科技信息资源内容监测与分析服务平台概况*

□ 徐硕 乔晓东 朱礼军 张运良 / 中国科学技术信息研究所 北京 100038

摘要：为了方便研究者分析领域的研究动态，了解领域内研究的重要研究者和重要文献，并对科技文献和科技工作者的工作进行准确的评价，作者借助国家科技图书文献中心（NSTL）雄厚的资源优势，联合清华大学等有关优势单位，共同开发了面向西文资源的科技信息资源内容监测与分析服务平台，该平台具有专家、期刊/会议和关键词统一检索功能，具有研究者关联路径发现、主题发现等功能，并且内嵌了专家和论文排名功能。

关键词：知识服务，话题模型，关联路径，排名，全文索引
DOI：10.3772/j.issn.1673—2286.2011.11.008

科技文献知识服务已经成为科技创新过程中的重要环节。科技文献是否全面、新颖影响到科技研究的质量，可以避免一些不必要的重复劳动，节省科研投入。传统的仅提供文献服务的方式已经难以满足科研工作者的需求，而智能的科技文献知识服务能够帮助研究者分析领域的研究动态，了解特定领域内的重要研究者和重要文献，并对科技文献和科技工作者的工作进行准确的评价，有利于形成良好的科技研究氛围，促进科研的良性发展。

为了满足国内科研人员的需要，我们借助国家科技图书文献中心（NSTL）雄厚的资源优势，联合清华大学等有关优势单位，共同开发了面向西文资源（主要是二次文献，包括具有西文二次文献的中文资源）的科技信息资源内容监测与分析服务平台，该平台具有专家、期刊/会议和关键词统一检索功能，具有研究者关联路径发现、主题发现等功能，并且内嵌了专家和论文排名功能。目前已经就“新能源汽车”领域建立了原型系统，本文将详细介绍这一原型系统的系统架构以及关键技术。

1 国内外研究现状及发展趋势

1.1 国内技术现状、发展趋势

北京理工大学从20世纪90年代后期开始从事基于科技文献信息的技术监测研究工作，提出以科技信息和数据分析为基础，以数据挖掘、信息萃取、知识发现、数据可视化技术等信息前沿技术为手段，综合集成各方面专家的战略性智慧，对科学技术活动进行动态监测、测量、分析和评估，为技术管理及决策提供动态、准确的科学技术发展状态，从而把握技术发展机会，降低风险，提高效率。

中科院文献情报中心开展了基于文献的知识发现的相关理论、方法与应用研究。将基于相关文献知识发现的共词与共引理论、基于非相关文献知识发现的Swanson理论和基于全文献的文本挖掘理论统一视为进行知识发现方法和手段，将其整合于基于文献的知识发现的理论框架之下，研究构建完整的基于文献的知识发现的方法和应用研究体系。将文本挖掘用于目标研究领域的背景分析，可获得比采用共词和共引方法进行研究背景分析更为全面深入的分析结果。

清华大学对科技文献监测算法展开了大量研究，并开发了一套面向计算机领域的英文科技文献监测系统——ArnetMiner。该系统从公开的文献数据库（如DBLP、Citeseer等）爬取相关的文献数据，从Web上抽取研究者的Profile信息，然后将其集成在一起，并在此基础上根据合作关系构建学术社会网络，并进行深入

* 本研究受“十二五”国家科技支撑计划“面向外文科技知识组织体系的大规模语义计算关键技术研究”（2011BAH10B04）、中国科学技术信息研究所预研项目“科技文献深层领域主题监测及主题演化规律揭示”（YY-201129）、江苏省社会科学基金项目“数字报纸的自动标引研究”（09TQC011）和教育部人文社会科学研究项目“电子报纸内容深加工研究”（09YJC870014）资助。

挖掘，提供如权威专家/期刊/会议发现、话题检测、关联路径发现等多种服务。

万方数据有限公司凭借自己强大的资源优势，搭建了中国科技分析评价服务平台。该平台是国家十一五科技支撑计划“科技文献信息服务系统关键技术研究及应用示范”项目的成果之一，主要提供以下服务：基于海量的科技信息，对期刊论文、科研人员、科研机构、科技期刊等科研实体进行多角度的分析统计；基于科技分析评价计量指标数据库，主要提供对国内的机构、人员、期刊的评价；科研实体（如作者、机构、期刊和项目等）、科技评价、学科观察、个性化。

CNKI学术趋势是依托于CNKI中国知识资源总库中的海量文献和千万用户的使用情况提供的学术趋势分析服务。它通过关键词在过去一段时间里的“学术关注指数”，提供其所在的研究领域随着时间的变化被学术界所关注的情况，以及有哪些经典文献在影响着学术发展的潮流；另外还可以知道在相关领域不同时间段内哪些重要文献被最多的同行所研读。

1.2 国外技术现状、发展趋势

国外科技信息服务机构中，汤姆森科技信息集团（以下简称Thomson）是本行业的世界领先机构，以SCI和网络为基础，集成多种数据库系统和相关信息资源，建立了知识链接系统Web of Knowledge，为科技类信息资源的深度开发和利用提供了崭新的途径和方法。目前该公司还与世界著名的数据挖掘公司开展战略合作，其服务模式和产品战略都已从提供高质量的数据信息服务向提供深层次的知识服务跨越。其中，Thomson集团和领先的高精度检索与知识挖掘软件开发商Collexis Holdings, Inc. (OTC Bulletin Board: CLXS) 合作，计划将Collexis的Knowledge Dashboard与汤姆森科技信息集团的Web of Science结合到一起，为科研界开创定制数据挖掘解决方案，以支持学术和研发机构的知识发现能力的提高。

美国Lehigh大学开发了用于文本挖掘的分层分布动态索引（Hierarchical Distributed Dynamic Indexing，HDDI）。HDDI通过信息/特征抽取、属性子集选择、文本挖掘和机器学习技术来从文本数据中进行前沿热点探测。该算法通过跟踪随时间变化概念出现频率的变化和相互关联来探测前沿热点。

Le Minh Hoang提出了新兴研究趋势探测模型，可从在线科学文献中发现新兴研究趋势。Le Minh Hoang在其博士论文中提出从在线科技文献中挖掘新兴研究趋势的模型。他依据文本挖掘技术来抽取名词性短语作为主题词，并采用与HDDI相同的主题词聚类方法来划分主题领域，通过衡量主题领域在文献集合中表现出的受关注程度和有用程度来划分主题领域。

美国德雷克赛尔大学在2003-2005年间开发了一个探测科学文献前沿热点的信息可视化软件Citespace。开发者利用Citespace进行两个研究领域的分析：集群灭绝（1981-2004）和恐怖主义（1990-2003）。通过在可视化网络中监测的明显趋势和关键点，结合领域专家意见，并在实际应用方面进行讨论，研究表明CiteSpace能够发现在未来研究工作中将要遇到的挑战和机遇。

美国佐治亚理工大学的Alan Porter教授为主的研究团队提出了“技术挖掘”理论，采用文献计量学和文本挖掘技术对海量的科技资源（包括科技文献和专利等）进行深度文本挖掘，服务于政府和企业对科技领域的信息资源的应用需求。

英国的陈超美教授（C. Chen，后转入美国德雷克赛尔大学）将潜在语义标引（Latent Semantic Indexing）和网络寻址定位（Pathfinder Network Scaling，简称PFNETs）融入ACA中。他开发了基于PFNETs的系列可视化工具，结合网络可视化，将数据挖掘和复杂网络分析方法综合集成，实现了科技文献的知识图景挖掘。

美国AUREKA公司和M.CAM公司均已掌握了先进的文本聚类技术，但其核心技术严格保密，且其文字主要针对的是英文，中文文字聚类技术没有开发。

2 服务平台的体系结构

根据三层体系结构（数据访问层、业务逻辑层和服务层）的设计原则[1]，本系统的技术框架如图1所示。相对于传统的二层体系结构而言，三层体系结构将业务逻辑层单独进行处理，从而使得应用层与业务逻辑层可位于不同的平台上，两者之间的通信协议可由系统自行定义，这样可使业务逻辑被所有用户共享。

具体来说，数据访问层存储大量的数据信息和数据逻辑， 所有与数据有关的安全、完整性约束、数据的一致性、并发操作等都在这一层完成；业务逻辑层对应应用服务器，所有的应用系统、应用逻辑、控制

都在这一层，系统的复杂性也主要体现在这一层，该层根据需要又分为三个子层（索引系统、检索系统和排名系统），而话题模型横跨这三个子层；服务层主要指用户操作界面，它要求尽可能简单， 使用户只需少量的培训就能方便地访问信息。该服务平台可对外提供的服务包括：权威会议/期刊发现、权威专家发现、权威论文检索、任意两个研究者间的关联路径以及话题发现等。

该服务平台的科技文献资源来源于国家科技图书文献中心（NSTL）[2]。NSTL是根据国务院领导的批示于2000年6月12日组建的一个虚拟的科技文献信息服务机构，成员单位包括中国科学院文献情报中心、工程技术图书馆（中国科学技术信息研究所、机械工业信息研究院、冶金工业信息标准研究院、中国化工信息中心）、中国农业科学院图书馆、中国医学科学院图书馆。网上共建单位包括中国标准化研究院和中国计量科学研究院。

该中心的宗旨目标是根据国家科技发展需要，按照“统一采购、规范加工、联合上网、资源共享”的原则，采集、收藏和开发理、工、农、医

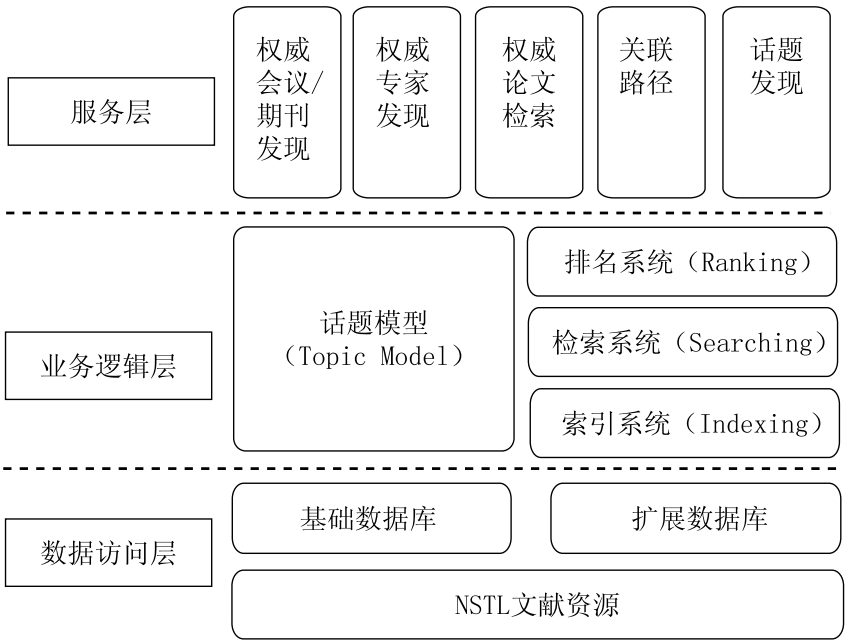

图1 服务平台的体系结构示意图

表1 NSTL西文库的数据抽取规则

"electric* vehicle*" or ”hybrid vehicle*” or ((EV or EVs or PEV or PEVs or HEV or HEVs or SHEV or SHEVs or PHEV or PHEVs or SPHEV or SPHEVs or FCEV or FCEVs or FCHEV or FCHEVs or hybrid−electric* or "hybrid electric*" or "hybrid power*" or hybrid−power*or "fuel cell*" or fuel−cell*) and (vehicle* or auto? or auto or automobile* or bus* or truck* or lorry* or autocar* or car* or motorcar* or motor? or motor or "wheeled machine*"))), **并根据选取文章的作者作二次抽取**

各学科领域的科技文献资源，面向全国开展科技文献信息服务。具体来说，NSTL收藏的科技文献资源包括中外文科技期刊、会议文献、学位论文、科技报告、专利、标准和计量规程等，拥有40多个数据库，文摘、引文、题录数据总量达1.18亿条。

3 服务平台的关键技术实现方法

3.1 NSTL数据抽取规则

为测试该平台基本运行和数据更新条件下的处理机制，从NSTL西文库中抽取数据的最初规则如表1所示，共抽取39，892条文献数据。表1中的关键词主要来源于我们已经建设好的“新能源汽车”领域的汉语科技词系统。确定服务平台运行稳定之后，我们根据选取文章所发表的期刊/会议作了进一步抽取，共抽取约850万条文献数据。

3.2 索引系统

为了加快信息的检索速度，本平台采用Lucene开源全文索引框架[3,4]。索引需要针对数据库中每个逻辑表进行建立，索引规则包括：需要建立索引的字段；是否将数据直接存储在索引文件中，以便在不访问数据库的情况下直接访问数据；是否对字段内容进行分词，一般只对描述性文字进行分词。本平台文献表的索引规则如表2所示。

3.3 话题模型

一篇文献讨论的通常不是一个话题，而是多个话题，而词汇在每

表2 文献表索引规则

索引字段	是否存储	是否分词	备注
ID	√	×	记录唯一标识
Title	√	√	正题名
Subtitle	×	√	其他题名
Abstract	√	√	原文语种文摘
Year	√	×	出版年
Keywords	√	√	关键词
Authors	√	√	作者列表
AuthorAddress	×	√	作者地址

个话题中的分布通常是不同的。因此，通常假设一个话题就是词汇上的一种概率分布，而话题模型（Topic Model）是一种产生式模型（Generative Model），它说明了一种简单的产生文档的概率过程。具体来说，为了产生一篇文献，假定为每个话题已经设置了一种概率分布，对于文献中的每个词，首先根据这个分布随机选择一个话题，然后从这个话题中随机抽取一个词汇。换句话说，文献与词汇间假设存在一个隐话题层$Z = \{z_1, z_2, \cdots, z_T\}$，这样从一篇文献d中产生一个词汇w的概率就可按下式计算：

$$P(w|d) = \sum_{z=1}^{T} P(w|z)P(z|d) \qquad (1)$$

当然，标准的统计技术也可以将这个过程倒过来，推导产生一个文献集合的话题集合。

目前话题模型比较多，从1999年Hofmann最初提出的PLSI（Probabilistic Latent Semantic Indexing）[5]，到2003年Blei等提出的LDA（Latent Dirichlet Allocation）[6]，以及LDA的各种扩展，像AT（Author-Topic）模型[7]、ACT（Author-Conference-Topic）模型[8]等。本平台采用的是ACT模型。之所以选用ACT模型，主要原因是它同时建模了作者、文献发表地点（会议、期刊等）以及话题，这与用户实际的检索体验是有关的[9]。比如用户输入“silicon surface”，他/她的意图可能不是要查找包含这两个单词的文献，更可能是要查找“silicon surface”相关的会议/期刊或者在这个领域活跃的研究人员等。

根据表1所示的规则，将NSTL西文库中抽取的文献数据分成了200个话题，图2为前三个话题的相关信息，其中包括每个话题的关键词（短语），活跃于特定话题的研究人员，以及与每个话题相关的会议/期刊。

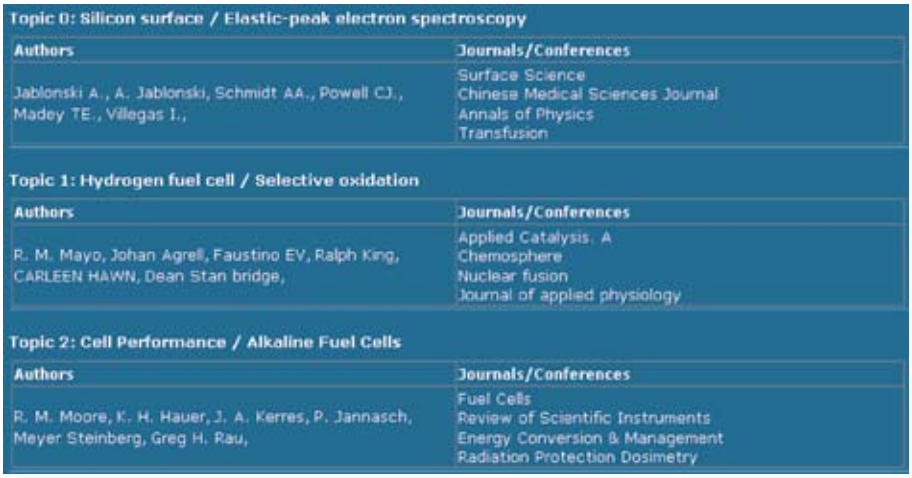

图2 “新能源汽车”领域200个话题中的前三个

为了将ACT模型应用于检索系统，还需定义文献d产生查询q的概率。但为了兼顾专指度与泛指度，通常需要将语言模型与ACT模型相结合：

$$P(q|d)=\prod_{w\in q}P(w|d)=\prod_{w\in q}P_{LM}(w|d)\times P_{ACT}(w|d) \qquad (2)$$

3.4 排名系统

本平台需要排名的对象比较多，既需要对文献进行排名，也需要对会议/期刊和作者进行排名。就像网页相互链接形成网页图一样，不同种类的对象通过它们之间的相互联系形成了对象图。在本平台中的链接有三种，例如，一篇文献对象可能被一个其他文献对象的集合引用（cited-by）；被一个作者对象的集合撰写（authored-by）；被一个会议/期刊对象出版（published-by），如图3所示。在网页图中，根据相互间的链接，不同的页面被赋予了不同的流行度。许多传统的链接分析方法，如PageRank[10]和HITS[11]等，已经被成功地应用于分析网页图的链接结构，从而区别不同网页的流行度。

相对于网页图来说，对象图中的链接是异质的，也就是说它们具有不同的语义，而传统的链接分析方法通常假定它们具有相同的“endorsement”语义，并且同等重要，如果直接应用这些方法很可能会导致不合理的流行度排名。例如，一篇文献的流行度不应该被作者的数量影响太大，被引的数量对流行度有较大影响。于是许多扩展随机行走（random walk）模型被提出，如文献[12,13]等，为了计算每个对象的流行度分数，它们同时考虑了一个对象的流行度和对象间的关系。文献[8]提出将话题模型与随机行走模型相结合，进一步改进了排名效果，这也是本平台采用的排名方法。

3.5 关联路径

给定一个学术社会网络G = (V, E)，其中v ∈ V表示社会网络中的一个研究者，$e^t_{i,j}$ ∈ E表示研究者v_i与v_j存在类型为*t*的关系（如co-author、co-project、co-college等）。研究者v_i与v_j间的关联路径定义为满足公式（3）条件的序列，其中v_i表示源研究者，v_j表示目标研究者。

$$\left\{e^t_{i_1,i_2},e^t_{i_2,i_3},\cdots,e^t_{i_{l-1},i_l}\right\},\ \text{s.t. } e_{i_m,i_{m+1}}\in E,\ m=1,2,\cdots,l-1, i=i_1, j=i_l \qquad (3)$$

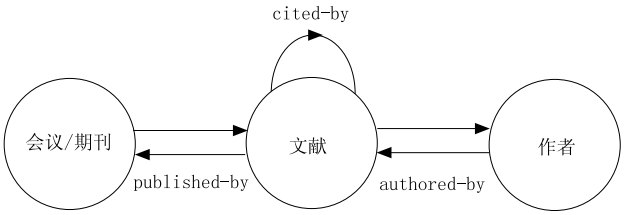

图3 科技文献对象链接关系图

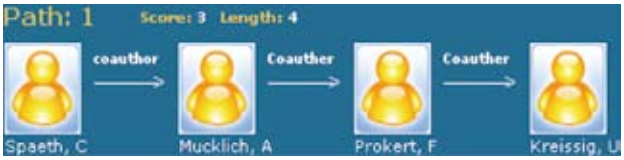

图4 研究者SpaethC与研究者KreissigU间的一条关联路径

在一个大的学术社会网络中，找到两个研究者间所有可能的关联路径是很难的。为了简化问题求解的复杂性，本平台目前只考虑了co-author关系，并且根据每个研究者的profile信息以及他/她与其他研究者间的co-author关系，为G中的每条边定义了一个权重，一条关联路径的权重定义为对应边权重之和。另外，将研究者v_i与v_j间具有最小权重的关联路径称为v_i与v_j间的最短关联路径，记最短路径的权重为L_{min}，该路径可由经典的Dijkstra算法[14]在$O(|V|^2)$时间复杂度内求解。为了给用户提供更多的洞察，本平台还可通过深度优先遍历（DFS）[14]网络G将v_i与v_j间权重低于$(1+\alpha)\times L_{min}$的关联路径输出，其中α为用户设置的阈值。图4给出研究者Spaeth, C与研究者Kreissig, U间的一条关联路径。

3.6 部分功能展示

由于版面空间限制，本文对部分功能进行了展示。图5为“Silicon surface”领域中最权威的专家、期刊/会议及论文示意图，其中最左面为该领域内最权威的专家；右上角是关于“Silicon surface”最权威的期刊/会议；而右下角是关于“Silicon surface”最权威的论文。而且所有的对象（专家、期刊/会议、论文）均利用话题模型，按与“Silicon surface”的相关程度进行了排名。图6为单个研究者的相关信息，最上面是该作者的研究兴趣，紧接着下面是该作者随时间变化的研究兴趣（时间间隔为两年），再下面是根据co-author关系绘制的该作者与其他研究人员的社会网络，最下面是

图5 “Silicon surface”领域最权威的专家、期刊/会议及论文

图6 单个研究者的相关信息示意图

该作者发表的所有文献。

需要注意的是，图5与图6中的信息是按照表1所示的规则从NSTL西文库中抽取的文献数据中得到的，约1000万条的文献数据正在处理之中。由于数据的有限性，图中的排名难免有些偏差。另外，由于NSTL中许多研究者的名字采用的是简写形式，姓名消歧变得更加困难，下一步我们打算结合可用的网络资源，自动抽取各个研究者的Profile信息，完善排名系统，使得本平台更实用。

4 总结和展望

近年来，利用各类科技文献（如期刊论文、专利文献、学位论文等）进行科技领域的分析与监测研究快速发展，且其发展方向正在从以往的基于数值分析的文献计量学方法向数值分析与内容分析相结合的分析方法转变。经调研发现，国内研究者迫切希望分析领域的研究动态，了解领域内研究的重要研究者和重要文献，并对科技文献和科技工作者的工作进行准确的评价。为满足国内研究人员的需要，我们借助国家科技图书文献中心（NSTL）雄厚的资源优势，联合清华大学等有关优势单位，共同开发了面向西文资源的科技信息资源内容监测与分析服务平台。该平台具有

专家、期刊/会议和关键词统一检索功能，具有研究者关联路径发现、主题发现等功能，并且内嵌了专家和论文排名功能。

由于NSTL的资源是多家单位共同加工完成的，在质量控制上难以做到统一。为了对科技文献资源进行深度分析，下一步我们将首先对NSTL提供的资源进行全面清洗，补全相关信息（比如目前许多研究者的名是缩写），并构建研究者间的co-project和co-college关系，这些关系可以为评审论文、项目申请书等推荐专家时解决避嫌问题提供支持，结合国家中长期发展规划，形成面向中长期科技发展纲要的重点领域及优先主题的动态科技监测能力。

作为服务于知识组织的术语符号系统，知识组织系统（KOS）逐渐成熟，而且目前已经有很多结合知识组织系统的应用，但主要使用的是词汇表、规范文档、文本分类与聚类等工具与技术，大多都是基于词的分析与处理，应用词的关系与概念、本体技术等开展对科技文献文本内容分析的相关研究非常少。目前我们已经建成了一个“新能源汽车”领域的汉语科技词系统，下一步将研究如何将该词系统引入服务平台之中，在提高查准率与查全率的同时，进一步细化话题模型、排名系统等。

参考文献

[1] 刘寅虓. 系统分析之路[M]. 北京:电子工业出版社,2005.
[2] 国家科技图书文献中心（NSTL）[OL]. [2011-07-01]. http://www.nstl.gov.cn.
[3] Lucene[OL]. [2011-07-01]. http://lucene.apache.org/.
[4] 吴众欣,沈家立. Lucene分析与应用[M]. 北京:机械工业出版社,2008.
[5] HOFMANN T. Probabilistic Latent Semantic Indexing [C]// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, California, USA, 1999:50-57.
[6] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research,2003(3):993-1022.
[7] STEYVERS M, SMYTH, ROSEN-ZVI, et al. Probabilistic Author-Topic Models for Information Discovery [C]// Proceedings of the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD). Seattle, Washington, USA, 2004:306-315.
[8] TANG JIE, JIN RUO-MING, ZHANG JING. A Topic Modeling Approach and its Integration into the Random Walk Framework for Academic Search [C]// Proceedings of the 8th IEEE International Conference on Data Mining (ICDM). 2008:1055-1060.
[9] HERTZUN M, PEJTERSEN A M.The Information-Seeking Practices of Engineers: Searching for Documents as well as People [J]. Information Processing and Management: An International Journal,2000,36(5):761-778.
[10] PAGE L, BRIN S, MOTWANI R, et al. The PageRank Citation Ranking: Brining Order to the Web [R]. Technical Report SIDL-WP-1999-0120, Stanford University, 1999.
[11] KLEINBERG J M. Authoritative Sources in a Hyperlinked Environment [J]. Journal of the ACM,1999,46(5):604-632.
[12] NIE ZAI-QING, ZHANG YUAN-ZHI, WEN JI-RONG, et al. Object-Level Ranking: Bringing Order to Web Objects [C]// Proceedings of the 14th International Conference on World Wide Web (WWW).Chiba, Japan, 2005: 567-574.
[13] XI WEN-SI, ZHANG BEN-YU, CHEN ZHENG, et al. Link Fusion: A Unified Link Analysis Framework for Multi-Type Interrelated Data Objects [C]// Proceedings of the 13th International Conference on World Wide Web (WWW). New York, USA, 2004: 319-327.
[14] 严蔚敏,吴伟民. 数据结构（C语言版）[M]. 北京:清华大学出版社,2007:156-192.

作者简介

徐硕（1979-），男，博士，助理研究员。研究方向：数据挖掘、信息抽取、生物信息等。E-mail: xush@istic.ac.cn
乔晓东（1965-），男，研究员，硕士生导师，中国科学技术信息研究所/信息技术支持中心主任，中国互联网协会理事，中国情报学会计算机应用分会副主任委员，CILIP会员，研究方向为信息资源管理工作和信息服务。E-mail: qiaox@istic.ac.cn
朱礼军（1973-），男，副研究员，硕士生导师。2005年起，在中国科学技术信息研究所/信息技术支持中心从事一线科研工作。目前主要研究Semantic Web、Web Service和知识技术在科技信息服务、电子政务/商务中的应用以及知识组织系统的集成与服务体系。E-mail: zhulj@istic.ac.cn
张运良（1979-），男，副研究员。主要研究方向：文本自动分类，概念层次网络，知识组织。E-mail: zhangyl@istic.ac.cn

The Service Platform on Monitoring and Analyzing the Content of Science & Technology Information Resource

Xu Shuo, Qiao Xiaodong, Zhu Lijun, Zhang Yunliang / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: Currently, many researchers are eager to analyze the trends of recent researches in respective domain, get detailed information about exporters and important literatures in respective domain, and precisely evaluate the science & technology literature and the researchers' work. For these reasons, on the basis of the rich resource in National Science and Technology Library (NSTL), Institute of Scientific and Technical Information of China (ISTIC) and Tsinghua University, etc., we develop a service platform on monitoring and analyzing the content of science & technology information resource. This platform has the following functions: experts, journals/conferences and keywords united searching, association paths finding between two researchers, topic finding, etc. as well as embedded experts and literatures ranking.

Keywords: Knowledge service, Topic model, Association path, Ranking, Full-text indexing

（收稿日期：2011-07-08）