基于角色划分的文献软聚类算法 *

马瑞新[a], 邓贵仕[b], 孟繁成[a]
(大连理工大学 a. 软件学院; b. 经济管理学院, 辽宁 大连 116621)

摘 要: 传统的文献聚类算法根据分析文献关键词进行,忽视了文献之间的引用关系,导致了主题漂移和搜索精度不高的问题。针对引文网络中的聚类问题,受到优先情节和增长定律的启发,提出了一种基于角色划分的分层次的文献软聚类算法。首先根据文献之间的引用关系构造引用矩阵,进行结构挖掘;然后根据结构挖掘的结果为每一聚类构造聚类主题,进而进行关键词分析,精化聚类。实验结果表明,该算法能够有效地提高搜索精度和效率。

关键词: 主题漂移; 优先情节; 增长定律; 角色划分; 聚类主题
中图分类号: TP309; TP301.6 **文献标志码**: A **文章编号**: 1001-3695(2012)03-0856-03
doi:10.3969/j.issn.1001-3695.2012.03.015

Soft paper clustering algorithm based on role assorted thoughts

MA Rui-xin[a], DENG Gui-shi[b], MENG Fan-cheng[a]
(*a. School of Software Engineering, b. School of Economics & Management, Dalian University of Technology, Dalian Liaoning* 116621, *China*)

Abstract: Traditional paper clustering algorithm focuses on the keywords analysis while ignores the "refer-to" relationship, which results in the problem of topic drift and low accuracy. This paper inspired by the complex priority and thegrouth theorem, in terms of the clustering in citation network, came up with a hierarchical soft clustering algorithm based on role assorted thoughts. It firstly constructed the "refer-to" matrix in accordance with the reference relationship, mined the structure communities; afterwards, it constructed the clustering theme on the basis of structure discovery, and then analyzed the keywords, refined the clustering. Experimental results show that this algorithm is able to greatly improve the search accuracy and efficiency.

Key words: topic drift; complex priority; thegrouth theorem; role assorted; clustering theme

0 引言

文献聚类作为科技论文管理平台数据整理和组织的重要手段之一,已经逐渐成为解决存储管理组织海量科技文献的有效途径。传统的硬聚类算法与软聚类算法的主要区别在于:在硬聚类算法中,一个个体只能属于某一个聚类;而软聚类算法却允许一个个体同时属于多个聚类[1]。近代的科学研究常常涉及多种学科的交叉和渗透,然而硬聚类算法却限制了对交叉学科文献的有效分析,因此,随着科技的发展和进步,软聚类逐渐取代硬聚类,成为文献聚类算法的主流。

目前,常用的聚类算法有 K 均值、K 中心值和层次聚类等。前两种算法人为设定聚类的数量,因此限制了聚类的精度;层次聚类算法不断地产生嵌套的簇集,在一定程度上增加了算法的复杂度。近年来,文献聚类算法受到了来自物理、数学、计算机等各领域专家的大量关注,然而绝大部分算法都是在现有算法的基础上进行优化和改进,并无新算法提出。本文结合层次聚类的思想,提出了一种无须任何先验信息的文献软聚类算法。

本文算法主要分为两个层次:基础结构挖掘和精细内容分析。第一个层次受到优先情节和增长定律[2]的启发,根据文献的被引用次数,以时间为轴动态模拟引文网络的形成演化机制,同时进行结构聚类;第二个层次的主要目的是提高聚类精度,结合结构聚类的结果,根据聚类中文献的关键词为每一个聚类构造一个聚类主题,对比文献与其所属聚类主题的相关相似性优化聚类结果。

1 基于角色划分的软聚类算法

1.1 算法思想

角色划分的思想来自于方守兴的特殊人物法则[3]。他提出,在网站的推广过程中,有三类人起到关键性作用:内行、联系员和推荐者。本文基于该思想,提出对引文网络中的文献进行角色划分,寻找目标文献的所属聚类、聚类中心及最近邻居,进而提高文献聚类的精度。

虽然大多数真实网络存在巨大差异,但它们都有一个共同之处:增长。从少数几个节点开始,随着节点的不断增长,网络的规模与日俱增,逐渐达到当前的数量。在增长的过程中,节点不断地与其他节点建立链接。优先情节便是指在建立链接的过程中,如果同时面对的两个节点中,前一个节点的链接数

收稿日期: 2011-08-29; **修回日期**: 2011-10-09 **基金项目**: 国家自然科学基金资助项目

作者简介:马瑞新(1975-),男,辽宁大连人,博士,主要研究方向为电子商务、社会挖掘、群智能(teacher_mrx@126.com);邓贵仕(1945-),男,辽宁大连人,教授,博士,主要研究方向为复杂系统分析、电子商务、经济管理;孟繁成(1989-),男,辽宁大连人,本科生,主要研究方向为群智能、社区挖掘.

量是后一个的两倍,那么选择前一个节点的概率将是后一个的两倍[4]。增长定律的实质在于,早出现的节点要比晚出现的节点有更多的机会积累链接。本文受到该启发,提出在引文网络中,被引用次数多的节点出现的时间要远远早于被引用次数少的节点。

传统的文献聚类算法往往忽略文献网络的增长特性,选择固定周期进行聚类和重聚类,而不对新加入引文网络中的文献进行及时聚类,导致新出现的文献不能及时地出现在用户的视野中,降低了文献的搜索精度。本文根据真实引文网络中文献之间的被引用次数,以时间为轴,模拟引文网络的形成及演化机制,同时进行文献聚类,有利于提高文献搜索精度。

1.2　算法详细步骤

1.2.1　引文网络结构聚类

根据文献之间的引用和被引用关系,构建一个有向引用矩阵。若资源 R_A 引用了资源 R_B,则表示为 $R_A \rightarrow R_B$。

首先,所有的文献按照自身被引用次数的降序排列,组成一张列表 L_{refer},聚类中心集初始化为空。其次,对列表中的文献从上到下进行检查,若列表中某个文献与已发现的聚类关联度小于最小关联度阈值 γ,那么这个文献成为新的聚类中心,加入到聚类中心集中;若某一文献与多个已发现的聚类关联度大于 γ,则标志为联系者,并将其加入联系者集合;若某一文献仅仅与一个已发现聚类的关联度大于 γ,则将其加入该聚类。因为对被引用次数高的文献优先进行检查,因此每个聚类种子都将是聚类的中心粒子。这些聚类中心分别引导本聚类的文献随自己落脚在多模态最优点[5]。

使用文献与聚类间的关联度来判定文献 i 的所属聚类。$\{C\}$ 表示聚类的集合,文献 i 与聚类 C 的关联度使用式(1)来计算:

$$R_C^i = \frac{\sum_{j \in C} [E_{i,j} \times d_j]}{\sum_{k \in \{C\}} [E_{i,k} \times d_k]} \tag{1}$$

其中:$E_{i,j}$ 为双极性阈值函数[6],当文献 i 和 j 存在关联时,$E_{i,j}$ 为 1,否则为 0;d_j 为文献 j 的被引用次数,d_j 越大,说明文献的位置越核心。与核心位置的文献关联权重越大,就与该核心文献所在的聚类关联度越密切。由式(1)可知,在为聚类外的游离文献寻找归属聚类的时候,不仅要考虑聚类种子对游离文献的吸引力,而且还要考虑所有处于聚类内部的文献对游离文献的综合吸引力。单个文献可能属于多个聚类,对于某两个聚类 A 和 B,假若 $R_A^i > \gamma \&\& R_B^i > \gamma$,则认为文献 i 同时属于聚类 A 和 B。

1.2.2　基于文献内容的聚类细化

文献之间的引用—被引用关系仅仅是文献属性的一方面,该特性是动态变化的、不稳定的。基于结构的聚类并不能准确地表示文献的内在特征,因此,本文在结构聚类的基础上,基于文献的关键词进一步对文献内容进行考察,并提出聚类主题的概念。本文在空间矢量模型理论[7]的基础上进行聚类分析和计算。

精细聚类的步骤如下所示:

a)根据聚类中文献的关键词向量,为每一篇文献构造一个特征向量。

b)根据每个聚类内部中文献特征向量,提取出现频率最高的 D 个关键词作为该聚类的特征主题。聚类主题用向量表示,对于社区 I,$\text{theme}_I = \langle \text{keyword}_1, \text{keyword}_2, \cdots, \text{keyword}_D \rangle$,$X_I = (X_{I,1}, X_{I,2}, \cdots, X_{I_i,D})$ 代表聚类主题中关键词出现的次数。聚类主题的特征向量长度为 D,并且 $\text{theme}_{SN1} \cdot \text{theme}_{SN2} = 0$。

c)计算文献特征向量与聚类主题之间的符合程度,将文献特征不符合聚类主题的文献驱赶出原聚类,为其寻找符合自身特征的聚类。

聚类主题是整个聚类中文献集中特征的反映,因此,它能明确地反映聚类的内容或者功能。在聚类网络中,用 D 表示文献特征的空间维度,则在 t 时刻,聚类中文献的运动速度可以表示为 $V_i^t = (v_{i1}, v_{i2}, \cdots, v_{iD})$,它代表在周期 t 内聚类特征的变化程度;$X_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,D})$,记录文献 i 的特征向量,$x_{i,D}$ 表示关键词 keyword_D 在聚类 i 中出现的次数。本文使用凝聚系数作为衡量聚类结构是否稳定的标准,d_M 表示聚类 M 中所有文献的引用情况,d_M^{in} 表示聚类内粒子的对内引用次数,d_M^{out} 表示聚类内文献的对外引用次数,则聚类 M 的凝聚系数[8]为

$$C_M = \frac{d_M^{\text{in}}}{d_M} \tag{2}$$

聚类的凝聚系数越大,表示聚类内文献的交流越密切,互动程度越高,文献之间的依赖关系越强烈,聚类对文献的吸引力越强,即聚类的内聚性越高,耦合性更低。

适应度函数用来评判文献与聚类主题的符合程度。若文献的特征向量与聚类主题不相符,说明虽然该文献在链接结构上属于该聚类,但是文献的行为特征、兴趣模型却不是该聚类主题能够表达的。

对于聚类 I 和位于其中的文献 i,定义 i 的适应度函数为

$$F(X_i^t) = \frac{\sum_{j=1}^{D} \text{keyword}_j \times X_{I,j} \times x_{i,j}}{\sqrt{\sum_{j=1}^{D} X_{I,j}^2}} \tag{3}$$

其中:keyword_j 为双极性阈值,若关键词 keyword_j 同时存在于文献 i 的特征向量和 i 所在的聚类主题中,则 $\text{keyword}_j = 1$,否则为 0。根据适应度函数计算文献与所属聚类的适应度值,若文献的适应度值为 0,说明文献的属性与聚类主题并不相符,应该重新对文献进行聚类划分。

对于新加入的文献,主要根据文献关键词进行聚类规划,同时,对于每一个聚类 I,其特征主题随着新文献 i 的加入会不断变化。式(4)所示的聚类主题未经过提取,提取之后,聚类主题的长度降为 D。

$$X_I^{T+1} = X_I^T + R_i^{T+1} \tag{4}$$

1.3　算法复杂度对比分析

本文采用快速排序法对文献的引用次数进行排序分析,复杂度为 $O(N \log N)$。相似度对比的最坏情况为每个节点单独成一个聚类,对于大小为 n 的聚类,该情况下的时间复杂度为 $O(n^2)$。对于网络规模为 N 的网络,假设 $N = m \times n$。m 为聚类数量,n 为聚类规模,则其时间复杂度为 $O(mn^2 + N \log N)$,即 $O(N(n + \log N))$。

表 1 展示了本文算法与 K-means 聚类算法和层次聚类算法的时间复杂度的对比结果。从表 1 中可以看出,本文提出的文献软聚类算法的时间复杂度要远远小于已有的其他算法。非增长类型算法假定数据是一次性提供的,因此,在算法运行过程中,该类算法不会自动检测是否有新的文献加入引文网络。本文提出的文献软聚类算法时刻监测新文献的增长状况,对于新加入的文献,根据其引用情况和关键词进行文献分类。

2 实验结果分析对比

为了验证本文软聚类算法的有效性,从中国知网上下载了数学类、物理类、政治类、生物类和计算机类各 100 篇文章作为测试数据集,对文献的引用关系和摘要进行了聚类分析,采用常用的性能评价方法:召回率、查准率[9],并且与 S2FCM[10] 和改进的 S2FCM 算法[11] 进行了实验结果对比。表 2 为算法召回率对比,表 3 为算法查准率对比。

表 1 算法时间复杂度对比

算法	类型	时间复杂度	备注
划分算法	K-means	$O(Nkt)$	迭代,非增长
	PAM	$O(kt(N-k)^2)$	非增长
层次聚类	单连接	$O(kN^2)$	非增长
	平均连接	$O(kN^2)$	非增长
	全连接	$O(kN^2)$	非增长
本文算法		$O(N(n+\log N))$	增长类型

对比表 2 和 3 可以发现,本文提出的算法在召回率和准确率上对比已存在的算法 S2FCM 和改进的 S2FCM 都有不同程度上的提高,值得注意的是,由于计算机属于服务类交叉学科,即计算机类的文章与很多数学、物理、生物上的文章相关,因此,在查询过程中,准确率受到了一定的限制;而政治类的文献与数学、物理等方面均没有交叉性,因此,形成了高内聚、低耦合的聚类,查询准确率就比较高。

表 2 算法召回率对比 %

算法	数学类	物理类	政治类	生物类	计算机类
S2FCM	79.4	81.3	83.5	76.2	72.5
改进 S2FCM	84.6	84.2	88.7	82.4	79.6
本文算法	91.2	92.1	95.3	91.3	87.3

表 3 算法查准率对比 %

算法	数学类	物理类	政治类	生物类	计算机类
S2FCM	82.5	83.4	88.6	81.4	76.8
改进 S2FCM	86.4	87.6	90.7	84.8	86.8
本文算法	93.7	94.5	96.8	91.5	92.6

3 结束语

本文提出了一种新颖的文献聚类算法,通过分析文献之间的引用关系进行角色划分,可以有效地查询到同时属于多种聚类的交叉文献;通过对文献关键词进行分析和研究,构造聚类主题,有效地抑制了“主题漂移”现象。实验证明,本文算法提高了对文献的搜索精度和搜索效率,具有较高的实用价值。该算法已在 SNS 科技论文管理平台[12] 中使用,并取得了不错的实验效果。

参考文献:

[1] 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.

[2] BIANCONI G, BARABSI A. Competition and multiscaling in evolving networks[J]. Europhysics Letters,2001,5(4):436-442.

[3] 王娴,谢弛,荣雪,等. SNS 走向何方——SNS 网站运营的现状和未来趋势研究[EB/OL]. 2008-12-03. 人民网.

[4] 艾伯特. 巴拉巴西. 链接网络新科学[M]. 徐彬,译. 长沙:湖南科学技术出版社,2007.

[5] LI Xiao-dong. Adaptively choosing neighborhood bests using species in a particle swarm optimizer for multimodal function optimization[C]// Proc of Genetic and Evolutionary Computation Conference. 2004:105-116.

[6] FAN Cong-xian, XU Ting-rong. Research and improved algorithm of HITS based on Web structure mining[C]//Proc of Computer Information. 2010:160-162.

[7] 高琪,张永平. PageRank 算法中主题漂移的研究[J]. 网络与通信,2010,3(3):117-119.

[8] 胡健,董跃华, 杨炳儒. 大型网络中社区结构发现算法[J]. 计算机工程,2008,34(19): 92-93.

[9] 范聪贤,徐汀荣,范强贤. Web 结构挖掘中 HITS 算法改进的研究[J]. 微计算机信息,2010,26(1-3):160-162.

[10] 裴继红, 范九伦, 谢维信. 一种新的高效软聚类方法:截集模糊 C-均值(S2FCM)聚类算法[J]. 电子学报,1998,26(2):83-86.

[11] 白似雪,陆萍. 一种基于文本分类的特征选择方法[J]. 南昌大学学报:工科版,2008,30(1): 87-90.

[12] http://www.linkscholar.com/[EB/OL].

(上接第 855 页)和均衡运行负载。但本文的交互协商策略只考虑了多 agent 系统中单一任务的委托,在将来的研究中,笔者希望进一步研究多任务的委托执行,也试图将这种策略应用在其他工程领域中。

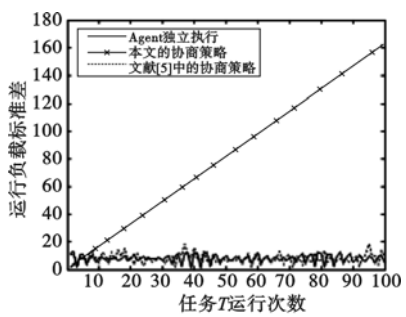

图6 运行负载标准差与运行次数关系

参考文献:

[1] RAJI F, LADANI B T. Anonymity and security for autonomous mobile agents[J]. IET Information Security,2010,4(4):397-410.

[2] FELBER P, NARASIMHAN P. Experiences, strategies, and challenges in building fault-tolerant CORBA systems[J]. IEEE Trans on Computers,2004,53(5):497-511.

[3] COLOMBO A W, SCHOOP R, NEUBERT R. An agent-based intelligent control platform for industrial holonic manufacturing systems [J]. IEEE Trans on Industrial Electronics, 2005, 53 (1):322-337.

[4] 吴菊华, 吴丽花, 甘仞初. 基于规范的多 agent 协同机制研究[J]. 计算机应用研究,2009, 26(5):1778-1781.

[5] 何炎祥,陈苹萌. Agent 和多 agent 系统的设计与应用[M]. 武汉:武汉大学出版社,2001.

[6] SANG-HOON K, JIN-SOO K, SEUNGRYOUL M. Modeling and evaluation of serial multicast remote procedure calls (RPCs) [J]. IEEE Communications Letters,2009,13(4):283-285.

[7] SUNGJIN C, HONGSOO K, EUNJOUNG B, *et al*. Reliable asynchronous message delivery for mobile agents [J]. IEEE Internet Computing Magazine,2006,10(6):16-25.

[8] ZHANG Wei, CHENG Gui-xue. A service-oriented distributed framework-WCF[C]//Proc of International Conference on Web Information Systems and Mining. 2009:302-305.