CrossRef 文本和数据挖掘服务
——《浙江大学学报(英文版)》的实践

■张欣欣 缪弈洲 张月红

收稿日期:2015-04-09
修回日期:2015-05-28

《浙江大学学报(英文版)》编辑部,杭州市浙大路 38 号 310027

摘 要 **【目的】**为国内出版界介绍如何使用 CrossRef 文本和数据挖掘。**【方法】**研究 CrossRef 文本和数据挖掘的流程,并分析《浙江大学学报(英文版)》的 CrossRef 文本和数据挖掘创新服务的实践。**【结果】**CrossRef 文本和数据挖掘可以为研究人员提供便捷的服务,支持科学研究。**【结论】**加入 CrossRef 文本和数据挖掘顺应国际出版业的潮流和发展,并可藉此扩大期刊的国际显示度,从多角度增强期刊的国际影响力。

关键词 CrossRef 文本和数据挖掘;CrossRef Metadata API;国际影响力;中国科技期刊国际影响力提升计划
DOI:10.11946/cjstp.201504090315

1 引言

文本和数据挖掘跨越多学科领域,并结合语言学、计算机科学和统计学技术来构建工具,可以有效地检索和提取数字化的文本信息。过去,无论是对于开放获取期刊还是基于订阅模式的期刊,研究人员进行文本和数据挖掘没有一个简单普遍的获取全文的方法。研究人员对学术内容进行数据挖掘的兴趣和需求与日俱增,这就需要对大量的文章全文进行自动地访问。研究人员发现,为获得对已经订阅内容进行数据挖掘的授权,他们需要与众多的基于订阅购买模式的出版商协商复杂的双边协议,但这显然不太实际,并常常被困于曲折的接洽和谈判中。比如加利福尼亚大学计算生物学家 MaxHaeussler,花费三年多时间与出版商争论要求获得许可以便从 300 万文章中抽取 DNA 数据为人类基因在线地图做注释。出版商也认为与大量的研究人员和众多的研究机构基于复杂的双边协议进行授权谈判,同样很难实现[1-3]。

CrossRef 公司于 2014 年 5 月启用的 CrossRef 文本和数据挖掘服务提供了一个简单通行的方法,即可用于文本和数据挖掘的标准应用程序界面 CrossRef Metadata API(Application Program Interface)。不论出版商的商业模式如何(开放获取、订阅或者二者兼而有之),都可以使用 CrossRef Metadata API,而且对于任何研究者都是免费使用的。CrossRef 文本和数据挖掘创新服务依托于出版业,不仅满足了研究人员对文本和数据挖掘的迫切需求,支持科学研究,解决了出版商与研究人员进行双边协议的谈判问题,同时也扩大了出版商期刊的显示度[1]。

中国科技期刊近年来愈发重视学术影响力的提升与国际化发展[4]。《浙江大学学报(英文版)》一直关注全球期刊行业的创新动态,争取与国际出版标准接轨。在中国科技期刊国际影响力提升计划项目的资助下,继 2014 年成为国内首家在网站平台与论文中同时标注 CrossMark、FundRef 和 ORCID 的期刊后[5],《浙江大学学报(英文版)》继续研究和实践 CrossRef 文本和数据挖掘创新服务,并藉此扩大期刊的国际显示度,从多角度增强期刊的国际影响力。本文将着重从出版商的角度介绍如何参与 CrossRef 文本和数据挖掘,并将从研究者角度使用数据挖掘应用程序界面 CrossRef REST API[1]。

2 CrossRef 文本和数据挖掘

对于出版商而言,数据挖掘很可能存在一个增长潜力巨大的市场和快速发展的机遇。英国政府

基金项目:2013-2015 年度中国科技期刊国际影响力提升计划
第一作者简介:张欣欣(ORCID:0000-0002-2852-803X),中级,编辑,E-mail:jzus_zxx@zju.edu.cn
通讯作者:张月红(ORCID:0000-0001-8702-909X),编审,主任,E-mail:jzus@zju.edu.cn

已于 2014 年 6 月实现了对非商业目的的文本挖掘的著作权费用的免除,这使得研究者能够挖掘他们已付费订购的任何内容。欧盟等虑及计算式研究的障碍可能阻碍科学创新,也在积极推进数据挖掘。这些都为出版业的数据挖掘铺平道路,给出版业的蓬勃发展带来了新的契机[6-7]。

出版商希望研究人员可以从他们的在线平台直接访问和抓取所需内容,这样不仅访问效率更高,同时防止短时间内的大量访问对其他使用者造成影响[6-7]。CrossRef 文本和数据挖掘应运而生,并于 2014 年 5 月 28 日正式启用。发起和推动这个项目的出版商和赞助者包括 American Institute of Physics (AIP)、American Physical Society (APS)、Elsevier、HighWire Press、Springer、Taylor & Francis 和 Wiley 等众多知名出版机构和组织。CrossRef 文本和数据挖掘服务使用一个可用于文本和数据挖掘的标准应用程序界面 CrossRef Metadata API。不论何种商业模式的出版商(开放获取、订阅或者二者兼而有之)都可以使用 CrossRef API,并对研究人员免费[1]。

CrossRef 拥有多达 4000 多家的出版商会员,这些会员都使用 DOI。每个 DOI 都有对应的元数据,带有描述了不同内容片段的信息片段,比如期刊文章、图书章节或者会议论文。这些存储的元数据可以扩展并识别哪些内容片段对应的全文是可以找到的,并且此信息可以被对数据挖掘感兴趣的研究人员所使用。CrossRef Metadata API 使用 CrossRef DOI 为研究人员提供在出版商页面的全文链接。出版商有义务保证满足研究人员获取全文链接的请求并可直接批量给予其全文。开放获取期刊的出版商可以简单地将请求的内容直接传送给研究人员,而基于订阅模式的出版商需要控制访问权限。CrossRef 文本和数据挖掘的流程如图 1 所示。

除了 CrossRef,Elsevier 和 IOP Science 等也提供对学术内容进行文本和数据挖掘的服务[8],并且仍积极更新文本挖掘政策以改善研究人员的获取状况[7,9,10]。

3 《浙江大学学报(英文版)》的参与和实践

《浙江大学学报(英文版)》从 2014 年 10 月开始,尝试实践 CrossRef 文本和数据挖掘服务。首先

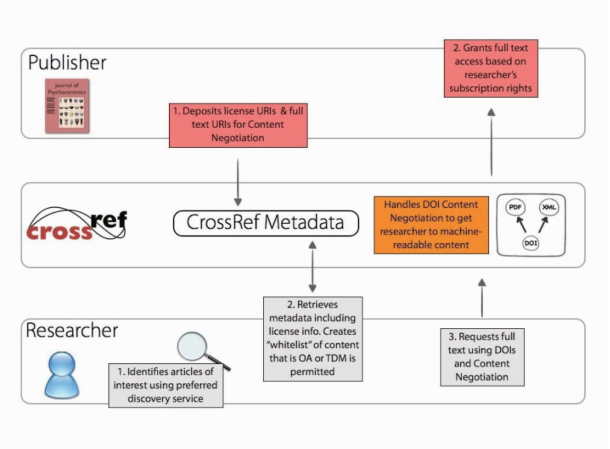

图 1 CrossRef 文本和数据挖掘流程图[1]

从出版商角度参与 CrossRef 文本和数据挖掘,包括申请参与 CrossRef 文本和数据挖掘、存储元数据、提供全文链接、明示版权信息以及提供 Click-through 服务等。并从研究者角度使用数据挖掘应用程序界面 CrossRef REST API 且成功获取所挖掘的全文。

3.1 注册

首先在 CrossRef 网站上注册,网址为 http://www.crossref.org/tdm/contact_form.html,申请参与 CrossRef 文本和数据挖掘(CrossRef Text and Data Mining Contact Form)(见图 2)。

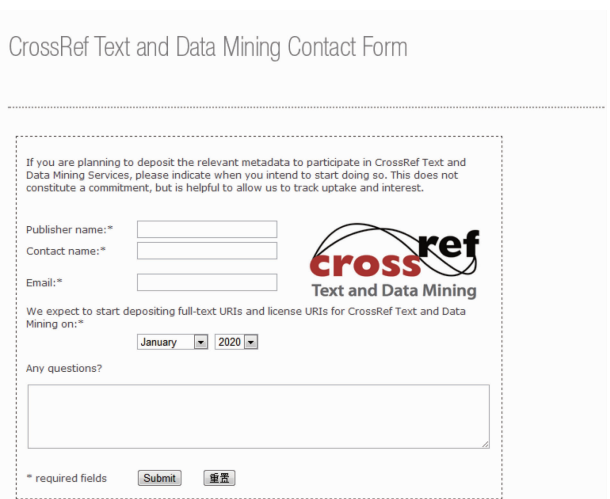

图 2 CrossRef 文本和数据挖掘注册页面[11]

3.2 元数据存储

作为出版商参与 CrossRef 文本和数据挖掘,需要做如下两件事情:(1)为每个 DOI 存储带有全文链接的元数据,使研究人员能够据此链接找到文章全文;(2)在上述的元数据中存储版权信息,方便研

究人员据此查询他们是否能够获取挖掘此内容片段的许可。存储上述信息的 xml 文件需要上传到 CrossRef 系统的 Metadata 处。

3.2.1 全文链接

根据出版商自身平台是否支持内容协商[12],存储内容的全文链接分为两种方式。绝大多数的出版商不在自身平台支持内容协商,则使用 CrossRef 提供的方法 1 (Method 1: Publisher provides specific URIs for each mime-type they support)[1]。以《浙江大学学报(英文版)》为例,提供数据的 xml 文件包含文章的基本信息(如 DOI、年、卷和页码等)、ORCID 和 FundRef 等信息,并且提供可以直接获取文章内容的全文链接(见图 3)。此 xml 文件信息高度丰富,直接体现文本和数据挖掘的真实价值[6,9,10]。

3.2.2 版权访问信息

元数据需要给文本和数据挖掘使用者一个明确的指示,告知其是否被允许使用 CrossRef DOI 所指向的内容。若研究人员不能自动得知其是否被允许访问全文,那么出版商仅为其提供全文链接是没有意义的。存储的数据中的版权信息部分需提供允许访问全文链接的时间范围。一般而言,出版商设定的允许访问时间为一年(见图 3)。开放获取期刊仅提供开放获取版权信息即可,如通用的 Creative Comments,并没有时间限制。

3.3 提供全文

出版商必须保证存储内容中所显示的链接与文章的实际链接一致。如果访问链接有所变动,必须随时更新以保证存储内容中的链接的有效性。

3.4 访问速度控制

文本和数据挖掘可能会增大网站的访问量,出版商的服务器必须能够应对和支持大流量的数据下载。出版商可以通过控制访问速度减轻网站负担,这取决于其自身情况。

3.5 附加版权条款

出版商可能要求研究人员同意一些额外的版权条款。这就必须使用 URI 指引使用者到 Click-through 服务。研究人员可以通过 Click-through 服务

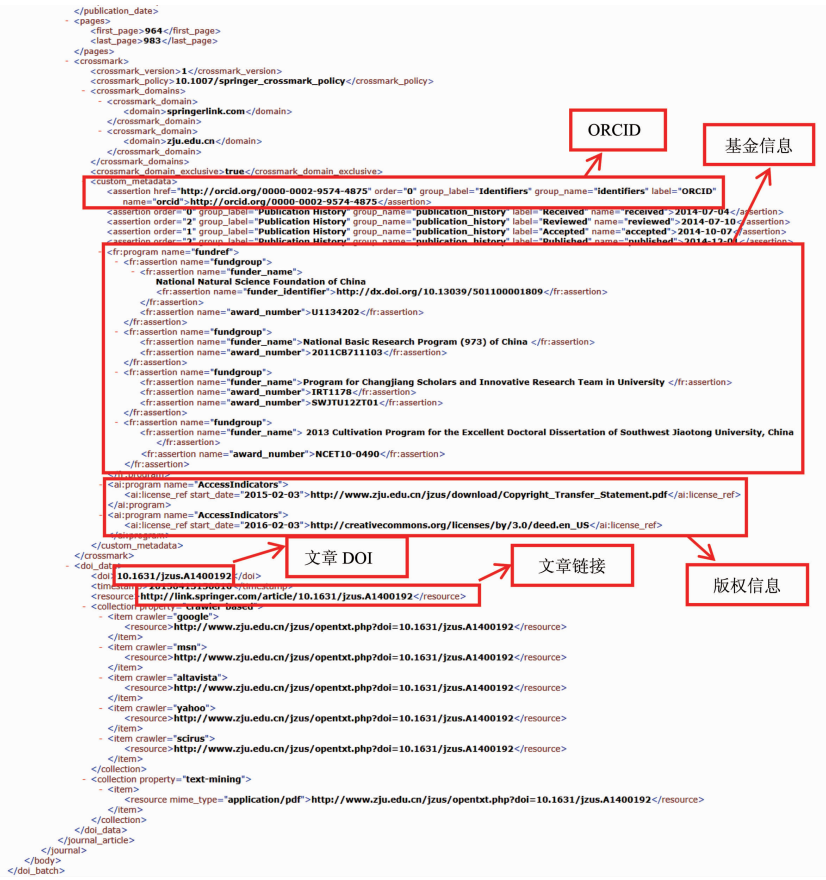

图 3 《浙江大学学报(英文版)》文本和数据挖掘的存储数据(xml)示例

阅读出版商的条款和限制条件(Terms and Conditions (T&Cs)),并判断是否接受或拒绝。出版商上传和管理 T&Cs,必须提供如下内容:(1)出版商的网站 URI;(2)出版商名称;(3)T&Cs 的简短描述;(4)T&Cs 的全文,并用 Markdown 格式显示。《浙江大学学报(英文版)》在 Click-through 中的 T&Cs 如图 4 所示。T&Cs 必须申明,文本和数据挖掘仅用于非商业目的,每次挖掘的片段内容不能超过 200 个字,且必须通过机器挖掘而非人工处理,同时遵守 CC-BY 3.0 协议等[13-15]。T&Cs 正式上线之后一旦被研究人员阅读并执行了同意或者拒绝命令,将不可修改;除非作废此 T&Cs,并提供新的版本。

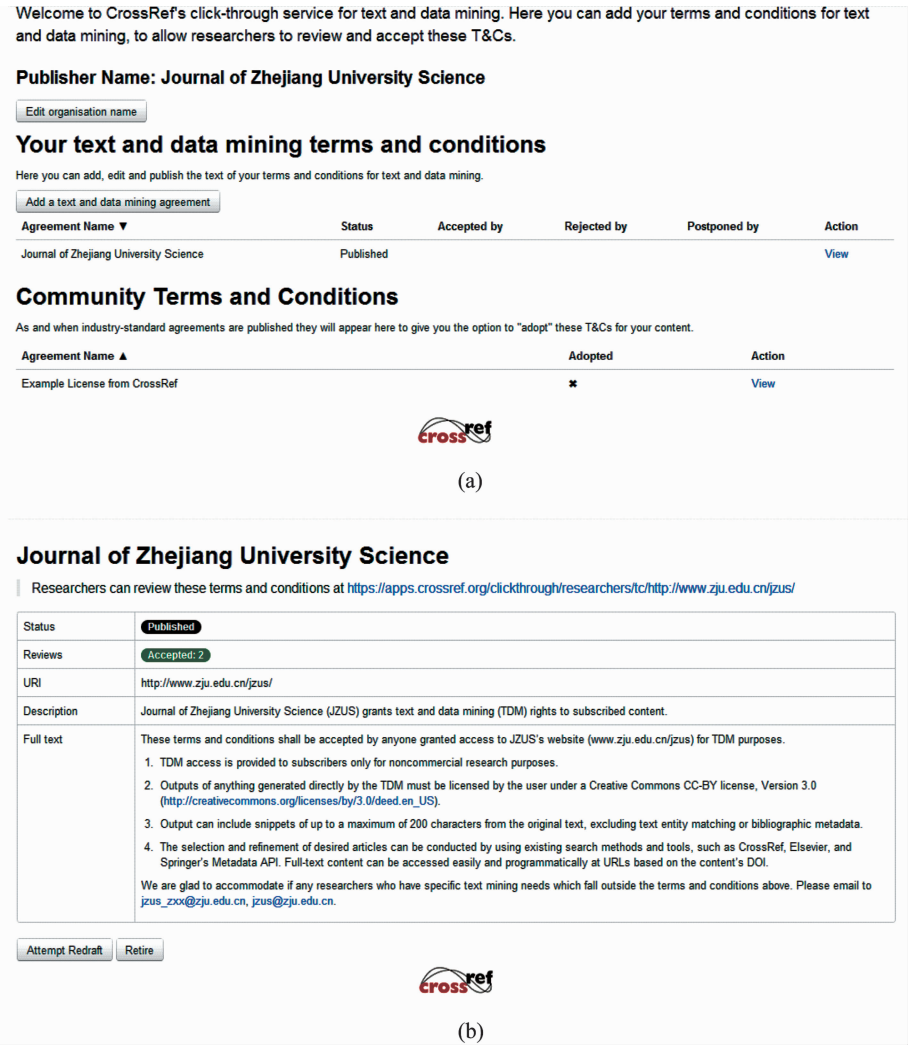

图 4 《浙江大学学报(英文版)》的 Click-through 页面(a)及其 Terms and Conditions (b)

出版商使用 CrossRef 系统的账号和密码使用 Click-through 服务,并获取其 API 验证码(Publisher API Token (PAT)),如《浙江大学学报(英文版)》的 PAT 为 e873add9-f850525e-4d233b2e-xxxxxxxx(最后八位数字隐去)。研究人员在爬取数据时发送了包含客户端 API 验证码(Client API Token (CAT))的内容(如本文作者的 CAT 为 9a9b2063-b57c021a-7fd7c9dc-xxxxxxxx(最后八位数字隐去))。出版商结合 PAT 和 CAT,可以很容易通过简单的 HTTP 请求(比如使用 Linux 系统中常见的访问网页命令 curl)来检查哪些条款被遵守,哪些没有。研究人员在发送 HTTP 请求时,在头部(header)包含 PAT,在 URI 研究人员对应的部分填写 CAT,形式如下:

curl-k-H " CR-Clickthrough-Publisher-Token: e873add9-f850525e-4d233b2e-xxxxxxxx"

" https://apps. crossref. org/clickthrough/api/licenses/9a9b2063-b57c021a-7fd7c9dc-xxxxxxxx" -D --L -O

返回信息(为 JSON 格式,故需在网页上安装 JSONView 插件)给出了研究人员所接受或拒绝的出

版商列出的条款。出版商可以通过迭代这样的结果,查看研究人员是否已签署相关内容的协议,以此判断是否同意其下载所请求的全文[16-17]。

3.6 研究人员使用 CrossRef REST API 简介

《浙江大学学报(英文版)》扮演研究人员的角色,实践了如何使用 CrossRef API 获取全文。研究人员使用 CrossRef API 的教程请参见 Geoffrey Bilder 的报告[17],使用简介请参见 https://github.com/CrossRef/rest-api-doc/blob/master/rest_api_tour.md,其参数说明请访问 https://github.com/CrossRef/rest-api-doc/blob/master/rest_api.md。

研究人员在 https://apps.crossref.org/clickthrough/researchers/#/login/处使用 ORCID 登录,在 Publisher-Specific Agreements 处可以查看、接受或拒绝各出版商已经发表的 T&Cs,并可获取 CAT。研究人员将包含接受或拒绝的条款信息的 CAT 提供给出版商,出版商即可知道该研究人员是否具有相应的许可。一旦研究人员接受或拒绝相应的 Click-through 许可,在发送 HTTP 请求并要求下载文章全文时,在 header 部分提供一个 CR-TDM-Client-Token。那些不需要 Click-through 功能或者开放获取期刊的出版商,可直接忽略这个头文件;而需要 Click-through 服务的出版商可以核对研究人员是否遵守和签订了条款。出版商使用研究人员提供的 CAT 来判断其是否已经接受了相应的条款和协议,如果研究人员接受,则将给予其全文[1-2]。

比如本文作者扮演研究人员角色通过 Click-through 服务请求获取某些特定 DOI 的文章,示例如下:

```
curl-k-H " CR-Clickthrough-Client-Token: 9a9b2063-b57c021a-7fd7c9dc-xxxxxxxx" "http://www.zju.edu.cn/jzus/opentxt.php?doi=10.1631/jzus.A1400195" -D - -L - O
curl-k-H " CR-Clickthrough-Client-Token: 9a9b2063-b57c021a-7fd7c9dc-xxxxxxxx" "http://www.zju.edu.cn/jzus/opentxt.php?doi=10.1631/jzus.A1400192" -D - -L - O
curl-k-H " CR-Clickthrough-Client-Token: 9a9b2063-b57c021a-7fd7c9dc-xxxxxxxx" "http://www.zju.edu.cn/jzus/opentxt.php?doi=10.1631/jzus.A1400263" -D - -L - O
```

表示本文作者使用的 CAT 为 9a9b2063-b57c021a-7fd7c9dc-xxxxxxxx,请求获取 DOI 为 10.1631/jzus.A1400195,10.1631/jzus.A1400192 和 10.1631/jzus.A1400263 这三篇文章的全文,并得以实现。

通过 CrossRef API 及其提供的 Click-through 服务,可以方便获取大量数据的 DOI 及其对应的全文链接。研究人员通过简单的 HTTP 等请求或语言,批量获取文章的全文。比如研究人员使用 CrossRef API,搜索在 CrossRef 元数据中,包含“血液(Blood)”这个关键词并且提供版权信息和全文链接的记录有多少条,要求如下所示:http://api.crossref.org/works?filter=has-license:true,has-full-text:true&query=blood&rows=0,由此可以获取大量记录,并可继续增加限制条件来缩小获取数据的范围从而精准地得到最符合要求的文献。如果需要从 API 的结果中获取特定的 DOI 及其全文链接,并且批量下载全文,则需要用 Python 或 Ruby 等语言编辑小程序来实现,在此不再赘述。

4 结语

文本和数据挖掘市场增长潜力巨大。CrossRef 文本和数据挖掘解决了传统数据挖掘手段存在的诸多问题,满足了研究人员对学术内容进行文本和数据挖掘的需求,方便地解决了出版商需要与大量的研究人员和众多的研究机构基于复杂的双边协议进行授权谈判的问题,并且出版商参与方式也简单便捷,必将会吸引越来越多的期刊加入此创新服务行列。《浙江大学学报(英文版)》一直关注全球期刊行业的动态并积极追求国际创新服务。在参与和使用 CrossMark、FundRef 和 ORCID 之后,在科技期刊国际影响力提升计划的资助成为国内首家实践 CrossRef 文本和数据挖掘创新服务项目的期刊。这不仅满足科研人员进行数据挖掘的迫切需求,支持科学研究,顺应国际出版业的潮流和发展;同时藉此扩大期刊的国际显示度,从国际创新技术服务等多角度提升期刊的国际影响力。

致谢:感谢浙江大学软件学院金小刚教授提供技术支持,并感谢 CrossRef 公司 Rachael 女士提供详细的咨询服务。

参考文献

[1] CrossRef. CrossRef Text and Data Mining [EB/OL]. [2015-01-21]. http://tdmsupport.crossref.org/.

[2] Lammey R. CrossRef Text and Data Mining Services. CrossRef ALPSP Annual Meeting, September, 2014, London, UK.

[3] Van Noorden R. Trouble at the text mine [J]. *Nature*, 2012,

483:134 – 135.
[4] 任胜利.《中国科技期刊国际化发展》专题序[J]. 中国科技期刊研究,2015,26(3):217 – 217.
[5] 张欣欣,张月红,缪弈洲,等. 创新与“棒”期刊——《浙江大学学报(英文版)》在科技期刊国际影响力提升计划中的思考与实践[J]. 科技与出版, 2015, 4:28 – 33.
[6] ALPSP. Member briefing text and data mining. ALPSP International Conference, 2014. London, UK.
[7] 史双青,彭乃珠. Elsevier 更新文本挖掘政策以改善研究人员的获取状况[EB/OL]. [2015 – 04 – 08]. http://www.open-access.net.cn.
[8] Elsevier. Text mining of Elsevier full-text content [EB/OL]. [2015 – 03 – 14]. http://dev.elsevier.com/text_mining.html.
[9] Chris Shillum. Elsevier updates text-mining policy to improve access for researchers [EB/OL]. [2015 – 04 – 08]. http://www.elsevier.com/connect/elsevier-updates-text-mining-policy-to-improve-access-for-researchers.
[10] Van Noorden R. Elsevier opens its papers to text-mining [EB/OL]. [2015 – 04 – 08]. Nature News, 2014. http://www.nature.com/news/elsevier-opens-its-papers-to-text-mining – 1.14659.
[11] CrossRef. CrossRef Text and Data Ming Contact Form [EB/OL]. [2014 – 10 – 08]. http://www.crossref.org/tdm/contact_form.html.
[12] 张善友. 内容协商[EB/OL]. [2015 – 3 – 6]. http://www.cnblogs.com/shanyou/archive/2012/06/12/2547019.html.
[13] Springer API. Springer's text-and data-mining policy [EB/OL]. [2015 – 02 – 13]. http://www.springer.com/gp/rights-permissions/springer-s-text-and-data-mining-policy/29056.
[14] IOPScience. Text and Data Mining (T&DM) [EB/OL]. [2015 – 02 – 13]. http://iopscience.iop.org/info/page/text-and-data-mining.
[15] Elsevier. Terms and conditions of text and data mining [EB/OL]. [2015 – 2 – 13]. http://www.elsevier.com/about/policies/content-mining-policies conditions-of-text-mining.
[16] Lammey R. CrossRef Text and Data Mining Webinar, June 3, 2014[EB/OL]. [2015 – 03 – 14]. https://www.youtube.com/watch?v=1BX6A0fshDw.
[17] Bilder G. Geoffrey Bilder's presentation from the 2014 CrossRef Workshops, 2014 [EB/OL]. [2015 – 03 – 06]. http://river-valley.zeeba.tv/text-data-mining-api-researcher-use/.

作者贡献声明:

张欣欣:材料收集、技术分析、综合整理及写作;
缪弈洲:技术分析及实现;
张月红:全面指导、论文设计和修改。

CrossRef text and data mining service: practice on *Journal of Zhejiang University Science*

ZHANG Xinxin, MIAO Yizhou, ZHANG Yuehong

Editorial Office of *Journal of Zhejiang University Science*, Hangzhou 310027, China

Abstract: [**Purpose**] This paper introduces the new innovation service CrossRef text and data mining. [**Methodology**] Supported by the Project for Enhancing International Impact of China STM Journals, *Journal of Zhejiang University Science* participated in the CrossRef text and data mining service and showed some experience. [**Findings**] CrossRef text and data mining service can be used by researchers to access to the full texts of contents identified by CrossRef DOIs across publishers' sites regardless of their business models. [**Conclusions**] CrossRef text and data mining service is very effective and is one of the new trends of the international publication industry.

Keywords: CrossRef text and data mining; CrossRef Metadata API (Application Program Interface); Innovation service; Impact of journal; Project for Enhancing International Impact of China STM Journals

(本文责编:李翠霞)