学术Web主题结构挖掘研究

杨 波 著

南京大学出版社

国家社科基金项目（13CTQ031）

学术Web主题结构挖掘研究

杨 波 著

南京大学出版社

图书在版编目(CIP)数据

学术 Web 主题结构挖掘研究 / 杨波著. —南京：南京大学出版社，2018.11

ISBN 978-7-305-19826-7

Ⅰ. ①学… Ⅱ. ①杨… Ⅲ. ①学术研究—网络信息资源—主题分析 Ⅳ. ①C795②N795

中国版本图书馆 CIP 数据核字(2017)第 330056 号

出版发行 南京大学出版社
社　　址 南京市汉口路 22 号　　邮　编 210093
出 版 人 金鑫荣

书　　名 学术 Web 主题结构挖掘研究
著　　者 杨　波
责任编辑 陈　佳　　编辑热线 025-83686308

照　　排 南京紫藤制版印务中心
印　　刷 江苏凤凰数码印务有限公司
开　　本 787×960　1/16　印张 13.5　字数 214 千
版　　次 2018 年 11 月第 1 版　2018 年 11 月第 1 次印刷
ISBN 978-7-305-19826-7
定　　价 45.00 元

网　　址 http://www.njupco.com
官方微博 http://weibo.com/njupco
官方微信 njupress
销售热线 025-83594756

目 录

图 目 录

表 目 录

第一章 引 言

1.1 研究背景

根据不同的研究目标和技术方案，学术 Web 结构挖掘的数据粒度可以是宏观的国家层面、区域层面或者网站层面，也可以是较为微观的网页层面，甚至是概念或关键词。因此最终获得的 Web 结构可以是比较宽泛的学科关系(discipline/subject/domain/field)，也可以是具体的主题关系(topic/theme)。由于涉及不同层面的研究对象，在数据粒度上也不尽相同，为了叙述方便，本研究将上述两种关系统称为主题关系。

由于频繁的跨区域科研合作和交流活动的需要，网上学术社区逐渐形成并且壮大。对万维网上学术社区的主题结构进行有效地揭示，是发现、评价和利用海量在线和离线学术资源的主要途径之一。社会网络分析理论认为，社会上人和人之间都是由一定的关系网络连接起来的，人和人之间的关系就是这个网络的边(Edge)，人是社会网络上的结点(Node)。整个社会是个巨大的网络，在这个网络中同时存在很多个子网，对子网的划分是通过社会网络中结点之间的距离(关系)和局部区域内结点(人)的疏密程度来进行的。每个子网代表了社会上的各个社区，这些社区具有较强的内聚性和较弱的耦合性，即社区内成员之间具有相似的兴趣，而不同社区的成员的兴趣差异比较大。可以试想，在万维网上是否也同样存在类似的社区？在社会网络中，人和人之间是通过模糊的兴趣爱好联系起来的，而在万维网上，可以把网页视为一个个结点，这些结点之间是通过精确清晰的 URL 联系在一起的。网页和网页之间的联系则体现为网页之间主题的相似性，这就给万

维网上的社区理论提供了最好的理论基础——我们称之为 Web 社区。

Web 社区是由多个成员组成的，社区内成员之间的链接关系明显要强于它们和社区外成员之间的链接关系。每个社区的主题以及链接特征都不同于其他社区，社区之间存在松散的链接关系，社区内部则具有紧密的链接关系。社区内部根据链接关系的紧密程度可进一步划分社区的边界。

利用 Web 信息的社会网络性进行知识发现是对复杂的网络学术社区中的主题关系进行识别、抽取、评价和再组织的新途径。对 Web 主题结构进行分析的本质是对网页(Page)、网站(Website)、域(Domain)等不同层面的网络对象之间的主题关系进行描述、分析和度量。目前对 Web 文档主题结构进行挖掘的研究主要从文本分析、链接网络分析、复合特征分析、宏观网络拓扑结构分析、用户聚类和基于 Web 主题图的信息检索可视化这六个方面展开。

1.1.1 基于文本的主题聚类

和传统的文本检索系统一样，以文本分析为基础的 Web 主题结构挖掘主要依赖的是网页之间在文本表示上表现出来的相似性。这种方法把网页的文本相似度近似等同于网页主题的相似度，从而实现 Web 资源的分类和聚类。早期著名的分类搜索引擎 Yahoo! 和 Vivísimo 等采用的就是这种方法，不同的是 Yahoo! 以人工来进行网站分类，其他分类搜索引擎大多是用自动分类的方式来构建主题目录。

Vivísimo 是由美国卡耐基梅隆大学三位科学家联合开发的分类搜索引擎，它的设计目的就是要解决互联网信息过载问题(Information overload)[1]。它与 Google 最大的不同在于，Vivísimo 不是追求文档库的庞大，而是注重让搜索结果更接近用户的信息需求。Vivísimo 对搜索结果进行实时聚类，并且以层级方式将结果显示给最终用户。和 Vivísimo 类似的分类搜索引擎 SNAKET 是由意大利学者开发的开源系统，可为 Web 搜索、图书、新闻和博客等提供等级聚类服务，它同时综合了常用的 16 个搜索引擎的检索结果，可对结果进行实时自动聚类[2]。由北京大学计算机系网络与分布式系统实验室开发的中文搜索引擎“天网”也利用中文网页自动分类

技术，开发了用于目录导航的中文网页分类目录[3]。虽然上述系统作为独立的产品，大多已经退出了历史舞台，但产业界一直在沿用和优化相关的技术策略。

在 Web 页面特征提取和表示方面，涉及自然语言处理以及人工智能方面的诸多技术。尤其是在中文信息处理方面，中文的语言特点决定了在 Web 页面特征提取和表示中必然遇到不同于西文语言的若干技术上的难点。在中文分词过程中，经常需要在汉语分词的粒度上做出选择。在自动标引环节是采用主题词标引还是采用元词标引？是采用赋词标引还是抽词标引？一般认为，在通用主题的中文信息处理中，宜采用元词标引，而在处理专业 Web 文档时，专指的主题词可能更为合适。在考虑同义词转换或者用专业分类法进行分类时，可能需要考虑使用赋词标引，而在处理混合主题的海量文本信息时一般采用抽词标引。在文本分类和聚类方面，设计和训练一个高性能的分类器(Classifier)是一项很有挑战的研究，尤其是要适应多主题和交叉主题的 Web 文本分类器。

用文本分析的方法进行 Web 主题挖掘具有很多优点，如文本特征提取比较方便，可以在比较小的范围内实现主题分类，可以借助分类知识库等经验知识来帮助文本主题的判断。但文本分析面对的是书写符号，因此在将书写符号转换为主题的过程中也存在很多固有的缺陷，如上述的页面特征提取、自然语言处理、机器翻译，以及垃圾信息过滤等问题。

1.1.2 基于链接分析的主题聚类

链接分析(Link Analysis)类似于文献之间的引文分析(Citation Analysis)，即利用网页、网站、机构、地区和国家之间的链接(Linking)与被链接(Linked)关系，通过定量的方法来评价和发现核心资源。链接分析算法以 PageRank 和 HITS 最为著名，它们之间的最大不同在于 PageRank 计算的是基于数据库中全部 Web 文档的稳定的全局值，而 HITS 是在用户查询的基础上进行动态权威值(Authority)和中心值(Hub)计算的。

利用 HITS 算法来发现 Web 主题社区有两个前提假设：① 由网页设计者所创建的超级链接(Hyperlink)隐含了足够的主题信息，通过这些信息可

以计算出该主题下的权威页面；② 在前提 1 的基础上，必然存在以相同宽泛主题为特征的、由多个相互链接的 Web 页面组成的主题社区。Gibson 和 Kleinberg 等人利用 HITS 算法[4]，针对“哈佛”“英语文学”“滑雪”等主题进行了主题社区挖掘的研究。实验结果表明，HITS 可以有效地发现隐含于 WWW 中的主题社区，并且在一定程度上可以反映其中子主题之间的关系，如“德国文学”主题社区和“欧洲文学”主题社区有相当明显的联系。Chakrabarti 等人将链接的锚点周边文字信息和 URL 中包含的字符信息的匹配度作为权重[5]，在 HITS 算法的基础上提出了 ARC 算法（Automatic Resource Compilation），实验证明该算法生成的资源目录质量与 Yahoo! 和 Infoseek 相当，甚至某些情况下质量要优于手工分类的主题目录。Dean 和 Henzinger 利用改进的 HITS 算法——Companion、共链算法（Co-link）分别做了发现相关网页的实验，并且将实验结果和浏览器 Netscape 提供的“相关网页”服务做了对比。结果显示，即便 Netscape 提供的“相关网页”服务利用了内容分析、使用模式分析和链接分析等多种分析方法，Companion 和共链算法还是明显占优[6]。Toyoda 和 Kitsuregawa 对 Companion 做了进一步改进，形成了新的 Companion—算法。利用该算法，他们成功地分析了日本国内和计算机产业有关的 Web 主题网络[7]。

1.1.3 基于复合网页特征的主题聚类

无论是 PageRank 还是 HITS，都是依靠 Web 页面之间存在的 URL 来构建关系网络，这些链接分析算法的理论假设借鉴于引文分析，即来源文献和被引用文献之间存在主题上的相关性。然而，引文分析也存在非主题相关引用的问题，如假引、误引等现象。网络上一个网站链接另外一个网站的动机更为复杂，很多商业网站知道搜索引擎的网页排名算法后，利用交换链接等手段来提高自己网站的排名，从而增加被用户点击的概率。SEO（搜索引擎优化，Search Engine Optimization）研究的一个主要内容就是如何通过设置合理的链接使企业在搜索引擎排名中获益。对于互联网垃圾制造者来说，买卖链接的方式让他们有了很好的机会来通过其他网站的声誉来操纵搜索引擎排名结果。Alastair G Smith 提出了实质网络影响因子

(Substantive Web Impact Factors)[8],他认为对网站的评价只有以指向真正信息资源的 Web 页面为基础才是有意义的。他把大学网站的链接按照链接目的分成九个大类,从分类来看,真正指向重要信息资源的、主题相关的链接很少。

为了防止主题漂移现象的发生,使某个网域(Domain)内具有相同主题的页面能够被顺利发现,从而提高主题社区分析的可靠性,Web 页面上除了链接以外的信息单元也有重要的参考意义,如页面结构、元数据、锚文本、标题和正文文本、正文字体特征等。页面的文字特征,如字体大小、颜色等虽然和主题没有直接的相关关系,但在页面主题特征的提取和表示方面有很好的辅助作用,页面结构也同样具有很重要的相似性判断的功能。Crescenzi 等人通过对网页内部结构建模的方式实现了对网页的分类。他们以 2002 年韩日世界杯的专题网站为样本,分析了每个网页的 DOM 树后,对每个网页建立了网页模板。如果两个页面是可达的(存在链接关系),则分析它们模板的相似度,如果相似度达到一定程度,可以认为两个页面的内容是相关的[9]。

1.1.4 基于宏观网络拓扑结构的 Web 主题社区发现

利用网络拓扑结构分析的方法严格来说也属于链接分析的研究范畴,但和 HITS 等不同的是,这种方法把 WWW 作为一个网络图(Web graph)来分析,网络中的结点和边分别代表网页和链接。基于网络拓扑结构来构建 Web 主题图的依据是,整个网络图中存在一个或者多个比较稠密的子图,这些子图明显不同于其他部分,每个子图就是一个 Web 主题社区。子图以及它们之间的关联关系组成了整个 Web 主题图。基于 Web 结构挖掘的链接图分析方法具有很多文本主题分析和日志挖掘不具备的优势,主要表现在以下几个方面:

(1) 避开了复杂的文本分析难题。

文本分析需要通过复杂的自然语言处理技术进行主题表示和匹配,而网络链接图中可以通过有无人工标注语义的网络链接来表示页面之间的相关与否。

(2) 解决了跨语言 Web 资源的主题相关性判断问题[10]。

文本分析技术在遇到跨语言 Web 资源的主题判断时，需要进行语料转换，也就是机器翻译，而网络拓扑结构分析具备了网络链接天然的语义相关性特质。

(3) 噪音信息过滤效果良好。

虽然网络链接的作假很容易，但主题图构建过程中所用到的社区发现算法以人工选定的起点出发，经过良好的链接评价算法，对垃圾页面的过滤效果比单纯的文本分析方法更加有效。

(4) 便于进行结构化的 Web 资源的评价、获取和利用。

基于 Web 结构挖掘的主题图构建方法可以实现不同层次的资源聚合，如按照主题进行聚合、按照网页、网站、域等信息单元进行聚合。

目前基于网络拓扑分析的主题图构建研究主要集中在两个方面，以资源评价和获取为目的的资源发现研究和以科学评价为目的的链接行为研究。

1.1.5 基于 Web 访问日志的用户社区发现

为了满足个体用户或者团体用户的个性化信息需求，Web 日志挖掘是解决问题的最好办法之一。Web 个性化是通过充分利用用户浏览行为，根据每个用户的特殊需要而定制网站内容和结构的方法。通过对站点用户访问记录的统计或者模式识别，可以建立用户社区(User Communities)，社区内的用户具有相似的兴趣。例如，当用户在电子商务网站进行一定的浏览操作后，会有一些同类用户感兴趣的商品被推荐给当前用户。通过用户社区的发现，以及对这些具有相似兴趣的用户所共同关注的主题进行跟踪和挖掘，可以发现具有相同或者相似兴趣和主题的用户社区和资源社区。

Almeida 等人以巴西的网上书店和音乐流媒体网站作为分析案例，使用访问日志对两个网站的用户进行了聚类分析，分别发现了不同的用户社区，其中以流媒体网站的用户社区分析效果最为明显。他们共发现了 10 个顶