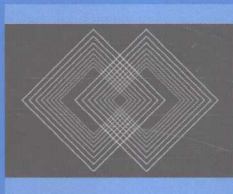


现代信息
科学译丛

XIANDAIXINXI
KEXUEYICONG



链接分析： 信息科学的研究方法

Link Analysis:
An Information
Science Approach

【英】迈克·塞沃尔 (Mike Thelwall) 著

孙建军 李江 张煦 等译

东南大学出版社

SOUTHEAST UNIVERSITY PRESS

现代信息科学译丛

链接分析:信息科学的研究方法

Link Analysis: An Information Science Approach

[英]迈克·塞沃尔(Mike Thelwall)著

孙建军 李 江 张 煦 等译

东南大学出版社
南 京

图书在版编目(CIP)数据

链接分析:信息科学的研究方法/(英)塞沃尔(Thelwall M.)著;
孙建军等译. —南京:东南大学出版社,2008.7

书名原文:Link Analysis:An Information Science Approach

ISBN 978-7-5641-1279-0

I. 链… II. ①塞…②孙… III. 信息学-研究方法 IV. G202-3

中国版本图书馆 CIP 数据核字(2008)第 099653 号

江苏省版权局著作权合同登记

图字:10-2008-195 号

@2004 Elsevier Inc.

This edition of *Link Analysis: An Information Science Approach* by Mike Thelwall is published by arrangement with EMERALD Group Publishing Limited, Howard House, Wagon Lane, Bingley, West Yorkshire, BD16 1WA, United Kingdom

All rights reserved including the rights of reproduction in whole or in part in any form.

链接分析:信息科学的研究方法

原 著:[英] 迈克·塞沃尔(Mike Thelwall)

主 译:孙建军等

出 版 人:江 汉

出版发行:东南大学出版社

社 址:江苏省南京市四牌楼2号(210096)

经 销:江苏省新华书店

印 刷:南京玉河印刷厂

版 次:2009年1月第1版 2009年1月第1次印刷

开 本:787mm×1092mm 1/16

印 张:15

字 数:384千字

书 号:ISBN 978-7-5641-1279-0/TP·116

定 价:38.00元

凡因印装质量问题,可直接向东南大学出版社读者服务部调换。电话:025-83792328

译者序

网络链接分析研究可以追溯到 20 世纪 90 年代中期。1995 年, Brazilian Marcia J. Bossy 首次提出可将信息技术应用于因特网。1996 年, Larson 在《万维网的文献计量:网络空间结构初探》一文中明确将信息技术从文献计量学移植到网络中。早期的链接分析研究同时出现在几个学科中,包括计算机科学领域中的搜索引擎开发,数学领域中的结构和复杂性分析。而在 1997 年, Almind 和 Ingwersen 提出了“网络计量学(webometrics)”一词,旨在定量分析网络现象。此后,链接分析便成了网络计量学的主要研究内容之一。1998 年, Google 的创始人 Brin 和 Page 公开了 PageRank 算法的核心部分,这一链接分析算法,作为 Google 的核心技术之一,支撑着 Google 在商业上取得了巨大的成功,同时,也彰显了链接分析研究的魅力。

在信息科学研究领域,按照 Mike Thelwall 的定义,链接分析就是采用并改进现有的信息技术,借助文档之间的相互关联,对文档自身的特征进行深入分析。链接分析涉及的文档包括四个层面:页面、目录、域名、站点。在理论方面,链接分析与文献计量学中的引文分析有高度相似性。

链接分析的内容主要包括:① Web 结构研究:将文档视为节点,将文档之间的链接视为连线,便可将 Web 理解为一张抽象的结构图(有向图),图中的节点与连线的属性都值得深入研究;② 链接增长规律研究:链接的建立不是随机的(“均匀链接”),而是服从某种规律的,Web 环境中,小世界现象已不再适用,不同类型的页面间的链接增长规律不同,纯粹的幂定律模型(“优先链接”)已难以概括这些规律;③ 链接分类研究:链接可以传达信息,因此,判断创建链接的动机可用于判断学术网络上信息交流的类型;④ 链接分析算法:Pagerank 算法、HITS 算法等链接分析算法应用于搜索引擎检索结果排序,极大地提高了检索效率,网络在发展,这些算法也在不断的更新;⑤ 链接分析工具研究:搜索引擎与网络爬虫一直是链接分析研究中获取数据的主要工具,但二者都有自身的缺陷,如何合理地使用链接分析工具以提高链接分析数据的有效性则一直是链接分析研究中讨论的主要问题。

链接分析研究中常用的工具与方法包括:搜索引擎、网络档案、网络爬虫、链接数据库、数据清理技术、网络空间分析法、虚拟民族志方法、社会网络分析法、网络可视化方法等。

链接分析可应用于多个领域,包括:① 搜索引擎与网站设计:链接分析算法用于对检索结果排序可有效提高检索效率,而网站根据搜索引擎的工作原理和排序算法改进网站结构、提高网站内容质量以增加其可见度;② 网站健康度检查:通过分析网站上的人链数与出链数、入链页面类型与出链页面类型等数据判断网站在网络中的影响力及健康状况;③ 知识挖掘:从链接分析算法与链接网络图中可以挖掘出网页、网站的潜在属性及潜在关联,以获得新知识,例如,可借助 Pajek 构建知识地图,将链接形成的网络关系可视化为一张二维图,从图中可以明显地判断出节点的重要程度与节点间关联的紧密程度。

信息科学有两个中心主题。第一个主题是“信息”。信息科学视角的链接分析的目标是传达有用的信息,而不仅仅是建立抽象的数学模型。第二个主题是“方法的稳健性”,主要是关于

结果的有效性和可靠性,而链接分析作为一种信息科学的研究方法,对于其稳健性的探讨自然必不可少。

国内链接分析研究起步较晚,至今没有该方面的任何专著,幸运的是,我们发现了英国知名学者 Mike Thelwall 教授的《Link Analysis: An Information Science Approach》一书。该书围绕信息科学的两个中心主题内容展开:

第一部分 理论:介绍信息科学视角的链接分析理论,包括一些基本的方法。

第二部分 网络结构背景:其他专业领域的研究可以为我们理解链接分析研究的结果提供有用的背景信息,并且有助于我们对网络中链接的用途形成一种直觉。

第三部分 学术链接:集中阐述学术链接分析。这部分主要有两个目的:一是主题本身,学术相关的链接是如何使用的,这方面的研究已经很多,在此全面概括其中最优秀的成果(2004);二是详细地解释和描述信息科学链接分析的方法,其中最核心的部分是讨论如何从链接中提取有用的信息。

第四部分 应用:列举了一系列链接分析的案例。这些案例是用来解释一系列不同的应用的,其中还包括个人搜索项目的详细内容。这部分可以跳过或者选读。

第五部分 工具和方法:介绍了链接分析中的方法和软件工具。关于各种工具的详细介绍可以在网络中查询,这部分主要是对这些工具和技术的链接分析性能进行一个大致的描述。本章适合那些打算自己开展链接分析研究的读者。

第六部分 总结:总结信息科学视角的链接分析的主要内容。

此外,书中提供了大量在线数据库、在线软件的相关信息,作为正文内容的补充。借助这些信息,读者便可以开始从事链接分析研究。

对于英文原著,从结构上看,逻辑严整,条分缕析;从内容上看,资料翔实,通过大量的方法、工具、技术介绍及案例分析,向读者展示了整个链接分析研究的全貌;从表达上看,文字表达深入浅出,用较简洁的语言描述了较复杂的理论与技术。

所有这些都是促成了我们蒙生将其译成中文的冲动,幸运的是,我们得到了 Mike Thelwall 教授的大力支持。他授予了我们独家翻译权,之后,又赠予我们全部电子书稿。在翻译过程中,我们力求保持原文原貌,将作者的观点完整地呈现给中国的读者,而在形式上,我们采用了国内学者更习惯的表达方式。出于对 Mike Thelwall 教授的感激,出于一种严谨的学术态度,我们兢兢业业地完成了这本译著,然而限于译者的水平,难免有翻译不当之处,我们愿意倾听读者的批评指正。

该书可作为国内情报学、图书馆学、信息资源管理、信息管理与信息系统、计算机科学与技术等专业的课堂教材,同时,其翔实的资料可作为国内情报学、计算机科学、传播学、社会学等领域学者从事链接分析研究的重要参考。

翻译过程中,各位译者的分工如下:第 1、4、25 章:孙建军、李江;第 2、3 章:叶晓飞;第 5、6 章:戴伟;第 7—11、26 章:曾雪娟;第 12—16 章:孙海霞;第 17—24 章:董珏、张煦。

最后,由李江、张煦负责全书的第一次审校、孙建军负责第二次审校并定稿。此外,邓中华、郑曦在语言处理方面付出了辛勤的劳动,在此一并致谢!

孙建军

2008 年 8 月于南京大学

目 录

第一部分 理论

第 1 章 前言	(3)
目标	(3)
链接分析	(3)
历史回顾	(3)
信息科学视角的链接分析是什么?	(4)
内容与结构	(5)
关键术语	(6)
小结	(7)
阅读与提高	(7)
参考文献	(7)
第 2 章 网络爬虫与搜索引擎	(9)
目标	(9)
引言	(9)
网络爬虫	(9)
查找网页	(10)
内容遍历与地址遍历	(11)
内容遍历	(12)
动态链接	(13)
遍历深度与人为限制	(13)
动态网页	(14)
道德规范和 robot.txt 文档	(15)
网页	(15)
网络爬虫小结	(16)
搜索引擎	(16)
公认的偏好	(17)
搜索引擎排序	(17)
网络档案	(18)
小结	(18)
阅读与提高	(18)
参考文献	(18)
第 3 章 链接统计的理论基础	(20)

目标	(20)
引言	(20)
链接统计的理论基础	(20)
异常	(21)
手工过滤和禁止列表	(22)
选择性文档模型(Alternative Document Model,简称 ADM)	(23)
网站和网络文档	(23)
ADMs 和标准 ADM 统计	(24)
ADM 域统计模型	(26)
选择链接统计方法	(26)
小结	(27)
阅读与提高	(27)
参考文献	(28)
第 4 章 对链接数的解释:随机样本与相关性	(29)
目标	(29)
引言	(29)
解释链接数	(29)
初步的可行性和有效性研究	(30)
全面的随机抽样	(31)
分类结果的置信度	(32)
相关性检验	(34)
文献回顾	(35)
小结	(35)
阅读与提高	(35)
参考文献	(35)
第二部分 Web 结构	
第 5 章 Web 图中的链接结构	(41)
目标	(41)
引言	(41)
Web 中的幂定律	(42)
Web 增长模型	(43)
链接拓扑结构	(45)
学术 Web 中的幂定律与链接拓扑结构	(46)
小结	(47)
阅读与提高	(48)
参考文献	(48)
第 6 章 Web 的内容结构	(50)
目标	(50)

介绍	(50)
Web 的主题结构	(50)
基于“链接—内容”的 Web 增长模型	(52)
链接文本	(52)
学术 Web 中的学科结构	(52)
共链	(57)
小结	(57)
阅读与提高	(57)
参考文献	(57)

第三部分 学术链接

第 7 章 大学:链接类型	(61)
目标	(61)
引言	(61)
引文分析	(61)
大学网站的作用	(62)
一国范围内的大学网站体系	(62)
页面类型	(63)
链接类型	(66)
小结	(68)
阅读与提高	(69)
参考文献	(69)
第 8 章 大学:链接模型	(71)
目标	(71)
引言	(71)
入链数和研究之间的关系	(71)
学术链接:质量与数量	(73)
备选的逻辑链接模型	(75)
数学模型	(76)
地理因素的影响	(76)
地区性群组	(77)
小结	(78)
参考文献	(79)
第 9 章 大学:国际链接	(80)
目标	(80)
引言	(80)
国内链接与国际链接	(80)
国际链接比较	(81)
语言的影响	(83)

小结	(84)
阅读与提高	(85)
参考文献	(85)
第 10 章 院系和学科	(87)
目标	(87)
引言	(87)
院系网站	(88)
链接类型中的学科差异	(88)
规模和相关性检验	(90)
地理和国际因素	(91)
小结	(91)
阅读与提高	(91)
参考文献	(91)
第 11 章 期刊和论文	(93)
目标	(93)
引言	(93)
期刊影响因子	(93)
期刊网站	(94)
期刊网站入链:存在的问题	(94)
期刊网站入链:案例研究	(95)
期刊论文中链接的类型	(96)
数字图书馆链接	(97)
与日志文件分析的结合	(97)
相关研究主题	(98)
小结	(98)
阅读与提高	(99)
参考文献	(99)

第四部分 应用

第 12 章 搜索引擎与网站设计	(103)
目标	(103)
引言	(103)
链接结构和爬虫爬行范围	(103)
网站中的文本和向量空间模型	(103)
PageRank 算法	(104)
案例研究:门户网站中的 PageRank 计算	(107)
HITS 算法	(109)
HITS 算法的原理示例	(110)
小结:根据 PageRank 算法和 HITS 算法进行网站设计	(113)

阅读与提高	(114)
附录:向量空间模型(Vector Space Model,简称 VSM)	(114)
参考文献	(115)
第 13 章 西班牙大学网站健康度检验	(117)
目标	(117)
前言	(117)
研究问题	(117)
研究方法	(117)
结果与讨论	(118)
结论	(123)
参考文献	(123)
第 14 章 链向大学网站的个人网页	(124)
目标	(124)
引言	(124)
网络信息发布与个人主页	(125)
研究问题	(126)
研究方法	(126)
数据搜集	(127)
数据分析	(127)
结果	(129)
ISP 偏好检验	(129)
ADM 匹配	(129)
链接与研究绩效的相关性	(130)
来自大学网站的入链与来自个人主页的入链之间的比较	(132)
个人网页分类	(132)
结论	(136)
小结	(136)
致谢	(137)
参考文献	(137)
第 15 章 学术网络	(140)
目标	(140)
引言	(140)
研究方法	(140)
大学网站地图	(140)
国内学术网络图	(143)
学科地图	(144)
小结	(147)
阅读与提高	(147)
参考文献	(148)

第 16 章 商业网站	(149)
目标	(149)
引言	(149)
网站覆盖范围检查	(149)
站点索引和排名	(149)
竞争情报	(150)
案例研究	(150)
Center Parcs	(151)
Hoseasons	(152)
Butlins	(152)
Pontins	(153)
Haven Holiday	(153)
通用查询	(154)
小结	(155)
阅读与提高	(155)
参考文献	(155)

第五部分 工具和方法

第 17 章 商业搜索引擎和网络档案的使用	(159)
目标	(159)
引言	(159)
检验结果	(159)
处理结果的变化	(160)
使用多个搜索引擎	(161)
使用网络档案	(161)
小结	(161)
在线资源	(162)
阅读与提高	(163)
参考文献	(163)
第 18 章 个人爬虫	(164)
目标	(164)
引言	(164)
个人爬虫类型	(164)
SocSciBot	(165)
检索到的网页	(165)
网页的限制条件	(166)
网络链接提取	(166)
来自 HTTP 的 URL	(167)
模糊的或未详细说明的 URL	(167)

(1.8) 动态页面·····	(168)
(1.8) 错误处理·····	(168)
(1.8) 爬行中的人为干预·····	(169)
SocSciBot tools ·····	(169)
小结·····	(170)
在线资源·····	(170)
阅读与提高·····	(170)
参考文献·····	(172)
第 19 章 数据清理 ·····	(173)
(19.1) 目标 ·····	(173)
(19.1) 引言 ·····	(173)
(19.2) 数据清理方法概述 ·····	(173)
(19.3) 识别异常 ·····	(173)
(19.4) TLD 光谱分析 ·····	(174)
(19.5) 小结 ·····	(175)
(19.6) 在线资源 ·····	(175)
(19.7) 参考文献 ·····	(175)
第 20 章 大学在线链接数据库 ·····	(176)
(20.1) 目标 ·····	(176)
(20.1) 引言 ·····	(176)
(20.2) 链接数据库概述 ·····	(176)
(20.3) 链接结构文件 ·····	(177)
(20.4) 禁止列表 ·····	(178)
(20.5) 数据分析 ·····	(178)
(20.6) 其他链接结构数据库 ·····	(178)
(20.7) 小结 ·····	(178)
(20.8) 在线资源 ·····	(179)
(20.9) 阅读与提高 ·····	(179)
(20.10) 参考文献 ·····	(180)
第 21 章 嵌入式链接分析方法 ·····	(181)
(21.1) 目标 ·····	(181)
(21.1) 引言 ·····	(181)
(21.2) 网络空间分析(Web Sphere Analysis, 简称 WSA) ·····	(181)
(21.3) 虚拟民族志(Virtual Ethnography) ·····	(182)
(21.4) 小结 ·····	(183)
(21.5) 阅读与提高 ·····	(183)
(21.6) 参考文献 ·····	(183)
第 22 章 社会网络分析 ·····	(184)
(22.1) 目标 ·····	(184)

引言	(184)
SNA 指标	(184)
软件	(186)
小结	(186)
阅读与提高	(186)
参考文献	(187)
第 23 章 网络可视化	(188)
目标	(188)
引言	(188)
网络图表	(188)
大型网络图表	(190)
多维尺度分析	(190)
自组织地图	(191)
认知领域可视化	(191)
小结	(191)
在线资源	(192)
参考文献	(193)
第 24 章 学术链接指标	(195)
目标	(195)
引言	(195)
作为过程指标的网络指标	(195)
规模和可靠性问题	(196)
基准指标	(197)
链接计量指标	(197)
相关指标	(198)
其他计量指标	(199)
小结	(199)
阅读与提高	(200)
参考文献	(200)
第六部分 总结	
第 25 章 总结	(205)
目标	(205)
引言	(205)
信息科学对链接分析的贡献	(206)
其他的链接分析方法	(207)
未来的方向	(207)
第 26 章 术语表	(208)
参考文献	(210)

附录:SocSciBot 使用指南

使用指南.....	(211)
第一步:安装 SocSciBot、SocSciBot Tools 和 Cyclist	(211)
第二步:安装 Pajek	(212)
第三步:使用 SocSciBot 爬取第一个网站	(212)
第四步:使用 SocSciBot 爬行另外两个网站	(216)
第五步:浏览 SocSciBot Tools 生成的关于“small test”项目的基本报告	(216)
第六步:使用 Pajek 生成网络图	(219)
第七步:浏览 Pajek 生成的站点图	(223)
第八步:使用 Cyclist	(224)
小结.....	(225)

第一部分

理 论

第1章 前言

目标

- 介绍本书的内容和结构,以及一些关键术语。
- 介绍信息科学研究方法——链接分析。

链接分析

链接分析在许多领域中有着广泛的应用,如计算机技术、理论物理、信息科学、传播学以及社会学等。之所以能有这样广泛的应用,一方面是因为网络的重要性;另一方面是因为人们普遍认为:从网页之间的超链接中能够提取各种有用的信息。这种认识主要源于一些相关的因素:① Google 的巨大成功,主要是利用一种基于链接的算法来判断网页的相关度;② 期刊引用、社会人际关系等类似现象;③ 网络用户每天都面对各种用于研究、或用于商业、或用于娱乐的链接。

在本书中,笔者的主要目的是向新读者介绍什么是信息科学视角的链接分析。之后,读者们就能够评价现有的研究,甚至从事自己的研究项目、形成自己的研究方法。在本书中,笔者深信信息科学的方法对于其他学科的研究人员同样有着广泛的实用价值,尤其是那些对在线分析感兴趣的社会学家。在研究过程中,如果将所有类型的链接分析都纳入考虑范围,必将一无所得,因为有些领域要求详细的数学算法,而另一些领域则只需要定性的分析。本书中至少有一半的内容是研究学术网络或学术交流的,因此,读者们也可以对学术交流有一些深入的了解。

本书主要探讨了以下 4 个问题:

- 通过分析网站或者网页之间的超链接可以提取什么样的信息?
- 应该使用哪种技术提取信息?
- 链接分析可能存在哪些缺陷?
- 是否应该进行链接分析? 怎样进行链接分析?

历史回顾

网络链接分析研究可以追溯到 1995—1996 年。当时,网络链接分析研究在几个学科中同时出现,包括计算机科学领域中的搜索引擎开发(Weiss, Velez, Sheldon et al., 1996),数学领域中的结构和复杂性分析(Abraham, 1996)。首次提出可将信息技术应用到因特网的信息科学家是 Brazilian Marcia J. Bossy(1995),其文章发表在《French online journal》上。而首次真