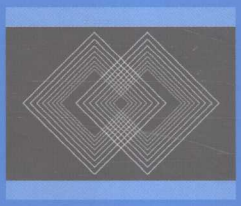
现代信息
科学译丛

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

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

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：通过分析网站上的入链数与出链数、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

第二部分 Web 结构

第三部分 学术链接

第四部分 应用

第五部分 工具和方法

第六部分 总结

附录:SocSciBot 使用指南

第一部分

理　论

第一部分

绪 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