ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2013/07(06)-0494-11
doi: 10.3778/j.issn.1673-9418.1212007

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-89056056

面向科学数据的PageRank排序算法*

黎建辉[1],兰金松[1,2],沈志宏[1],滕常延[1,2],周园春[1+]
1. 中国科学院 计算机网络信息中心,北京 100190
2. 中国科学院大学,北京 100190

PageRank Algorithm for Scientific Data Ranking*

LI Jianhui[1], LAN Jinsong[1,2], SHEN Zhihong[1], TENG Changyan[1,2], ZHOU Yuanchun[1+]
1. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China
2. University of Chinese Academy of Sciences, Beijing 100190, China
+ Corresponding author: E-mail: zyc@cnic.cn

LI Jianhui, LAN Jinsong, SHEN Zhihong, et al. PageRank algorithm for scientific data ranking. Journal of Frontiers of Computer Science and Technology, 2013, 7(6)：494-504.

Abstract: With the development of scientific research, scientific data have been exploding increasingly. Faced with the challenges of big data, data retrieval service is beginning very important. However, traditional scientific data retrieval system just performs key words matching between record and query, leading to unreasonable ranking results. Focusing on this issue, this paper proposes a technique to extract link information from structured scientific data and applies PageRank to scientific data ranking. This algorithm considers the importance of scientific data during the ranking stage and can get better ranking results. The experimental results of Voovle, which is a typical scientific data retrieval system, indicate that the ranking result is more reasonable and can improve user experience.
Key words: scientific data; search engine; link extraction; PageRank

摘 要：随着科学研究的发展,科学数据资源日益激增。在海量数据的情况下,数据检索服务变得极其关键,传统的科学数据检索系统只进行关键词匹配,检索结果的排序效果很差。为此,提出了针对结构化的科学数

* The National Natural Science Foundation of China under Grant No. 91224006 (国家自然科学基金); the National Science and Technology Support Program during the Twelfth Five-Year Plan of China under Grant No. 2012BAK17B01-1 (国家“十二五”科技支撑计划); the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No. XDA06010202 (中国科学院战略性先导科技专项); the Special Project of Informatization of Chinese Academy of Sciences in the Twelfth Five-Year Plan under Grant No. XXH12504 (中国科学院“十二五”信息化专项).
Received 2012-11, Accepted 2013-01.
CNKI网络优先出版：2013-03-01, http://www.cnki.net/kcms/detail/11.5602.TP.20130301.1554.003.html.

据的链接提取技术，并基于此把PageRank链接分析应用于科学数据排序。该算法在排序阶段考虑了各个科学数据资源的重要性以获得更好的排序结果。在科学数据检索系统Voovle中的实验结果表明，结合PageRank的科学数据排序更能满足用户的需求，排序结果更加合理。

关键词：科学数据；搜索引擎；链接提取；PageRank
文献标志码：A 中图分类号：TP391

1 引言

数据是与自然资源、人力资源一样重要的战略资源，隐含着巨大的经济价值，已引起科技界和企业界的高度重视。如果能有效地组织和使用大数据，将对经济发展产生巨大的推动作用，并带来前所未有的机遇。

数据根据其来源，可以粗略地分成两大类[1]：一类来自物理世界；另一类来自人类社会。前者多半是科学实验数据或传感数据。基因组学、蛋白组学、天体物理学和脑科学都是以数据为中心的学科，这些领域的基础研究产生的数据越来越多。例如，用电子显微镜重建大脑中的突触网络，1 mm^3大脑的图像数据就超过1 PB。

面对海量的科学数据，数据的管理和检索服务变得越来越重要。提供管理和检索服务的前提条件之一就是使用统一的标准来描述和存储科学数据。为此，中国科学院在科学数据库项目中研制了“科学数据库元数据框架”，并基于此整合了化学、物理、天文、材料等62家相关研究所的538个数据库，数据资源量达148 TB（http://www.csdb.cn/prohtml/0.aboutus.introduction/pages/2015.html）。该框架采用RDF[2]（resource description framework）描述科学数据资源。RDF是一种描述资源的语言，适用于任何领域，它是W3C组织推荐的描述网络上资源元数据信息的标准。一条完整的RDF表达式包含了三种对象类型，即资源（resource）、属性（property）和描述（statement），也就是“主-谓-宾”三元组。资源就是要描述的事物，属性是资源具有的某一项特征，描述就是这个属性的取值。科学数据库以元数据框架描述资源，以XML（extensible markup language）文档的形式管理和存储数据。对科学数据资源的检索和排序也转化为对相应XML文档的检索和排序。

传统的科学数据检索系统只提供与用户查询词匹配的记录，并不能保证用户最想要的记录排在检索结果的前面，因此不能帮助用户从过载的海量数据中快速地选取真正相关的记录。如何将更相关的科学数据记录排在检索结果的前面，减少用户浏览数据的数量，帮助其快速找到需要的信息，是一项很有意义且富有挑战性的工作。

本文对科学数据资源检索中的排序方法进行了研究，将在网页文本排序领域应用成熟的链接分析技术引入到科学数据排序中。通过科学数据库内部的表之间的关系和科学数据资源的文本内容信息，来研究科学数据之间的链接关系，并且把这种链接关系结合PageRank[3]算法应用到科学数据检索结果排序中，解决了科学数据检索的排序问题。

2 相关工作

2.1 XML文档的检索与排序研究

科学数据采用RDF框架进行描述，以XML格式进行存储，因此对科学数据资源的检索最终转化为对XML文档的检索。已有很多研究人员在XML文档检索和排序的工作上取得一定进展，可以借鉴这些工作以提高科学数据的排序效果。

传统上，基于关键字的信息检索系统（information retrieval，IR）针对的是“扁平”文档，最典型的为互联网上广泛存在的HTML文档。这些文档没有内在的结构，或者网页检索系统暂时不能利用这些结构来提高检索效果和性能，这样导致即使用户只想检索这些文档的某些部分，还是不得不对整个文档进行搜索。相反，XML文档含有层次的结构语义标记，用户可以指定搜索的上下文环境，故使得搜索更加高效和准确。

由于XML在结构信息上的优势，越来越多的学

者开始关注针对XML的索引、排序和检索的研究。Botev等人[4]定义并提出了一种上下文环境敏感的排序方法,为了达到要求,又提出了一个新的检索排序框架和增强的倒排索引结构。在上下文环境敏感的检索中,用户可以指定搜索的上下文,搜索可以只针对上下文相关的内容进行,而不需要处理其他非相关的内容,这样能提高搜索的效率和准确性。

Guo等人[5]设计、实现并评测了一个XML检索系统XRANK。XRANK是第一个有效利用XML中的层次结构的XML检索系统。XRANK在对搜索结果排序时,不只是计算搜索关键词之间的直接距离,还考虑了关键词在不同的XML元素中的情况,使得排序更为高效和准确。XRANK能把针对XML的检索结果转化为超链接引用,故XRANK能够检索XML和HTML混合的文档集合。

Kimelfeld等人[6]参加了INEX 2006 XML文档检索竞赛,并获得了优异的成绩。他们的检索系统包含两个重要的步骤:第一步是文档过滤;第二步是XML元素排序。两个步骤中都使用统计语言模型和HITS(hypertext-induced topic search)链接分析算法。在文档过滤步骤中,获得一个比较小的待排序文档集合。在排序步骤中,结合统计语言模型得分和HITS得分进行结果排序。需要指出的是,INEX 2006使用的是Wikipedia XML文档集合,文档中存在显式的链接,故能采用HITS链接分析算法。

综览关于XML排序的研究工作,在XML文档中没有显式链接的情况下,还没有研究者使用一些隐式链接分析方法,把链接分析应用到XML的排序方法中去。众所周知,链接分析在传统的Web检索系统中有着极其重要的作用,因此链接分析对XML文档的排序也会有非常重要的作用。

2.2 链接分析算法

链接分析最著名、成功的算法是PageRank[3]算法。PageRank是Google的两位创始人Brin和Page于1998年提出的。这个算法的基本思想是一个网页被多次引用,则它可能是很重要的;一个网页虽然没有被多次引用,但是被重要的网页引用,则它也可能是很重要的;一个网页的重要性被平均地传递到它所引用的网页上[5]。这种重要的网页被称为权威(authorities)网页,相应的其计算出来的PageRank值也高。

假定用户一开始随机地访问网页集合中的一个网页,以后跟随网页的向外链接向前浏览网页,不回退浏览,浏览下一个网页的概率就是被浏览网页的PageRank值。PageRank值的计算方法如下:

$$PageRank(v)=p+(1-p)\times\sum_{b\in B}\left(\frac{PageRank(b)}{N_b}\right) \quad (1)$$

其中,B为所有链向v的网页集;$PageRank(x)$表示网页x的PageRank值;N_b表示网页b链出的链接总数;p为跳转概率,物理意义为用户随机跳转到网页的概率。

迭代求取每个网页的PageRank值直至收敛。因为网页的集合是个有限集合,所以根据马尔科夫平衡原理,最后经过有限次迭代,PageRank值一定会收敛。最后收敛的值就是各个网页的PageRank值。

因为PageRank值与查询词的主题无关,且只与网页的链接有关,所以使用PageRank算法会出现“主题漂移”,新网页PageRank值偏低及其偏重综合性网页等问题[7]。

为了解决PageRank存在的不足,研究者又提出了各种基于PageRank的改进算法,其中一个是Google工程师Bharat提出的Hilltop算法[8]。Hilltop的指导思想和PageRank是一致的,都是通过网页被链接的数量和质量来确定搜索结果的排序权重。但是Hilltop认为只计算来自具有相同主题的相关文档链接对于搜索者的价值会更大,即主题相关网页之间的链接对于权重计算的贡献比主题不相关的链接价值要更高。

Hilltop算法的过程如下:首先计算与查询主题最相关的“专家”网页集;其次在选中的“专家”集中找出相关网页的链接,并根据这些链接找到目标网页;最后根据指向目标网页的“专家”数量和相关性,对目标网页进行排序。可见,目标网页的得分反映了中立的专家们的集体观点。Hilltop算法在无法得

到足够的专家网页子集时（小于2个专家网页），返回为空。Hilltop适合于对已有的查询排序进行更精细的排序。这就意味着Hilltop算法适合与其他页面排序算法结合，用于提高精度，而不适合作为一个独立的页面排序算法。

Hilltop另外一个特性是在对两个具有同样主题，而且PageRank值相近的网页排序过程中显得非常重要，并且能避免通过增加许多无效链接来提高网页PageRank值的作弊方法。

另外一个非常重要的链接分析算法是HITS算法[9]。HITS算法是由康奈尔大学的Kleinberg博士于1998年首先提出的。该算法把网页分为Hub和Authority两类。Hub通过超链接指向其他网页，Authority则是被指向的网页。通常它们之间存在一个相互加强的关系，即好的Hub指向多个好的Authority，好的Authority则被多个好的Hub所指向。

HITS的主要思想是利用网页的链出和链入为每一个网页计算两个值，即Authority值和Hub值，指向别的网页定义为Hub值，被指向定义为Authority值。结果按照Authority值进行排序。HITS算法中Hub值的计算方法如式（2），Authority值的计算方法如式（3）。在式（2）和式（3）中 B 是链接到 v 的网页集合，N_b 是网页 b 的链出链接的数量。

$$Hub(v)=\sum_{b\in B}\left(\frac{Authority(b)}{N_b}\right) \tag{2}$$

$$Authority(v)=\sum_{b\in B}\left(\frac{Hub(b)}{N_b}\right) \tag{3}$$

与PageRank值的计算一样，HITS算法中的Hub值和Authority值的计算也是通过迭代完成的。收敛时的值即网页最终的Hub值和Authority值。排序算法根据计算出的Authority值进行排序。

2.3 三种链接分析算法比较

首先介绍的是PageRank算法，它能离线地计算出各个页面的等级，是非查询相关的排序算法。用户通过输入关键词进行检索，根据关键词匹配得到结果集，并按照PageRank值进行排序。PageRank算法的最大问题在于倾向于老网站和综合性网站，因为老网站被引用的可能性要远远大于新网站，而综合性网站能比专业性网站获得更多的链接，但是一般来说专业性网站更具针对性，是用户更想得到的检索结果。同时PageRank对所有的链接一视同仁，没有区分主题相关与否，造成了一些网站通过增加不相关链接来恶意提高PageRank的现象。

HITS算法是查询相关的排序算法，是在查询执行时进行排序的，而不像PageRank算法是根据线下计算好的PageRank值进行排序。这样可能会对检索速度产生影响，减慢用户响应时间。同时它完全忽略了网页的内容，仅仅考虑网页之间的链接关系进行排序，如果集合中出现了一些主题不相关的网页，并且这些网页之间有比较多的链接关系，那么就容易发生Authority集中在这些链接稠密的非相关的网页中，也就是主题漂移现象。

Hilltop算法也是Google正在使用的排序算法。Hilltop算法虽然解决了通过增加许多无效链接来提高网页PageRank值的作弊方法，但是对专家页面的选择上同样存在着如何保证每次选取的专家页面都是高质量的网页的问题。在Hilltop的原始模型中，专家页面只占到整个页面的1.79%，忽略了除专家页面的大部分其他普通网页的影响，不能全面反映整个互联网的情况。同时Hilltop中根据查询主题从专家页面集合中选取与主题相关的子集也是在线运行的，同样也会影响查询响应时间。

3 科学数据之间的链接发现与生成

3.1 科学数据通用表示及其链接定义

本文采用一套元数据模型来对科学数据进行表示和存储。元数据是关于数据的数据，在数据共享过程中元数据具有数据发现、数据获取、数据管理与交换等功能。元数据为各种形态的数字化信息单元和资源集合提供规范、描述方法和检索工具，为分布的、由多种数字化资源有机构成的信息体系提供整合的工具与纽带[10]，为各种信息的集成提供支持，为集中检索提供保障。当前国内外数据共享平台多以元数据为核心[11]。本文提出的元数据模型包含通用和定制的两部分数据元素。通用的元数据元素如表1所示。

Table 1 An example of common metadata elements

表1 通用元数据元素示例

术语名称	含义	术语名称	含义
Title	标题	Creator	创建者
Keyword	关键词	Classification	数据分类
Abstract	摘要	Type	类型
Created	创建日期	Relation	关联
url	统一资源定位标示符	Organization	单位
Source	数据来源	Contact	联系人信息

表2和表3是采用通用元数据元素表示的科学数据的两个实例。这两组科学数据都来自于南京土壤所数据库。表2是关于农田环境资源的科学数据，表3是关于省市关系的数据。

Table 2 The resource of farmland environment in soil database

表2 土壤数据库中的农田环境资源

名称	描述
标题	258
联系人	周静
关键词	长期生态学 ;农田土壤环境 ;鹰潭
地点名称	江西
数据子库名称	农田土壤环境
地点代码	JX
支撑项目描述	选择主要农田生态系统类型和主要土壤类型，统一监测当地典型的农田管理模式下的土壤环境质量现状
东经	116°55'
结束年份	2005
数据集名称	农田土壤环境数据库
Meta Id	258
Dataset Id	cn.csdb.soil.agroenvironment
支撑项目名称	中国生态系统研究网络监测项目
数据名称	农田土壤矿质元素，农田土壤硝态氮与铵态氮，农田土壤微量元素和重金属元素
数据提供组织	南京土壤所鹰潭站
起始年份	2005
专题名称	农田土壤环境
路径	showItem.vpage?id=cn.csdb.soil.agroenvironment.province/8
支撑项目来源	院重大项目
北纬	28°15'
Url Copy	江西
数据类型	Table

Table 3 The resource of cities and provinces in soil database

表3 土壤数据库中省市资源

名称	描述
标题	江西
地点代码	JX
省市名	江西

不同于网页之间的链接关系，科学数据资源之间的链接是一种隐式的链接关系，需要对它进行挖掘才能获取。

具体而言，科学数据之间的隐式链接关系分为两种：

⑴ 资源的关系型引用链接。由于原始的科学数据是一种关系型结构化数据，它们之间存在主外键的引用关系，这种引用关系就是一种链接关系。原始的数据经过收割、处理后变成由XML格式表示的RDF文档，但是这种链接关系还是保持着，并能够分析出来。

在原始的南京土壤所的数据库表中，表2和表3分别存储在数据库表agroenvironment表以及provin表中，而这两个数据库表之间存在着主外键关系，这种关系如图1所示。

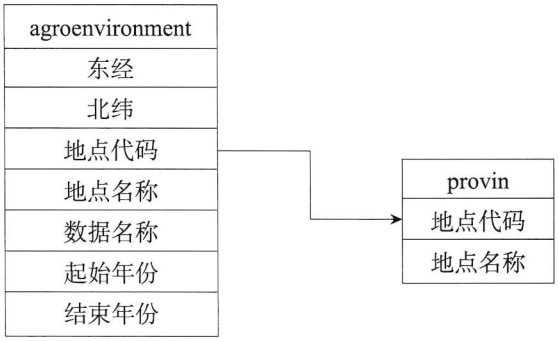


Fig.1 The relationship of provin and agroenvironment tables

图1 数据库表provin和agroenvironment之间的关系

其中，provin表的主键是agroenvironment表的外键，因此数据库表agroenvironment对应的科学数据到数据库表provin对应的科学数据存在链接关系。

⑵ 科学数据资源属性的文本引用链接关系。文本引用链接关系分析在互联网领域被大量使用，

其召回率很高，通过使用上下文相关信息，可以控制准确率的范围。这种链接关系体现在描述科学数据资源的最主要信息的标题摘要之间的互相引用关系。最直接的形式为一个科学数据的资源标题为另外一个数据的资源标题的字串。本文定义第二个科学数据对第一个科学数据存在链接关系。

图2所示为两个科学数据资源的标题，第一个资源的标题为“白桦”，第二个资源的标题为“栓皮白桦”。因为第二个资源的标题包含第一个资源的标题，所以第二个资源存在对第一个资源的链接关系。

Fig.2 The example of text reference of link relationship

图2 文本引用链接关系示例表

3.2 链接发现算法

基于文本引用分析链接的方法的前提是科学数据资源的标题能够标识这条资源，能够概括这条资源的主要内容，因此要求科学数据资源的标题应该是有意义的中文或者英文。如果科学数据资源的标题是数字或者是一些没有意义的英文字符，那么就不能够通过文本分析的方法得到链向它的其他的科学数据资源。通过统计实际的科学数据发现，无意义的标题绝大部分仅含数字或英文字母。为了方便计算，定义无意义标题为只包含数字和英文字母的标题。如果标题为无意义标题，则采用一定的策略方法自动修改其合理标题。针对不同研究所的科学数据来源，可以为其制定一个修改标题的规则。图3为中科院沈阳应用生态研究所的科学数据资源。

图3对应的修改标题规则为 选取摘要属性中的中文名字段。此处即为“早花忍冬”。现在参与建库单位有37家，针对每一家建库单位的数据都制定好其修正标题的规则。自动校正标题的算法如下。

算法1 自动校正标题的算法

1. CorrectTitle(data, ruleset) //校正标题

//data为科学数据元数据，ruleset为所有建库单位的校

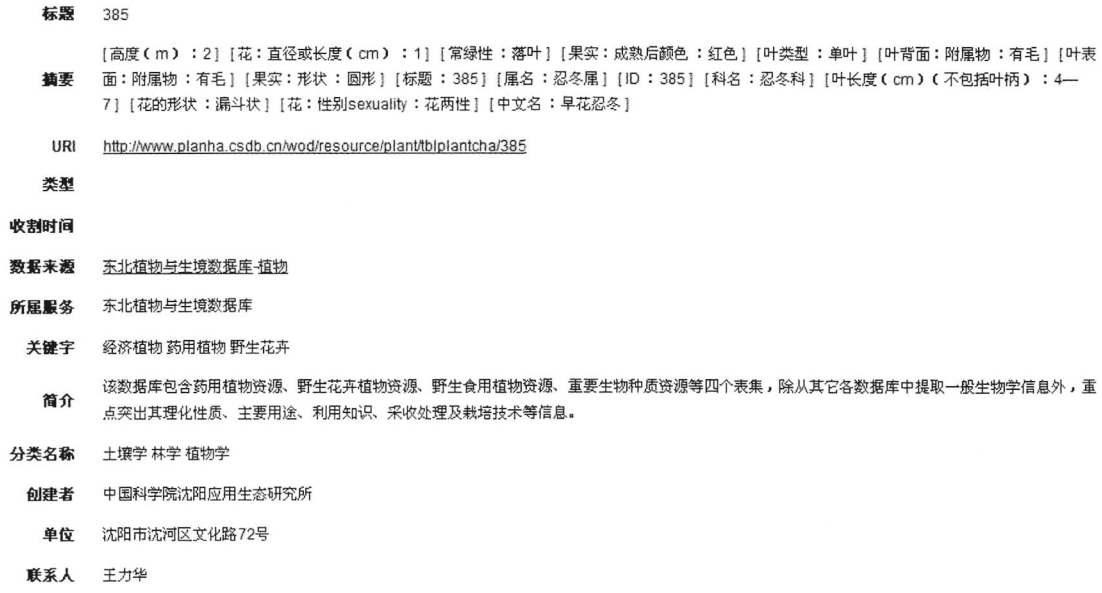


Fig.3 A scientific data resource of the Institute of Applied Ecology, Chinese Academy of Sciences

图3 中科院沈阳应用生态研究所的科学数据资源

正标题的规则集

2. If IsMeaningless(data.Title) //该条数据的标题没有意义,需要修正

3. rule=SelectRule(ruleset, data) //根据数据的建库单位选择相应的校正规则

4. data.Title=GetNewTitle(rule,data.Title) //根据规则获取新的标题,并校正

5. Return

6. IsMeaningless(title) //判断给定的标题是否无意义

7. If title只包含数字或字母

8. Return true

9. Return false

本文用计算机科学中的图论来描述科学数据资源之间的链接关系。设有向图 $G=<V,E>$,V表示图的节点集合,即所有的科学数据资源,E表示图的边集合。边 $<u,v>$ 表示资源u对v存在链接关系。分别对两种链接关系进行挖掘发现,可以得到科学数据资源的链接关系图。在此基础上,使用PageRank链接分析算法获得各个资源的PageRank值,结合传统排序的tf-idf[12]方法,可以对科学数据进行更准确、有效的排序。链接关系发现算法如下所示。

算法2 科学数据链接发现算法

//初始化算法所需参数

RuleSet //各个建库单位的资源标题校正规则

$G=<V,E>$ //使用图模型表示科学数据资源

Set $V=$ 所有的科学数据,E为空集

1. LinkDiscover(G) //链接发现例程, $G=<V,E>$

2. For u in V

3. For v in V

4. If $u\neq v$ and $<u,v>$ 不在E中 //如果 $u\neq v$ 且 $<u,v>$ 链接不在E中

5. If KeyAnalysis(u,v) or TextAnalysis(u,v) //主外关键字段和标题文本分析

6. 把 $<u,v>$ 加入E中

7. Return

8. KeyAnalysis (u,v) //主键-外键字段分析,u、v都为XML表示的科学数据资源

9. 获取u的所有外键字段保存至foreign_key_set

10. For foreign_key in foreign_key_set

11. If foreign_key$=v$ 的主键字段

12. Return true

13. Return false

14. TextAnalysis(u,v) //科学数据资源标题文本分析,u、v都为XML表示的科学数据资源

15. CorrectTitle(u, RuleSet)

16. CorrectTitle(v, RuleSet)

17. If u.Title包含v.Title

18. Return true

19. Return false

4 实验

4.1 实验背景介绍

本文对科学数据排序算法的评价是在科学数据搜索引擎上进行的。“科学数据搜索引擎Voovle”[13]是由中国科学院计算机网络信息中心科学数据中心基于收割上来的科学数据资源描述信息,进行语义转化,入库索引,并面向用户提供科学数据检索的服务。目前,科学数据搜索引擎已经推出2.0版本,共收录了37个库,5 646 706条资源描述信息,9 000多万条RDF三元组信息。Voovle的具体工作流程如图4所示。

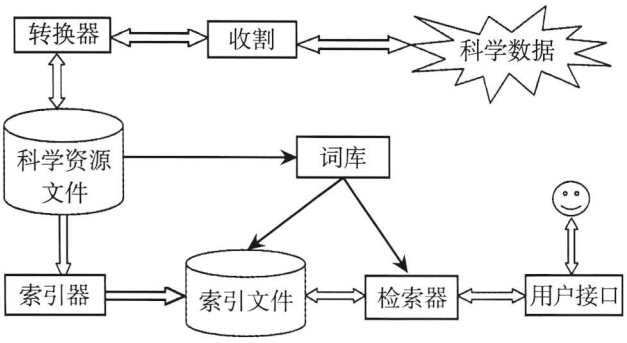


Fig.4 The workflow of scientific data search engine

图4 科学数据搜索引擎工作流程

现有的科学数据搜索引擎Voovle已经实现了基本的检索服务,但是该搜索引擎中还没有一个完善、高效的排序算法,所有的检索结果只是简单的堆积,用户最关心的数据没有放在用户容易得到结果的前几页,造成了虽然结果集中包含了用户想要的数据,

但是因为排序算法不完善，把用户最想得到的数据排在了比较靠后的位置，降低了用户体验效果，也大大降低了系统的可用性、准确性和高效性。可以结合科学数据的特点，运用现在市场上成熟的排序算法，对科学数据的检索结果进行排序，优化检索结果，提高用户体验。

另外一方面，科学数据资源有着自己独有的特点，比如结构性强，内容可信度高。但是资源之间的关联关系并不像网页资源之间的关系那样通过超链接进行关联，因此需要从科学数据资源的最初来源，即各合作建库单位的数据库入手，根据各个数据库表之间的关系以及对数据资源进行文本分析的方法，发现科学数据之间的链接关系，最后把这种关联关系与这些排序模型结合起来，共同作用，改进检索结果。

4.2 实验设置

本文主要采用的实验数据包含东北植物与生境数据库、中国土壤数据库、武汉植物与引种数据库、西双版纳热带植物数据库、青海湖联合科研基地基础数据库等5个数据库30个数据集。这些科学数据共计6.1 GB，用XML文档存储。实验平台是科学数据检索引擎Voovle2.0。通过检索不同的关键词，对不同算法的排序结果进行评测。为了使评测更准确客观，评测人员包括建库单位的工作人员、建库单位数据发布平台的开发人员、普通的科研工作者。采用了多种检索系统最常用的评价体系，包括precision@N（P@N），average precision@N（AP@N），NDCG（normalize discount cumulative gain）[14]。

4.3 实验结果

经过链接分析后，每个XML文档，即每个科学数据资源都有一个PageRank值，这个值与查询无关。但是针对每个检索的关键词，XML文档计算出的tf-idf值与查询相关。本文采用线性插值来衡量最后一个XML文档对一个查询的得分，如式（4）：

$$score(q,x)=\alpha\times\sum_{i=1}^{N}tfIdf(q_i,x)+(1-\alpha)\times pageRank(x) \quad (4)$$

其中，$score(q,x)$指XML文档x对查询q的总得分，用以确定文档x在q的查询结果中排序所在的位置；N是查询q的关键词个数；$tfIdf(q_i,x)$是文档x在关键词q_i上的tf-idf值；$pageRank(x)$是文档x的PageRank值；α是一个超参数，其取值范围为0到1之间，它对文档x的PageRank值和tf-idf值进行加权。α越大，表明查询的tf-idf值在排序中起更重要的作用，而文档的PageRank值作用更少。反之亦然。两个极端的情况分别是当α为0和1。当α为0时，查询的排序只与文档的PageRank值有关；当α为1时，查询的排序只与查询词的tf-idf值有关。

本文采用不同的α值来验证实验结果。不同的α值有不同的评测结果。下面给出在3个评测体系下相对α取值的最好的结果。每次实验随机选取n（本文中$n=4$）个用户查询，获得各个查询的性能指标后进行归一化，然后取n个结果的平均值，作为该性能指标的实验测量值。图5~图7分别表示了新的排序

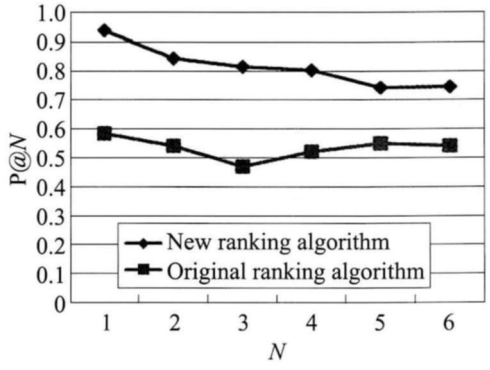


Fig.5 Precision@N by changing the parameter N

图5 N变化下的P@N值

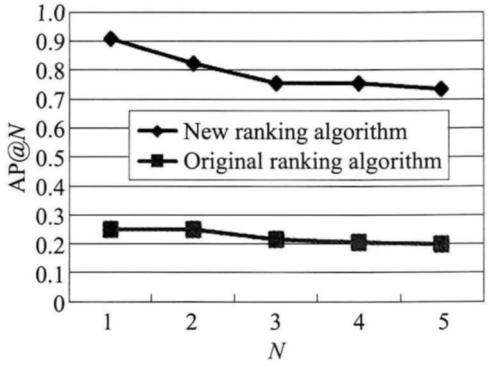


Fig.6 Average precision@N by changing the parameter N

图6 N变化下的AP@N值

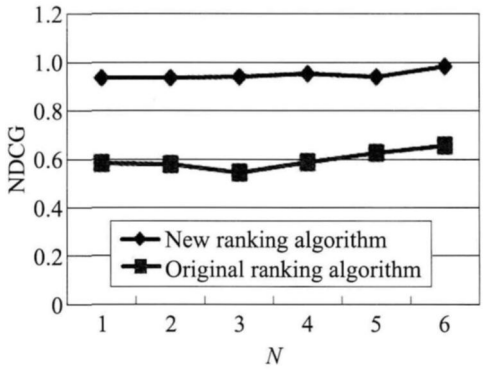


Fig.7 NDCG by changing the parameter N

图7 N 变化下的NDCG值

算法和旧的排序算法在P@N、AP@N和NDCG排序评价指标上的结果。

4.4 实验结果分析

通过加入链接分析，并结合链接分析的PageRank值，以XML文档表述的科学数据检索排序结果质量在3个常用的评测体系下均得到明显提升。3个评测体系的评测值均能反映检索系统用户对检索排序结果的满意度。评测值的提升，体现了用户体验满意度的提升。

通过实验，证实了在网页搜索引擎中起到关键作用的链接分析，通过一些技术手段转化，同样也可以将其成功地应用到类似XML的结构化数据的检索系统中。

5 结论和展望

本文针对科学数据资源之间的关联关系无法通过传统的网页超链接分析得到这一问题，提出了XML文档之间链接关系的发现策略，即联合应用XML文档的原始数据之间的主外键关系和XML文档的文本分析策略，为在XML文档集合上进行链接分析打下基础。

本文结合传统检索系统的tf-idf值排序方法，并引入PageRank值，通过实验证明了链接分析的应用是非常成功的。科学数据搜索引擎Voovle利用新的搜索排序算法，在一定程度上解决了搜索结果的排序问题，提高了用户的体验满意度。

科学数据搜索引擎是科研工作者进行科学数据检索的平台与工具，本文对科学数据检索结果的排序进行了研究，但目前还处于原型实验阶段。如何针对不同领域、不同形态的科学数据，提供不同粒度、不同层次的搜索功能，具有较大的挑战。首先，科学数据具有强烈的学科差异性，无法采用统一的描述模型，但又需要提供统一的查询接口，因此需要研究一种变粒度的科学数据资源描述框架及多层次的检索服务架构。其次，需要研究规模化的、存储格式各异的科学数据资源的元数据提取与统一存取技术，采取可伸缩的、大容量的存储架构，以满足海量数据的存储与高效查询的需求。最后，还需要综合考虑科学数据的质量、学科的差异性，以及搜索用户的使用行为，基于此增强数据搜索及关联推荐的效果。

References:

[1] Li Guojie. The scientific value of big data research[J]. Communications of the China Computer Federation, 2012, 8(9): 8-15.

[2] Abbas M A, Ahmad W F W, Kalid K S. Resource description framework based intelligent tutoring system[C]//Proceedings of the 11th IEEE/ACIS International Conference on Computer and Information Science (ICIS '12), Kuala Lumpeu, 2012. Washington, DC, USA: IEEE Computer Society, 2012: 324-328.

[3] Francheschet M. PageRank: standing on the shoulders of giants[M]. New York, NY, USA: ACM, 2012: 92-101.

[4] Botev C, Shanmugasundaram J. Context sensitive keyword search and ranking for XML[C]//Proceedings of the International Workshop on Web and Databases (WebDB '05), Baltimore, Maryland, USA, 2005: 115-120.

[5] Guo Lin, Shao Feng, Botev C, et al. XRANK: ranked keyword search over XML documents[C]//Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data (SIGMOD '03), San Diego, California, USA, 2003. New York, NY, USA: ACM, 2003: 16-27.

[6] Kimelfeld B, Kovacs E, Sagiv Y, et al. Using language models and the HITS algorithm for XML retrieval[C]//LNCS 4518:

Proceedings of the 5th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX '06), Dagstuhl Castle, Germany, 2006. Berlin: Springer, 2006: 253-260.

[7] Jiang Yonghui, Wu Hongli. New PageRank optimization algorithm[J]. Computer Engineering and Applications, 2012, 48(6): 94-104.

[8] Mo Yunfeng. A study on tactics for corporate website development aiming at search engine optimization[C]//Proceedings of the 2010 2nd International Workshop on Education Technology and Computer Science (ETCS '10), Wuhan, China, 2010: 673-675.

[9] Awekar A C, Mitra P, Kang J. Selective hypertext induced topic search[C]//Proceedings of the 15th International Conference on World Wide Web (WWW '06), Southampton, UK, 2006. New York, NY, USA: ACM, 2006: 1023-1034.

[10] Fang Yuan, Du Zhuping, Zhou Gongye. New kind of metadata management strategy based on object storage[J]. Computer Engineering, 2012, 28(3): 25-27.

[11] Han Chunhua, Liang Jianfeng, Zhang Junming, et al. Design of ocean data management and sharing platform[J]. Computer and Modernization, 2012(7): 218-221.

[12] Ramos J. Using tf-idf to determine word relevance in document queries[C]//Proceedings of the 1st Informational Conference on Machine Learning (iCML '03), Piscataway, NJ, USA, 2003: 56-60.

[13] Li Chengzan, Shen Zhihong, Li Jianhui. Voovle: a scientific data-oriented search engine[J]. e-Science Technology & Application, 2011, 2(5): 36-43.

[14] Pan Yan, Luo Haixia, Tang Yong, et al. Learning to rank with document ranks and scores[J]. Knowledge-Based Systems, 2011, 24(4): 478-483.

附中文参考文献：

[1] **李国杰. 大数据研究的科学价值**[J]. **中国计算机学会通讯**, 2012, 8(9): 8-15.

[7] **蒋永辉，吴洪丽. 新的**PageRank**优化算法**[J]. **计算机工程与应用**, 2012, 48(6): 94-104.

[10] **方圆, 杜祝平, 周功业. 基于对象存储的新型元数据管理策略**[J]. **计算机工程**, 2012, 28(3): 25-27.

[11] **韩春花, 梁建峰, 张俊明, 等. 大洋数据管理与平台设计**[J]. **计算机与现代化**, 2012(7): 218-221.

[13] **李成赞, 沈志宏, 黎建辉. 面向科学数据的搜索引擎**Voovle[J]. **科研信息化与应用**, 2011, 2(5): 36-43.

LI Jianhui was born in 1973. He received his Ph.D. degree from Institute of Computing Technology, Chinese Academy of Sciences in 2007. Now he is a professor and Ph.D. supervisor at Computer Network Information Center, Chinese Academy of Sciences. His research interests include semantic-based data integration, large-scale distributed databases process and analysis, data intensive computing and scientific application, etc.

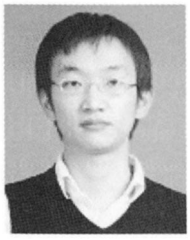
黎建辉（1973—），男，湖北通城人，2007 **年于中国科学院计算技术研究所获得博士学位，现为中国科学院计算机网络信息中心科学数据中心主任、研究员、博士生导师，主要研究领域为基于语义的数据集成，大规模分布式数据处理和分析，数据密集型计算和科学应用等。发表学术论文**40**余篇，主持国家自然科学基金、中国科学院创新方向、中国科学院信息化专项等**10**多项课题。**

LAN Jinsong was born in 1988. He is a master candidate at Computer Network Information Center, Chinese Academy of Sciences. His research interests include data mining and information retrieval.

兰金松（1988—），男，湖南茶陵人，中国科学院计算机网络信息中心硕士研究生，主要研究领域为数据挖掘，信息检索。

SHEN Zhihong was born in 1977. He received his Ph.D. degree from Chinese Academy of Sciences in 2012. Now he is a senior engineer at Computer Network Information Center, Chinese Academy of Sciences. His research interest is the management, organization and semantic integration of scientific data.

沈志宏(1977—),男,安徽东至人,2012年于中国科学院文献情报中心获得博士学位,现为中国科学院计算机网络信息中心高级工程师,主要研究领域为科学数据的管理、组织与语义化集成。



TENG Changyan was born in 1988. He received his M.S. degree from Computer Network Information Center, Chinese Academy of Sciences in 2012. His research interests include information retrieval and semantic Web.

滕常延(1988—),男,安徽亳州人,2012年于中国科学院计算机网络信息中心获得硕士学位,主要研究领域为信息检索,语义网。

ZHOU Yuanchun was born in 1975. He received his Ph.D. degree from Institute of Computing Technology, Chinese Academy of Sciences in 2006. Now he is an associate professor at Computer Network Information Center, Chinese Academy of Sciences, and the member of CCF. His research interests include data mining and data intensive computing.

周园春(1975—),男,江西余江人,2006年于中国科学院计算技术研究所获得博士学位,现为中国科学院计算机网络信息中心副研究员,CCF会员,主要研究领域为数据挖掘,数据密集型计算。发表论文60多篇,主持国家自然科学基金、中国科学院创新方向、中国科学院信息化专项、国家基础条件平台等多项课题。