基于搜索的科研论文自动评价

罗江琴,阳小华,马家宇
(南华大学,湖南 衡阳 421001)

摘 要:互联网上信息量的增加,迫切需要一些新的科研论文评价方法来使科研工作者能更便捷地查阅有价值的文献。分析了传统的人工科研论文评价方法存在的缺点,介绍了Google Scholar学术搜索及其特点。提出一个新的基于搜索的科技论文自动评价方法,通过搜索获取原始论文集、引文和被引用文献,构造论文网络,利用基于Web社区发现技术形成论文社区,获得社区的核心论文,对指定的论文进行评价。该方法在评价论文时避免了人工操作的缺点,结果也很准确。

关键词:自动评价;元搜索引擎;论文社区
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2007)11-0080-04

Paper Auto-Evaluation Based on Search Engine

LUO Jiang-qin, YANG Xiao-hua, MA Jia-yu
(Nanhua University, Hengyang 421001, China)

Abstract: The growth of Web contents increases the need for some new methods of paper evaluation to help reader to find the important paper. In this paper, analyzed the shortcoming of traditional paper evaluation approach that relies on manual work. Google scholar and its characteristic were introduced. Lift a new method of paper auto-evaluation that extract citation, reference, original papers from search result; form paper network; build paper community based on Web community's identification, then obtain core paper of the paper community and evaluate the specifically papers by the mean time. The method avoided the disadvantage on manual work to get exact result.

Key words: auto-evaluation; meta search engine; paper community

0 引 言

科研工作者研究的过程经常需要查阅大量的文献资料,尤其是科研论文。通常查阅科研论文的方法是到各种相关期刊或者索引上进行手工检索,随着计算机和互联网的发展,通过关键字信息在互联网上可以搜索到很多相关的论文。但在这大量的论文当中,并不是每一篇论文都有参考价值,很可能在读者读完大量文献后发现有用的文章可能就只有几篇。为了避免这一类问题的出现,就需要对科研论文进行准确的评价,需要读者对每篇论文进行浏览,选择出质量高的文献,这一过程人工操作既复杂又费时,而且对知识渊博经验丰富的读者而言比较容易,对普通的读者尤其是当他们要进入一个新的领域研究的时候则是一件困难的事。

要解决这个困难,必须面对两个问题:第一,怎样才能快速简便地获得领域内的重要文献资料;第二,如何对科技论文进行准确的评价。

关于科研论文的评价,国内外研究人员近年来进行了许多有益的研究和尝试,目前对科研论文的评价主要有同行专家评价和文献计量学分析方法[1]。科研论文的评价除了可以对已发表的论文进行学术价值的评价外,还可以对尚未发表的论文进行评价,比如编辑在审稿过程中需要评价论文的学术价值来判定是否有发表的价值。

对于上面提出的第一个问题,也有很多科研论文检索系统已经开始研究并付诸实现,比如Web of Science[2],Scopus[3]以及Google Scholar,但是都没有很好地解决。

1 Google Scholar学术搜索及其特点

专用的科学论文搜索引擎Google Scholar是Google公司专门针对科研人员提供的搜索引擎,它所包含的资源除包括原Google普通搜索引擎已经包含的一系列出版商和集团,如计算机协会(ACM)、IEEE等外,还包括一些学术出版物、专业学会、预印本库、大学及网上学术论文。搜索内容覆盖期刊论文、学位论

收稿日期:2007-01-17
基金项目:衡阳市科技局项目(2005KR01-002)
作者简介:罗江琴(1976-),女,四川合江人,讲师,研究方向为信息检索、评价与决策支持系统。

文、图书、预印本、文摘、技术报告等学术文献[4]。Google Scholar 按照查询关键字进行搜索,检索结果主要按照文章与查询关键字的相关度进行排序,它采取自动分析与抽取引文的方法,首先按照检索词出现在文献题目中的位置来进行排列,同样出现在标题中的文献再按照引用次数来排列,未被引用的文献排列在后,其次是按照检索词出现在文献的其它部分排列。

Google Scholar 的特点[5,6]如下:

1)提供高级检索功能。可以直接通过作者检索某一作者的文章;能够限定检索来自某一专门出版物的文章,如果确信查找什么,采用出版物限制检索非常方便;能够限定检索某一(或一段)时间内出版的论文等。

2)提供被检索文章的引用信息,说明一篇文章在学术文献中被引用的频率。Google Scholar 本身也是一个引文数据库,对许多引用来说,用户会发现一个直接的链接到 Google Scholar 数据库中其他文章的链接,引用了用户所选择的文章。

Google Scholar 虽然有以上优点,但是也存在不足:

(1)对科研论文价值的评价不准确。排在前面的文献是检索词出现在文献标题中且被引用次数相对较高的文献或者引文,这样的论文往往不是最有价值的。Google 学术搜索的排名技术还考虑到每篇文章的完整文本、作者、刊登文章的出版物以及文章被其他学术文献引用的频率。其相关度排序主要是以被引用次数为评价原则。虽然按被引次数来评价文献的价值是评价论文质量的重要因素之一,但是不能准确地评价一篇论文,因为被引次数基本上随着时间的变化会相对的增加,比如,一片新发表的高质量的论文,它的被引次数远远低于发表日期较长的论文。虽然 Google 学术搜索增加了最新文章的搜索,但是评价新论文的原则和老论文的原则不加区分,也不能得到准确的评价。

(2)被引用次数的统计结果不准确。虽然 Google Scholar 包含了多个数据源,但和 Web of Science 比较起来还是少很多[7],所以统计结果偏低。同时 Google Scholar 根据超级链接来对文献进行区分,有些引用文献相同但是往往来自不同的数据源所以其链接不一样,这样在计算被引用次数时经常出现重复计算的情况。

2　利用元搜索引擎技术进行论文检索

通过对 Google Scholar 的分析,在此提出了一种新的科研论文搜索和评价的方法。科研论文的搜索是利用元搜索引擎技术实现。由于在科研论文之间存在引用和被引用的关系,可以通过这样的链接来获取与主题相关的论文网络。

元搜索引擎(meta search engine),它是多个搜索引擎的集成,是建立在已有的搜索引擎(成员搜索引擎)服务之上的一种搜索引擎[8]。元搜索引擎工作过程可以总结为如下几步:

(1) 接收用户提供的原始查询;

(2) 把原始查询分别转换为各个成员搜索引擎能够接受的形式;

(3) 向成员搜索引擎发送查询;

(4) 收集各个搜索引擎的原始查询结果;

(5) 对原始查询结果进行合成,形成最终结果;

(6)把最终查询结果递交给用户[9]。

对用户而言只发出了一次查询,使用了一个搜索引擎,获得的结果是多个搜索引擎的综合,提高了查询的全面性和准确性。

利用元搜索引擎技术建立基于搜索的科研论文自动评价,其搜索部分的结构原理和元搜索引擎基本一致,原始查询是待评价论文的相关信息,比如论文的标题、关键字信息、所属领域及其评价指标名称等。由于搜索的信息是与科研论文相关的,所以不必对互联网进行全网搜索,只需要确定与论文相关的数据源即可,比如中文期刊数据(CNKI),科研论文社区等,Google 学术搜索和 Citeseer 等都可以作为数据源之一。这里的数据源可以看作是在元搜索引擎中的成员搜索引擎,原始查询经过查询转换后转发到各个数据源,最后将各个数据源返回查询结果进行结果合并处理,结果处理包括有用信息的提取、评价等。处理完毕后将评价结果及相关信息提交给用户。

每一个用户在进行文献搜索的时候希望搜索回的文献比较全面并且都是对自己最有用的(即最有学术价值的文献),利用元搜索引擎获得的搜索结果并不一定能够满足用户的需求。通过统计分析,发现用户在浏览结果页面的时候对前 5 页的结果点击率比较高,也就是说前 50 条结果里面可能有比较多让用户满意的结果。因此取前 50 条结果作为原始论文集,在这里论文集的选取也可以由用户来进行操作,用户可以将检索结果里面感兴趣的文献列入到原始论文集里。

3　论文网络的形成

论文网络的形成是在原始论文集的基础上形成的,假设原始论文集为 S,形成的过程如下:

1) 对 S 里的每一篇文献,将它的参考文献和被引用文献提取出来加入到集合 S 里面,得到一个增集 S'。通过实验证明,增集 S' 虽然包含了比 S 更多的结

果,但是其全面性还不够,同时有些本领域内的重要文献还没有被包含到该集合中来。

2) 对增集 S' 再一次进行扩充,同样加入它所有的参考文献和被引用文献,最后得到的集合为 G。

对 S 进行两次扩充后得到集合 G,G 里的一个元素就是引文网络的每一个节点,节点之间的联系主要是引用和被引用关系,这种关系可以用有向图来表示,也可以用矩阵来表示。文中为了计算方便,在这里用引证矩阵和被引矩阵[10]来表示该引文网络,引证矩阵如图 1 所示,其中行标题代表引用文献,列标题代表被引用文献,第 i 行,j 列位置的数据代表论文 i 是否引用了论文 j,如果有引用关系则为 1,否则为 0。被引证矩阵如图 2 所示,其中行标题代表被引用文献,列标题代表引用文献,第 i 行,j 列位置的数据代表论文 i 是否被论文 j 引用,如果有引用关系则为 1,否则为 0。事实上,引证矩阵和被引矩阵互为逆阵。

	论文 1	论文 2	论文 3	论文 4	论文 5	论文 6	论文 7	论文 8
论文 1	0	0	0	0	0	0	1	1
论文 2	1	0	0	0	0	0	0	0
论文 3	1	1	0	0	0	0	0	0
论文 4	1	0	0	0	0	0	0	0
论文 5	1	1	1	0	0	0	0	0
论文 6	0	1	1	0	0	0	0	0
论文 7	0	0	0	0	0	0	0	0
论文 8	0	0	0	0	0	0	0	0

图 1 引证矩阵

	论文 1	论文 2	论文 3	论文 4	论文 5	论文 6	论文 7	论文 8
论文 1	0	1	1	1	1	0	0	0
论文 2	0	0	1	0	1	1	0	0
论文 3	0	0	0	0	1	1	0	0
论文 4	0	0	0	0	0	0	0	0
论文 5	0	0	0	0	0	0	0	0
论文 6	0	0	0	0	0	0	0	0
论文 7	1	0	0	0	0	0	0	0
论文 8	1	0	0	0	0	0	0	0

图 2 被引矩阵

4 科研论文社区发现算法

在上述的论文网络里,可以发现有些论文之间链接密度比较大,有些则很疏,其中紧密相连的那些论文所描述的主题应该是一致的,这就是想要找的科研论文社区。在实现自动社区发现算法之前,为了简化矩阵,可以把去掉那些孤立节点或者只有很少链接与它相连的节点,在矩阵里面也就是那些对应的行和列上“1”的个数都比较少的论文,可以把这些行和列都去掉。在这里,把行和列上“1”的个数总和不超过两个的都删除。

在集合 G 上,基于 HITS 算法来计算每一篇论文的 hubs 和 authorities 值。Kleinberg 认为 hubs 和 authorities 是相互增强的关系。一个好的 hub 页指向许多好的 authorities,同时,一个好的 authority 页也有多个好的 hubs 指向它[11]。HITS 算法应用于论文网络里相互增强关系体现在引用和被引用关系上,一篇好的 hub 论文引用了很多好的 authorities 论文,同样,好的 authorities 论文也有好的 hub 论文指向它。因此,利用 HITS 的叠代算法来实现论文社区的发现以及 hub 和 authorities 值的计算,可以得到社区的中心点和权威点。

具体算法如下[11]:

Iterate (G,k)

G:一个包含 n 篇有链接关系的论文集合

k:自然数,叠代计算的次数

$z:=(1,1,1,\cdots,1)\in R_n$

$x_0:=z$

$y_0:=z$

For $i=1,2,\cdots,k$

$$x_i^p \Leftarrow \sum_{q:(q,p)\in E} {y_{i-1}}^p$$

$$y_i^p \Leftarrow \sum_{q:(q,p)\in E} {x_{i-1}}^p$$

规范化 x_i

规范化 y_i

End

Return(x_k,y_k)

通过该叠代算法可以获得论文社区的中心论文和权威论文,与这两个点紧密相连的就是所计算的科研论文社区,实际上,可以根据上述的引证矩阵和被引证矩阵的主特征向量得到论文社区。在社区内可以根据中心值和权威值来对论文进行简单排序。

5 科研论文自动评价方法

对一篇待评价的科研论文 A,自动评价方法如下:利用搜索引擎对论文主题关键字进行搜索形成论文社区,如果该论文 A 在社区内部,且有较高的权威值(authorities)或者中心值(hub),那么这篇论文的学术价值较大,且可读性好。如果该文不在社区内,则它的参考价值不大,这时可以把社区内有较高的权威值(authorities)或者中心值(hub)的论文推荐给读者。

6 小　结

主要介绍一种自动的科研论文搜索和评价方法,利用元搜索引擎实现一个基于搜索的科技论文自动评价过程,该评价过程不仅适用于已发表的论文,还可以用于尚未发表论文的评价。

参考文献:

[1] 叶继元,朱　强.论文评价与期刊评价——兼及核心期刊的概念[J].学术界,2001(3):63－71.

[2] 陈江帆.Web of science数据库检索及其科学研究价值试析[J].情报探索,2002(3):41－42.

[3] 樊怡菁.SCIE 和 Scopus 引文功能的比较分析[J].现代情报,2006(3):80－82.

[4] 岑俏玲.学者专用型搜索引擎——Google Scholar[J].科技情报开发与经济,2005,15(22):56－57.

[5] 李明伟.免费学术数据库 GoogleScholar 浅析[J].情报探索,2005(9):78－79.

[6] 李志荣,沈利华.站在巨人的肩膀上——Google Scholar 搜索引擎的评介[J].现代情报,2005(10):205－206.

[7] Jacso P. Comparison and analysis of the citedness scores in Web of Science and Google Scholar[J]. In Proceeding of Digital Libraries: Implementing Strategies and Sharing Experiences, Lecture Notes in Computer Science, 2005,3815:360－369.

[8] Meng Weiyi, Yu Clement, Liu Kinglup. Building Efficient and Effective Metasearch Engine[J]. ACM Computing Surveys, 2002,34(1):48－84.

[9] 阳小华,刘振宇.元搜索引擎系统合成算法的约束条件[J].软件学报,2002(13):1264－1270.

[10] Leydesdorff L. Clusters and Maps of Science Journals Based on Bi－connected Graphs in the Journal Citation Reports[J]. Journal of Documentation, 2004,60(4):317－427.

[11] Kleinberg J M. Authoritative Sources in a Hyperlinked Environment[J]. Journal of the ACM,1999,46(5):604－632.

(上接第 76 页)

一步工作准备考虑更合适优化的存储方案,以使本算法不仅在算法运行时间上而且在数据存储空间上都得到优化处理。

参考文献:

[1] Han Jiawei, Kamber M. 数据挖掘概念与技术[M].北京:机械工业出版社,2001.

[2] Cheung D W, Han J, Ng V T, et al. Maintenance of discovered association rules in Large database : An incremental updating technique[C]//In Proc 12th Int Conf on data engineering. New Orleans, Louisiana: IEEE Computer Society, 1996: 106－114.

[3] Cheung D W, Lee S D, Benjamin K. A general incremental technique for maintaining discovered association rules[C]//In Proceedings of the Fifth International Conference on Database Systems for Advanced Applications. Melbourne, Australia: [s. n.], 1997: 185－194.

[4] 闫　炜,崔杜武,付长龙.基于幂集的关联规则挖掘算法研究[J].计算机工程与应用,2004(1):192－193.

[5] 徐章艳,刘美玲.Apriori 算法的三种优化方法[J].计算机工程与应用,2004(36):190－192.

[6] 杨　明,孙志挥.频繁项目集的快速增量式更新算法[J].应用科学学报,2003(4):368－372.

(上接第 79 页)

参考文献:

[1] 都志辉, 陈　渝, 刘　鹏. 网格计算[M].北京:清华大学出版社,2002.

[2] Buyya R, Murshed M, Abramson D. A Deadline and Budget Constrained Cost－Time Optimization Algorithm for Scheduling Task Farming Applications on Global Grids[C]// The 2002 International Conference on Parallel and Distributed Processing Techniques and Applications(PDPTA'02). Las Vegas: IEEE Press,2002:540－552.

[3] Foster I, Kesselman C. Globus: A Metacomputing Infrastructure Toolkit[J]. Intl J. Supercomputer Applications, 1997, 11(2):115－128.

[4] Fudenberg D, Tirole J. Game Theory[M]. Cambridge: MIT Press,1991.

[5] Sun X H, Wu M. GHS: A Performance Prediction and Task Scheduling System for Grid Computing[C]// Proc. of 2003 IEEE International Parallel and Distributed Processing Symposium(IPDPS 2003). Nice: [s. n.], 2003: 123－135.

[6] Varianl H R. Microeconomic Analysis[M]. 3rd ed. New York: W W Norton & Company, 1992: 398－401.

[7] 傅晓明,张尧学,马洪军,等. 一种基于市场模型的网络带宽分配方法[J]. 电子学报, 1999(9):127－129.

[8] Foster I, Kesselman C, Tsudik G, et al. A Security Architecture for Computational Grids[C]// Proc. 5th ACM Conference on Computer and Communications Security Conference. Chicago: IEEE Press, 1998: 83－92.

[9] 曹鸿强,肖　侬,卢锡城,等.一种基于市场机制的计算网格资源分配方法[J].计算机研究与发展,2002,39(8):913－916.

[10] Basney, Livny M. Deploying a High Throughput Computing Cluster[M]// Buyya R. High Performance Cluster Computing. Vol. 1. Chpter 5. [s. l.]: Prentice Hall PTR, 1999.