基于社会网络分析的信息检索结果可视化呈现方法研究*

周姗姗 毕 强 高俊峰
(吉林大学管理学院 长春 130022)

【摘要】分析传统检索结果呈现方法的两个缺陷,提出一种基于社会网络分析法的数字图书馆学术信息检索结果可视化呈现方法,从科研作者的中观维度出发,以检出文献作者间的相互关系为切入点,构建基于作者科研网络结构的邻近矩阵,探寻检索问题下学术信息的隐形知识聚合与关联,并将重新聚合后检索结果以图谱的形式直观展现给用户,达到改善数字图书馆知识服务的目的。

【关键词】数字图书馆 社会网络分析 信息检索 可视化 知识聚合
【分类号】G250

A Method of Information Retrieval Results Visualization Based on Social Network Analysis

Zhou Shanshan Bi Qiang Gao Junfeng
(School of Management, Jilin University, Changchun 130022, China)

【Abstract】This paper analyzes the two defects of traditional presentation methods for retrieval results, and proposes a method of information retrieval results visualization based on social network. Starting from the medium dimension of scientific research authors and with the purpose of figuring out the relationships among the document authors, the paper builds an adjacent matrix based on the authors' research network structure and explores the invisible knowledge retrieval problem under the academic information aggregation and association. The re-aggregation of the retrieval results in form of a graph are directly shown to the users in order to achieve the purpose of improving digital library knowledge service.

【Keywords】Digital library Social network analysis Information retrieval Visualization Knowledge converge

1 引 言

数字图书馆提供最终服务是向用户展示检索结果,但检出信息量过大,使得用户在徒劳而又机械的翻页浏览模式下迷失。可视化检索技术虽然能够在一定程度上解决检索结果冗余而导致的迷航问题,但因当前可视化检索的主要思想是基于主题再分类,以及以经典排序算法 PageRank 为基础的检索结果可视化方式,导致形成检索结果的可视化主题图对文献之间隐性知识的挖掘以及作者之间学术思想的相互影响与借鉴方面表现不足,无法

收稿日期:2013-08-07
收修改稿日期:2013-10-09
*本文系国家自然科学基金项目"语义网络环境下数字图书馆资源多维度聚合与可视化研究"(项目编号:71273111)、国家社会科学基金重大项目"基于语义的馆藏资源深度聚合与可视化研究"(项目编号:11&ZD152)和吉林大学 985 工程项目的研究成果之一。

满足用户对数字图书馆知识服务的需要。因此,如何将检索结果多维度地、直观可视化地呈现给用户,成为数字图书馆发展过程中急需解决的问题。本文从科研作者的中观角度出发,运用社会网络分析方法中基于邻近关系矩阵而形成的社会网络图谱,力图以直观的视觉方式勾画出社会体之间的二元关系,构造基于作者的社会网络图谱,利用 SNA 图谱节点和有向线段直观地描述出检索问题下作者科学合作的互为关系,并通过引入 Folksonomy 标签云,从微观角度(关键词)揭示文献特征单元之间的必然联系,从而有效解决传统检索结果呈现方法的困局,最终达到提高数字图书馆服务质量的目的。

2 研究背景与相关工作

目前较为流行的可视化检索主要从语义、概念相互关系、主题分类等角度切入,构造了主题树、概念地图等图形体系,且在检索过程可视化、语义呈现可视化、可视化的用户分析等方面均有一定的研究成果。时至今日,国内外学者更致力于探索检索结果的图形化表现。

在数字图书馆文本信息可视化的研究方面,施乐帕克研究中心(Xerox Palo Alto Research Center, Xerox PARC)做出了代表性的研究成果,如在美国加州大学伯克利分校(University of California, Berkeley)数字图书馆项目中的 TileBars 系统,采用结构化检索模型,着眼于文档内部结构,根据检索词在文献中出现的频率来确定相关度。检索结果界面能提供每篇文献相对长度、查询条件集合在文献中的频率、查询条件集合关于文献和它们彼此的分布状况[1]。另外其开发的 Scatter/Gather 系统,提供基于集群的文档浏览方法,替代组织排名标题查看检索结果[2]。Chen 等[3]提出一种新的工具以解决现有共引可视化工具的缺陷。此外,Sokhn 等[4]提出的 HELO(High – level modEL for cOnference)模型可用于细粒度的搜索条件和复杂的查询,以提高知识检索和可视化。检索结果集可视化方面,Howard White 等开发的 ConceptLink 系统[5]通过检索结果的词频分析来检出关联文献。2011 年 TPDL 会议中,Wong 等[6]展示的可视化检索系统 INVISQUE,使用索引卡片方式来揭示图书馆内容以发现最新发表的文献与引用最多的文献。Groxis[7]作为一种信息管理与搜索工具,通过实现动态聚类、可视化结果地图,来提高用户探索、组织与共享数字信息的水平。AquaBrowser[8]作为 Serials Solutions 公司的商业产品,通过共现分析找出关联词汇,以可视化“词云”的方式揭示相关资源,使检索过程更方便快捷。王畅[9]分析了文献数据库 EBSCO 检索结果的可视化界面,对其特点、模块功能以及适用群体进行了详细的分析。许德山等[10]也将检索的过程与结果进行了可视化。数字图书馆可视化应用方面,Shen 等[11]发表了 VIDI 协议,提出了加强互操作性数字图书馆(DLS)和可视化系统(VIS);Ha 等[12]分析设计了一个基于本体驱动的语义检索可视化系统;许德山等[13]提出了面向本体知识库可视化检索的实现思路以及可视化平台的功能设计。

综上所述,无论是国内还是国外,信息检索的可视化研究都是当前学术界关注的重要课题。尽管在检索结果可视化方面积累了较为丰富的成果,但鲜见与社会网络分析法结合的实例,本文提出的检索结果可视化方法中,凝聚子群、点度中心性与结构洞能分析出检出文献的作者之间的学术交流关系、核心作者群、核心文献群等内容,有利于用户通过直观生动的方式了解主题知识的全貌,快速筛选出最有价值的信息,获得更理想的检索结果,获取更有深度的知识服务。

3 基于社会网络分析的检索结果可视化聚合与呈现

3.1 研究方法

本文提出基于社会网络分析的可视化检索结果呈现方法,不仅可以利用 SNA 图谱节点和有向线段直观地描述出检索问题下作者科学合作的互为关系,进而弥补用户构造检索表达式时的片面性,还可以通过测量图谱中节点的“度(Degree)”量化地刻画出作者在问题领域中权威性,使得用户在检索结果中进行二次查询时有的放矢。

3.2 技术路线

基于社会网络分析法的数字图书馆检索结果可视化呈现方法的实施流程如图 1 所示。

(1)检出数据清洗

针对数字图书馆的检索结果一般是规范度较高的学术文献资源,因此该步骤只是将资源的外部特征加以筛选和整理,以便确定可用于分析的特征单元。

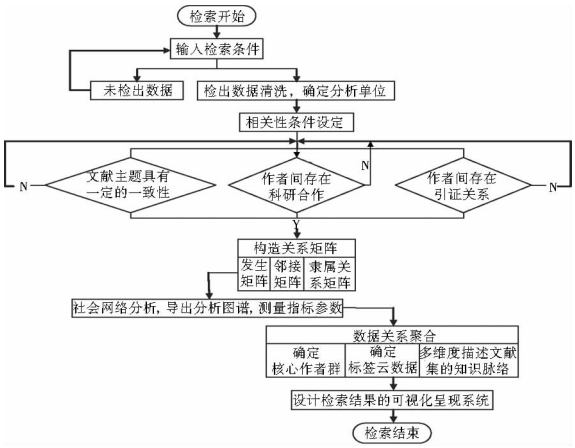

图 1　基于 SNA 的检索结果可视化实施流程

(2)相关性条件设定

基于检出文献的作者之间的相互关系构建关系矩阵。设定作者之间的相关关系主要有三种:主题之间存在关联关系、作者之间存在合作关系、作者之间存在引证关系。

(3)构造关系矩阵

社会网络分析法中的关系矩阵包括邻接矩阵(Adjacency Matrix)、发生矩阵(Incidence Matrix)和隶属关系矩阵(Affiliation Matrix)三种:

①邻接矩阵的第一行和第一列都代表相同的角色,一般为二值方阵,即两个行动者是否具有关系在矩阵中用"0 和 1"体现[14]。

②发生矩阵的"行"代表图谱中节点,而"列"代表图谱中连接各节点的线,即发生矩阵表达的是哪个点连接在哪条线上,本文采用社会网络分析软件 UCINET 绘制图谱,在此不构造发生矩阵。

③隶属关系矩阵描述的是行动者的隶属关系,即角色的自然属性和社会属性。隶属关系矩阵的"行"为行动者,"列"为各种属性[11]。

本文基于检出文献的作者关系构造邻接矩阵。

(4)社会网络分析

将构造出的关系矩阵导入社会网络分析软件中,诱导出分析图谱,再根据图谱测量出目标群体中各行动者的指标参数,这些分析指标多以数据集合的形式呈现。

(5)数据关系聚合

根据社会网络分析步骤中得出的各项参数,量化确定检出文献的核心作者群、标签云数据,多维度地描述检出文献集的知识脉络。

(6)可视化呈现

设计检索结果的可视化呈现系统,建立网络链接、基于社会网络分析法得出的图谱构造用户友好的检索结果可视化呈现界面。

4　实证研究

4.1　基于社会网络分析法的检索结果呈现分析

以"概念格 and 数字图书馆"为检索主题,在 CNKI 数据库中进行检索,截至 2013 年 7 月 2 日共计检出 28 篇相关文献,以此作为基础数据进行研究。

首先对检出文献作者关系进行分析,构造关系矩阵。鉴于第二作者科研产出率较低,因此不予考虑,通过赋予相关关系得出作者关系矩阵,如表 1 所示:

表 1　检出文献的作者关系矩阵

	…	郭　强	汪胜楠	黄琳颖	宋绍成	王　磊	董　洁	…
…	…	…	…	…	…	…	…	…
郭　强	…	0	0	0	0	0	0	…
汪胜楠	…	0	1	1	0	1	0	…
黄琳颖	…	0	0	1	0	1	0	…
宋绍成	…	0	0	0	0	0	0	…
王　磊	…	0	1	1	0	0	0	…
董　洁	…	0	0	0	0	0	0	…
…	…	…	…	…	…	…	…	…

矩阵中行与列均为相同行动者,作者间的相互关系由"1","0"直观体现。碍于篇幅有限,在此并没有将关系矩阵全部列出,有些作者群体的全部关系在图谱中体现。信息检索结果导航不应该仅仅局限于反映信息资源物理存储地址的静态链接,更应该在用户选择的浏览路径上提供检出作者之间科研相关性的逻辑关系揭示以及知识指导。利用数字图书馆多维聚合可视化检索系统进行检索后,基于社会网络分析图谱的检索结果可视化呈现如图 2 所示:

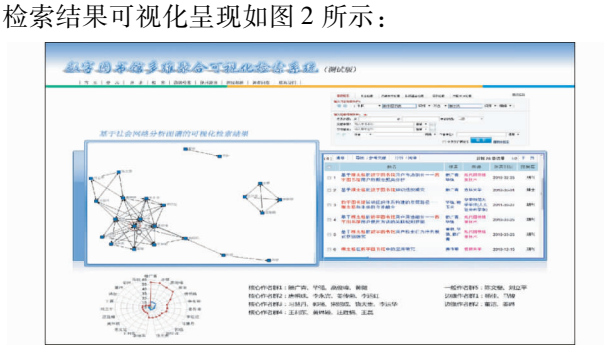

图 2　基于社会网络分析图谱的检索结果可视化呈现

图2左侧即通过作者关系矩阵诱导出的关系图谱，作者之间的相互关系由有向线段列举，检出文献作者的凝聚子群一目了然。通过观察，发现作者毕强、滕广青、黄微在图谱中具有较高的权威性，各自的D(K)值分别为0.55、0.55与0.5。这表示上述几位学者在“概念格and数字图书馆”研究领域有较高的建树，其科研成果与社会网络中其他作者相关度较高，能够在一定程度上满足用户的需要。

图2的右侧以传统的列表形式呈现单独作者的检索文献，用户可根据选择左侧图谱的不同节点而浏览该节点作者的检出文献群，是对可视化图谱的一种补充。

通过计算，分析出文献作者在“概念格and数字图书馆”主题网络下的各自绝对中心度、相对中心度以及分享度如表2所示：

表2　文献作者社会网络分析权重

作者	1 Degree	2 NrmDegree	3 Share
1　滕广青	12.000	54.545	0.100
2　毕　强	12.000	54.545	0.100
4　黄　微	11.000	50.000	0.092
9　高俊峰	11.000	50.000	0.092
16　唐明珠	8.000	36.364	0.067
22　李永宾	6.000	27.273	0.050
3　姜传菊	6.000	27.273	0.050
19　李运红	6.000	27.273	0.050
17　习慧丹	5.000	22.727	0.042
10　郭　强	5.000	22.727	0.042
13　宋绍成	5.000	22.727	0.042
8　饶天贵	5.000	22.727	0.042
5　李运华	5.000	22.727	0.042
18　王利东	3.000	13.636	0.025
20　陈文斐	3.000	13.636	0.025
12　黄琳颖	3.000	13.636	0.025
11　汪胜楠	3.000	13.636	0.025
23　刘立平	3.000	13.636	0.025
14　王　磊	3.000	13.636	0.025
21　杨　佳	2.000	9.091	0.017
15　董　洁	1.000	4.545	0.008
6　姜　琳	1.000	4.545	0.008
7　马　骏	1.000	4.545	0.008

表2根据全部23个作者的权重高低降序排列，D(K)取值区间为(0.5,1)的节点包括{1,2,4,9}，D(K)取值区间为(0.2,0.5)的节点包括{16,22,3,19,17,10,13,8,5}，D(K)取值区间为(0,0.2)的节点包括{18,20,12,11,23,14,21,15,6,7}，根据图谱显示的相互关系以及各自权重，产生的7个作者群落分别为：

(1)滕广青、毕强、黄微、高俊峰；

(2)唐明珠、李永宾、姜传菊、李运红；

(3)习慧丹、郭强、宋绍成、饶天贵、李运华；

(4)王利东、黄琳颖、汪胜楠、王磊；

(5)陈文斐、刘立平；

(6)杨佳、马骏；

(7)董洁、姜琳。

上述7个作者群落按权重高低分为核心作者群、一般作者群及边缘作者群，并以降序排列到图2下方，从数据的主题分析、内容传递等方面入手，使得信息资源按照作者研究内容间的相关性进行动态的聚类，从而将内容从作者权威性的角度有序表达出来提供给用户，将背景知识整合到检索结果呈现的页面中指导用户导航，提高检索表现。

检索结果通过社会网络分析法的形式加以聚合，然而社会网络分析图谱在体现文献资源内容特征方面还有欠缺，因此，本文通过引入Folksonomy Tag作为补充，将文献内容特征主题词的权重按字号大小直观体现，如图3所示：

图3　检索结果内部特征的可视化呈现

从图3可以看出，在检出的28篇文献中关键词“概念格、数字图书馆、数据挖掘，本体”等词汇在文献中标引的频率较高，并分别指向以作者群为类目展现的文献个体。

4.2　基于社会网络分析法的检索可视化结果分析

通过分析检出文献作者，发现作者“王磊、王利东、汪胜楠、黄琳颖”的科研产出成果主要为概念格算法方向，与检索词“数字图书馆”不符，属误检文献，在数据清洗步骤中并没有精准地筛选错检文献，但是由于社会网络分析法以中观层次的作者相关关系为出发点，因此在图谱中，作者“王磊、王利东、汪胜楠、黄琳颖”由于社会网络关系相近而被归并为一类子群中。在海量检索结果处理过程中，按社会网络分析图谱形式呈

现的可视化检索结果能将错检的文献作者群落准确地排斥,满足了用户一步到位的检索需求,并且减少了无效查询时间。

经过测试,不难看出,该方法不仅可以有效提高检索效率,使得检索结果生动直观,而且有助于用户方便快捷地获取信息。但该方法更适用于规范度较高的数字图书馆资源,且按照检索关键词检出的文献数量不大,而对于检出文献数量较大、知识背景复杂的情况,该方法不能获得预期的呈现效果,需进一步设定检索条件,缩小检索范围,以达到最佳效果。

5　结　语

通过实践应用发现,基于社会网络分析图谱的检索结果呈现方法能够有效地解决传统检索结果呈现方法的两方面缺陷:

(1)SNA 图谱克服了导航系统浏览路径单向,系统无法从多视角与用户之间形成有效交互的局限;

(2)Folksonomy Tag 呈现的资源内部特征弥补了用户信息筛选过程中受短时记忆限制增大迷航可能性的不足。

因此该方法不仅可以从多种维度出发,以检出文献作者间的相互关系为切入点,构建基于作者科研网络结构的邻近矩阵,探寻检索问题下学术信息的隐形知识聚合与关联,还可根据聚类与关联分析结果,将重新聚合后的检索结果以图谱的形式直观展现给用户,进而为个性化信息服务提供支持,满足用户个性化的知识需求,并以此改善数字图书馆知识服务。然而,随着检索用户对于检索效率要求的提高,如何更好地利用社会网络分析法可视化文献深层语义将是进一步深入研究的方向。

参考文献:

[1] TileBars: Visualization of Term Distribution Information in Full Text Information Access [EB/OL]. [2013 - 07 - 05]. http://people. ischool. berkeley. edu/ ~ hearst/papers/tilebars - chi95/chi95. html.

[2] Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results [EB/OL]. [2013 - 07 - 05]. http://www. researchgate. net/publication/2633414_Reexamining_the_Cluster_Hypothesis_ScatterGather_on_Retrieval_Results.

[3] Chen T T, Yen D C. CociteSeer: A System to Visualize Large Citation Networks[J]. *The Electronic Library*, 2010, 28(4): 477 - 491.

[4] Sokhn M, Mugellini E, Khaled O A. Knowledge Modeling for Enhanced Information Retrieval and Visualization[J]. *Advances in Intelligent and Soft Computing*, 2010, 67: 199 - 208.

[5] ConceptLink: Visual Exploration of Medical Concepts[EB/OL]. [2013 - 11 - 03]. http://project. cis. drexel. edu/conceptlink/.

[6] Wong B L W, Choudhury S, Rooney C, et al. INVISQUE: Technology and Methodologies for Interactive Information Visualization and Analytics in Large Library Collections[C]. In: *Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries: Research and Advanced Technology for Digital Libraries (TPDL'11)*, Berlin, Germany. Berlin, Heidelberg: Springer - Verlag, 2011: 227 - 235.

[7] Groxis [EB/OL]. [2013 - 07 - 18]. http://www. groxis. com.

[8] AquaBrowser [EB/OL]. [2013 - 07 - 22]. http://www. serialssolutions. com/en/services/aquabrowser.

[9] 王畅. 可视化信息检索技术的应用——以 EBSCOhost2.0 为例[J]. *图书馆学刊*, 2010(6): 87 - 89. (Wang Chang. An Application of Visual Information Retrieval Techniques——Case in EBSCOhost2.0 [J]. *Journal of Library Science*, 2010(6): 87 - 89.)

[10] 许德山,张智雄, 邢美凤. 面向本体知识库的可视化检索研究[J]. *情报理论与实践*, 2010, 33(8): 114 - 117. (Xu Deshan, Zhang Zhixiong, Xing Meifeng. Research on Visual Retrieval Oriented to Ontology Knowledge Base [J]. *Information Studies: Theory & Application*, 2010, 33(8): 114 - 117.)

[11] Shen R, Wang J, Fox E A. A Lightweight Protocol Between Digital Libraries and Visualization System[C]. In: *Proceedings of JCDL Workshop on Visual Interfaces to Digital Libraries*. London: Springer - Verlag, 2002: 217 - 225.

[12] Ha I, Oh K, Hong M, et al. Ontology - driven Visualization System for Semantic Searching[C]. In: *Proceedings of International Conference on Information Science and Applications (ICISA'11)*, Jeju Island, South Korea. 2011: 1 - 6.

[13] 许德山,张智雄. 面向本体知识库的可视化检索研究[J]. *情报理论与实践*, 2010, 33(8): 114 - 117. (Xu Deshan, Zhang Zhixiong. Research on Visual Retrieval Oriented to Onology Knowledge Base[J]. *Information Studies: Theory & Application*, 2010, 33(8): 114 - 117.)

[14] 戴维. 诺克, 杨松. 社会网络分析[M]. 李兰译. 上海: 格致出版社, 2012: 17 - 23. (David Knoke, Yang Song. Social Network Analysis[M]. Translated by Li Lan. Shanghai: Truth & Wisdom Press, 2012: 17 - 23.)

(作者 E - mail: zhou33jl@126. com)