社交网络中的Web数据挖掘技术

谢宗彦[1] 黎 嬡[2] 周纯洁[1]
(北京联合大学北京市信息服务工程重点实验室 北京 100101)[1]
(北京联合大学旅游信息化协同创新中心 北京 100101)[2]

摘 要 社交网络已经成为Web 2.0时代最流行的应用。文中阐述了社交网络和Web数据挖掘的时代背景,并介绍了Web数据挖掘和社交网络的技术及概念。重点讨论了应用于社交网络分析的Web数据挖掘算法,以及集中于文本内容的相关聚类算法。对目前该领域内的常用算法进行分析和比较,并分析算法目前存在的问题及研究情况。

关键词 Web数据挖掘,社交网络分析,聚类算法
中图法分类号 TP311 **文献标识码** A

Survey of Web Data Mining Technology in Social Networks

XIE Zong-yan[1] LI Nao[2] ZHOU Chun-jie[1]
(Beijing Key Laboratory of Information Services Engineering, Beijing Union University, Beijing 100101, China)[1]
(Collaborative Innovation Center of eTourism, Beijing Union University, Beijing 100101, China)[2]

Abstract Social network has become the most popular application of Web 2.0 era. This paper described the social network and Web data mining era background, and introduced the Web data mining and social networking technology and concepts. This paper focused on the Web data mining algorithm applied to social network analysis, focusing on the related clustering algorithm of text content. The existing algorithms in this field were analyzed, compared, and the existing problems and research situation of the algorithm were analyzed.

Keywords Web data mining, Social network analysis, Clustering algorithm

1 引言

近年来,Internet飞速发展并得到广泛应用,从而使得其中包含的Web站点的数目呈指数级别增长,Web已经成为人们获取信息的重要手段。而每个Web站点就是一个数据源,这些数据源信息丰富,涉及的内容广泛,包括政治、经济、新闻、娱乐、消费、财经、教育、宗教等。通过超级链接,这些内容和组织各异的Web站点就构成了一个巨大的异构数据库环境,可以看成广泛意义上的数据库。与传统数据库相比,Web数据源更大、更复杂。由于Web上的数据正以每天一百万个页面的速度增长,页面数目已超过10亿,使得人们正处在数据信息爆炸但知识贫乏的时代。另外,万维网(World Wide Web, WWW)具有动态、开放、异质、分布广泛等特点。如何在Web这样的分布式环境中快速、准确、有效地发现潜在的有价值的信息,并从中提取出知识内容已经成为目前信息检索、数据挖掘和知识管理等研究领域的重要课题。同时,信息过载、安全、真伪等已经成为一个越来越需要迫切解决的问题。

社交网络分析主要是对社交网络中的关系和结构进行分析[1]。具体包括分析结构中节点的密度、集中性以及群体的特征。社交网络主要由网络中的个体日常持续不断地交互形成,因此它包括了网络中人们之间不同种类的关系,比如位置、个体和团体之间的关系等。为了更好地理解社交结构、社交关系以及社交行为,社交网络分析成为一项基本且重要的分析技术。

本文受国家自然科学基金青年项目:基于Agent的景区游客游憩行为仿真建模研究(41101111),北京联合大学学术(科研)创新团队资助项目:旅游大数据研究方法、关键技术与应用研究(Rk10020150)资助。

谢宗彦(1992—),女,硕士生,主要研究方向为文本分析与挖掘;黎 嬡(1975—),女,博士,副教授,主要研究方向为旅游大数据分析与挖掘、游客行为仿真等,E-mail:lytlinao@buu.edu.cn(通信作者);周纯洁(1995—),女,硕士生,主要研究方向为文本分析与挖掘。

社交网络分析的目标是从 Web 的大量用户数据中获取有用的信息和知识，比如用户在网络中发布的内容、社交网络的结构以及用户在网站中的访问行为。Web 挖掘技术是最适合用来进行社交网络分析的信息技术。Web 挖掘就是将数据挖掘技术应用在 Web 环境中，从而得出对互联网企业以及用户有益的知识和模式。从形式上来讲，Web 挖掘主要分为 Web 内容挖掘、Web 结构挖掘、Web 使用挖掘。

2　Web 挖掘技术

Web 挖掘是数据挖掘技术在 Web 中的应用。它从网络中的大量数据以及数据库中提取和发现有用的模式和信息，因此可被定义为从 Web 中发现和提取有用的信息。

针对不同的分析对象和资源，Web 挖掘技术分为 3 种不同的类型：Web 内容挖掘、Web 结构挖掘、Web 使用挖掘。

2.1　Web 内容挖掘

Web 内容挖掘是指挖掘 Web 页面内容、后台交易数据库中的信息，即从 Web 文档内容或其描述中的内容信息中抽取有用知识的过程。Web 上的内容挖掘多为对 Web 上大量文档集合的内容进行聚类、总结、分类、关联分析等，是文本信息的挖掘。除此之外，Web 内容挖掘还包括多媒体挖掘，这种挖掘通常采用关联规则法和特征提取法。

2.2　Web 结构挖掘

Web 结构挖掘是指挖掘 Web 内部组织结构和文档间的链接关系并从中推导知识的过程。对 Web 结构的挖掘，可以用来指导对页面的排序，找到重要的页面，提高检索的性能和网页采集的效率[2]。Web 结构挖掘可分为 Web 内部文档结构挖掘和 Web 文档间超链接结构挖掘两种。

2.3　Web 使用记录挖掘

Web 使用记录挖掘是通过挖掘服务器端上记录的客户访问日志和相关数据来获取有用信息的过程，如获取站点上的浏览者的访问模式等。

3　Web 挖掘技术在社交网络分析中的应用

3.1　3 种 Web 挖掘的类型在社交网络中的应用

(1)Web 内容挖掘。文本挖掘或者自然语言处理在社交网络分析中占据非常重要的地位。Web 内容挖掘可以对社交网站的文档进行分类或分级，特别针对博客、微博或是以文字内容为主的论坛。文章主题的分类通常是很多社交网站很重要的分析应用。

Web 内容挖掘还可以被用来分析用户的阅读兴趣和习惯上通过运用分析工具得出他们最感兴趣的阅读内容，在得到结果后网站可以为用户推送最精准的阅读内容[3]，而大部分的分析通常都需要 3 种类型的 Web 挖掘技术结合起来使用。例如，在上面的例子中若要得到理想的结果，就需要将 Web 内容挖掘和使用挖掘结合起来。

(2)Web 使用挖掘在社交网络挖掘中也扮演了重要的角色。用户在社交网站上的使用交互数据被收集后，对相关的数据进行使用挖掘，其结果可以为社交网站的改进建设提供有价值的建议。使用挖掘还可以用来作为网络中节点的中心度度量的工具，比如：

$$Closeness=(f*(w*b)+f*(w*r))+(f*(w*I)) \quad (1)$$

利用式(1)，我们可以通过微博用户的 3 种行为的权重得出该博主的中心度情况。f 表示博主行为的频率 W 针对每一个博主行为 *closeness* 的权重。微博用户的 3 种行为可以被定义为：b 为访问，r 为阅读，i 为关联。这只是 Web 使用挖掘最简单的应用[4]。

(3)Web 结构挖掘是第三种 Web 挖掘方法，它可以被用来在社交网站、邮件等系统中抽取有用的用户之间的链接结构信息；还可以被用来分析路径长度、可到达性以及找出结构洞等。以上的内容在社交网络分析中是非常基本和传统的。结构挖掘经常将图和可视化的方式作为挖掘的工具来表示社交网络中的数据，这种方式能够使分析者更容易理解和分析问题。

对于大部分的社交网络分析，3 种挖掘方法都是相互结合起来进行工作的，即对于某个特殊的社交网络分析，3 种挖掘类型被共同用来进行分析。

3.2　Web 挖掘技术在社交网络中的应用

本节举例介绍两个最典型的 Web 挖掘技术在社交网络分析中的应用，它们分别是聚类和关联规则。

3.2.1　聚类

在社交网络分析中，在社交网络内部或者跨网络找到属性最相近的群体是一项非常有意义的研究。以往它的实现主要依靠在一个小型的社交网络中应用可视化技术，但这种方法无论是在发现的群体数量上还是在适用的范围上都不能达到理想的状

态。聚类技术可以在一个更大的社交网络环境中确认更多的团体。也就是说，聚类技术可以提供比可视化技术更详细的信息，它包括关系密切的群体，以及群体和社交网络之间的关系[5]。

3.2.2 关联规则

关联规则是一种流行的数据挖掘技术，它可以应用到生活中的很多方面。在社交网络分析中，关联规则可以被用来发现社交网络和跨网络节点之间隐藏的关系。该规则应用于社交网络中，可以得到类似甲认识乙那么甲也认识丙的关系，在这种关系中 *support* 的值为 0.9，*confidence* 的值为 0.5。联系到社交网络中的情况，也可这样理解：关注到甲博客文章的人也关注到乙博客文章的人之间也有一定的关联关系，其中 *support* 的值为 0.9，*confidence* 的值为 0.5。关联规则一次可以提供不同的分析、转化更多的关联数据、确认更多的节点以及网络中的关系。此外，关联规则对社交网络分析中的推荐系统和信息过滤系统的建设也有很大的辅助作用[6]。

4 社交网络分析中的聚类算法

我们可以把聚类方法看作一个独立工具，也可以将它作为一个预处理步骤嵌套到其他算法中。聚类分析方法可以帮助我们解决以下现实问题：哪类顾客群体偏爱何种类别的商品，偏爱同一类型商品的顾客之间又有何共同点（包括收入、年龄、性别等），哪些用户购买时比较看重价格，哪些用户购买时更多地是关心商品的质量[7]。

4.1 SCAN 算法

该算法在传统基于链接稠密度的方法的基础上，同时考虑了结构相似度并且分析了节点功能，然而其仅针对无向网络聚类，未考虑社交网络的有向性。

4.2 结构相似度的有向网络聚类算法(Dir SCAN)

为了发现网络中隐藏的簇结构，传统的网络聚类方法主要基于链接的稠密度(linkdensity)，使得簇内节点距离较近而簇间节点距离较远，如经典的 Newman 快速算法[8]和 Kernighan-Lin 算法[9]。然而，以上算法忽略了社交网络的有向交互性和节点具有不同功能的特点。

DirSCAN 算法[10]仅需遍历有限次节点和边，一次遍历即可获得节点的到达邻居、判断核节点，从而基于核节点进行簇扩展。因此，若网络中存在 n 个节点，则遍历节点的复杂度为 $O(n)$。在遍历边时，需要计算节点的每条出边是否为到达邻居关系，最差情况为所有节点都相连，复杂度为 $O(n(n-1))$。由于实际社交网络通常为稀疏网络，遍历边的次数可近似为遍历节点的次数，因此 DirSCAN 算法的时间复杂度近似为 $O(n)$。相对于经典的 Newman 快速算法和 Kernighan-Lin 算法，该算法运行速度更快，并且同时考虑到了社交网络的有向交互性和节点具有不同功能的特点。

4.3 基于概率的图聚类算法

PGC 的主旨思想是通过比较各节点邻居集的相似度来实现聚类。综合来说，PGC 主要针对现实世界中真实的大型图数据进行聚类，主要分为 3 个阶段，即降维、哈希和验证。其中降维技术采用改进后的 Minhash 算法，哈希部分利用 LSH 实现，验证过程利用贝叶斯推断估计。

鉴于大型图数据的复杂性，PGC 主要采用了概率运算而非精确运算。在社交网络聚类技术中，很多时候并不需要精确的聚类，只需概率保证计算结果即可，因此 PGC 本质上是对聚类的效率和质量进行均衡。

4.4 谱聚类

谱聚类方法近几年受到了学者的广泛关注，它是数据挖掘领域的又一个研究热点[11]。与那些传统聚类算法相比，它的优点是能够发现任意形状的聚类，且最终收敛在全局最优解[12]。谱聚类算法实现简单，它以谱图理论为基础，通过 Laplacian 矩阵将原数据空间进行重构[13]，降低样本数据的维度，这使得数据在子空间上的分布结构更为清晰。

结束语 本文研究了 Web 挖掘的概念和技术在社交网络中的应用，回顾了关于 Web 挖掘和社交网络相关的研究现状；然后介绍了社交网络中用到的聚类算法、算法之间的比较及发展现状。

把 Web 挖掘技术应用于社交网络分析领域，是当今一个热门的研究方向，然而该研究领域里有一些困难需要克服，比如如何从海量的数据中进行抽样。当然在其他 Web 挖掘应用中，数据抽样是一项简单的工作，它的目的是减少分数据的总量。但是，对于社交络，从社交网络庞杂的数据中抽样出能够代表真实社交网络原貌的数据并不是一项简单的工作。进一步，针对某个特殊的社交网络问题，把不同类型的 Web 挖掘类型结合起来进行挖掘也是一种通用的解决方法。本文提出的解决方法和实施步骤将会有助于解决某些社交网络的分析问题，但有些挖掘过程在实际的应用中还需要进一步细化。

参 考 文 献

[1] 高华. Web 挖掘技术在社交网络分析的应用研究[J]. 科技信息，2013(9)：91-92.

[2] NEWMAN M E J. Detecting community structure in networks[J]. European Physical Journal B, 2004, 38(2)：321-330.

[3] 黄钢石，陆建江，张亚非. 基于 NMF 的文本聚类方法[J]. 计算机工程，2004，30(11)：113-114.

[4] 蒋玉婷. Web 数据挖掘及其在微博话题检测中的应用研究[J]. 现代电子技术，2016，458(3)：115-119.

[5] 伍育红. 聚类算法综述[J]. 计算机科学，2015，42(s1)：491-499.

[6] 杨震，王来涛，赖英旭. 基于改进语义距离的网络评论聚类研究[J]. 软件学报，2014(12)：2777-2789.

[7] 彭敏，黄佳佳，朱佳晖，等. 基于频繁项集的海量短文本聚类与主题抽取[J]. 计算机研究与发展，2015，52(9)：1941-1953.

[8] NEWMAN M E. Fast algorithm for detecting community structure in networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics，2004，69(6 Pt 2)：066133.

[9] SHEN H，CHEN X，CAI K，et al. Detect overlapping and hierarchical community structure in networks [J]. Physica A Statistical Mechanics & Its Applications，2009，388(8)：1706-1712.

[10] SHEN H W. Detecting the Overlapping and Hierarchical Community Structure in Networks [M]// Community Structure of Complex Networks. Springer Berlin Heidelberg，2013：19-44.

[11] 宋传超. 社交网络中基于概率的可伸缩聚类算法研究[D]. 济南：山东建筑大学，2013.

[12] 严俊. 谱聚类算法改进及在社交网络中的应用[D]. 桂林：广西师范大学，2014.

[13] 孔万增，孙志海，杨灿，等. 基于本征间隙与正交特征向量的自动谱聚类[J]. 电子学报，2010，38(8)：1880-1885.

(上接第 34 页)

[2] 冯志伟. 自然语言的计算机处理[M]. 上海：上海外语教育出版社，1996：3-4.

[3] HUBEL D H，WIESEL T N. Receptive fields and functional architecture of monkey striate cortex[J]. Journal of Physiology，1968，195(1)：215-243.

[4] FUKUSHIMA K. Neocognitron：A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol[J]. Cybern，1980，36(4)：193-202.

[5] 李彦冬，郝宗波，雷航. 卷积神经网络研究综述[J]. 计算机应用，2016，36(9)：2508-2515.

[6] LECUN Y. Generalization and Network Design Strategies[C]// Connectionism in Perspective. 1989.

[7] BENGIO Y，LECUN Y. Convolutional Networks for Images，Speech，and Time-Series[C]// The Handbook of Brain Theory and Neural Networks. 1995.

[8] LÉCUN Y，BOTTOU L，BENGIO Y，et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE，1998，86(11)：2278-2324.

[9] KRIZHEVSKY A，SUTSKEVER I，HINTON G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc.，2012：1097-1105.

[10] KIM Y. Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv，2014.

[11] KALCHBRENNER N，GREFENSTETTE E，BLUNSOM P. A Convolutional Neural Network for Modelling Sentences[J]. Eprint Arxiv，2014，1.

[12] DONG L，WEI F，ZHOU M，et al. Question Answering over Freebase with Multi-Column Convolutional Neural Networks[C]// Meeting of the Association for Computational Linguistics and the，International Joint Conference on Natural Language Processing. 2015：260-269.

[13] RUMELHART D E，HINTON G E，WILLIAMS R J. Learning rep-resentation by back-propagating errors [J]. Nature，1986，323(3)：533-536.

[14] BENGIO Y，COURVILLE A. Deep Learning of Representations[M]// Handbook on Neural Information Processing. Springer Berlin Heidelberg，2013：1-28.

[15] BENGIO Y. Learning Deep Architectures for AI[M]. Now Publishers，2009.