虑两样品间的相似性，而不考虑各指标的差值大小。Jaccard 距离考虑两样品间取非零

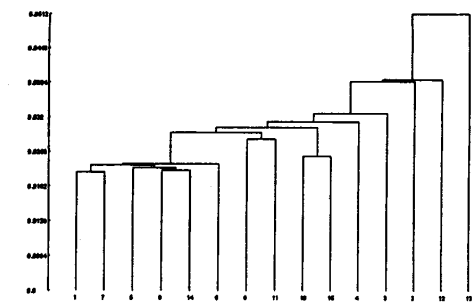

图 4 选Manhattan 距离时的聚类树状图（指标加权）

值的重叠程度，适于稀疏度中等的取值矩阵，即取值矩阵中的非零值既不多也不少，它更适合于定性取值矩阵（取值 0 或 1）。Manhattan 距离和 Euclidean 距离则考虑了两样品间各指标的差值大小。在实际应用时，可根据该原理和经验选用合适的聚类结果。

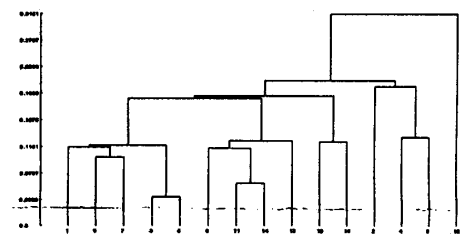

图 5 选夹角余弦时的聚类树状图

参考文献

1 夏 旭，李健康，葛 驰 网络计量学研究：现状、问题与发展 图书馆论坛，2001，21（6）：44～47

2 吴华香 网络计量学研究：互联网上的文献计量学 图书馆杂志，2001，20（1）：33～36

3 金 岩 网络信息计量学方法研究 图书情报工作，2001（2）：29～31

4 齐艳红，王德轩，张文军 应用计算机分析小麦品种抗赤霉病指标及批量鉴定的初步研究 植物病理学报，1990，20（2）：147～152

5 冯德益，等 模糊数学方法与应用 地震出版社，1988.175～179

6 齐艳红，张文军 有害生物侵扰在多样化生境中的一种随机扩散过程及网络计算软件 现代计算机，2002（133）：16～19

7 张文军，齐艳红 生物多样性和均匀度显著性的统计检验及网络计算软件 现代计算机，2002（145）：6～9

8 张文军，齐艳红，Schoenly KG. 泛函连接网络计算软件及其在生物多样性研究中的应用 生物多样性，2002，10（3）：345～355

9 程焕文 趋势种种—图书馆数字化网络化研究札记 图书馆论坛，2001，21（2）：17

（下转第 1079 页）

运行，同时在反馈、循环中实现功能优化和效益最大化目标。

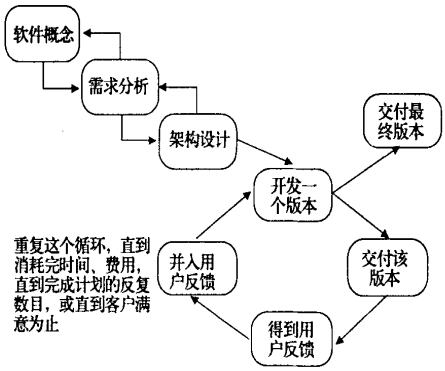

图 3　渐进交付模型

在系统投入使用中，配置管理是一个重要方面。对于配置管理，指定了专门的配置管理人员负责配置管理的工作。通过建立配置管理系统，确保在软件开发的整个生命周期自始自终地建立和维护软件产品的完整性。配置管理项主要包括三类信息：

程序（包括源代码和可执行代码）

文档（包括开发流程文档和用户手册等使用文档）

数据（程序内数据、例如配置文件；程序外数据、例如外部系统输入的数据）

配置管理的主要工作包括：配置管理计划，识别配置项，控制配置项的变化，配置审计，提供配置状态报告。

我们使用 CVS 工具来支持配置管理工作，因为 CVS 是自由软件，不需要付费，并且其功能能满足绝大部分软件企业的需求。CVS 是现在使用最为广泛的配置管理工具，可在 Linux、Unix 及 Windows 2000 等平台上运行。

信息软件开发应有严格的质量保证，质量保证由 SQA（软件质量保证工程师）负责，SQA 要制定质量保证计划，定期审核软件过程和软件产品是否符合规范、流程和相关标准。SQA 在整个软件项目生命周期中检讨项目工作并审核项目产品，并且为管理层就项目是否遵循所建立起的计划，标准和步骤提供可视性。对使用和运行状况，SQA 要进行跟踪。

在大学图书馆面向用户的数据库开发和推进服务中，我们部署、试行了 CMM 3 管理，不仅沟通了开发者与使用者之间的关系，而且提高了系统的可靠性，有利于在现有技术平台下的项目优化。

参考文献

1　Mark C. Paulk, Bill Curtis, Mary Beth Chrissis, and Charles V. Weber, "Capability Maturity Model, Version 1.1," IEEE Software, Vol. 10, No. 4, July 1993, pp. 18-27

2　Project Management Institute Standard Committee, A Guide to the Project Management Body of Knowledge, PMI, 2000

3　Pankaj Jalote, CMM in Practice: Processes for Software Development at Infosystems, Addison-Wesley, 2000

4　(美) 凯西·施瓦尔贝著. 王金玉, 时郴译. IT 项目管理. 机械工业出版社, 2002

5　胡昌平. 信息服务的社会监督—信息服务的社会化发展与社会监督体系的构建. 情报学报, 2001 (3)

6　胡昌平. 信息管理科学导论. 高等教育出版社, 2001

(责任编辑：刘凤勤)

(上接第 1071 页)

10　沈向若　网络信息安全防患及防范对策　图书馆论坛, 2001, 21 (1): 30～32

11　朱淑华　论网络环境下图书馆的信息服务工作　图书馆论坛, 2001, 21 (1): 71～73

12　袁嘉秀　普及型用户界面的理想模式　图书馆论坛, 2001, 21 (5): 23～24

13　孔志辉, 徐守志, 等　图书馆网络建设的问题研究　现代图书情报技术, 2000 (3)

14　吴传炉, 李业龙　数字图书馆信息资源建设探讨　现代图书情报技术, 2000 (5)

15　郭　瑜　Internet 上专业性信息资源指引库的建设　现代图书情报技术, 1997 (2): 20～33

16　Kim P, Eng TR, Deering JM, et al. 1999 Published criteria for evaluating health related web sites: review. BMJ, 318 (7184): 647～649

(责任编辑：徐　波)