面向科技情报的互联网信息源自动发现技术

高 辉，陈 钧，牛海波，罗 威
中国国防科技信息中心 北京 100142

摘要
自动获取高质量互联网信息源是科技情报工作的一项基础性研究内容。以网站/网页类信息源和Twitter信息源为研究对象，基于共引关系以及关注关系和文本内容，分别提出了两类信息源的自动发现方法，并面向科技情报领域进行了实验。对信息源自动发现技术应用形式进行了研究，分析了科技情报工作对信息源服务的具体要求，提出了3类应用场景。

关键词
科技情报；互联网信息源；Twitter；共引；社会网络分析

doi: 10.11959/j.issn.2096-0271.2015040

Internet Information Sources Automatic Discovery Technology for Scientific and Technological Intelligence

Gao Hui, Chen Jun, Niu Haibo, Luo Wei
China Defense Science & Technology Information Center, Beijing 100142, China

Abstract
It is a basic work to discover high quality internet information sources automatically for scientific and technological intelligence. The technology of website/webpage information sources discovery was presented based on the co-citation relationship, and the technology of Twitter information sources discovery was presented based on the following relationship and content analysis. Then, the application forms of automatic discovery of information sources were discussed. Three kinds of application scenarios were presented based on the analysis of the requirements of scientific and technological intelligence.

Key words
scientific and technological intelligence, internet information source, Twitter, co-citation, social network analysis

1 引言

科技情报人员通常通过两种方式获取互联网信息：一是通过搜索引擎对某一主题相关的信息进行全面搜索；二是对所关注的领域积累大量有价值的网络信息源，通过对这些信息源持续跟踪而获得领域动态。第二种方式是一个长期而持续的工作，是进行技术预警、技术热点发现与跟踪、技术发展趋势预测等重要工作的基础。因此，全面掌握所关注领域相关的互联网信息源，对科技情报人员来说至关重要。

互联网信息源是指互联网上能够提供信息的各类媒体，各种机构、院校、企业几乎都拥有自己的网站甚至社交媒体账号，大量科技工作者通过各种社交媒体向外界发布着科技类消息，因此这些网站和社交媒体账号都是科技工作人员潜在的信息源。随着互联网的发展，互联网信息源的数量也不断增长，截至2014年7月全球网站数目超过9.7亿个[1]，2015年5月Twitter用户数量超过5亿户，活跃用户超过3亿户[2]。传统人工积累搜集互联网信息源的方式已经不能满足大数据时代对科技情报工作的要求，因此必须对互联网信息源的自动发现技术开展研究。

互联网信息源是互联网数据的生产者，信息源种类和数量的增加以及活跃度的提高，导致了数据的爆炸式增长。全世界数据总量以每两年翻一番的速度递增，而近十年来增长最快的当属互联网数据。未来的任务主要不是获取越来越多的数据，而是数据的去冗分类、去粗取精，提高知识发现的产出率[1]。要在不明显增加采集成本的条件下尽可能地提高数据的质量。这就要求在采集互联网信息时尽量选择与研究领域紧密相关的信息源，减少不必要的数据采集。如何获取相关性强、权威性高、时效性强的信息源，并能够及时有效地把信息源提供给科技情报研究人员，是一个重要的研究课题。

目前公开的互联网信息源服务主要有Yahoo Directory、Open Directory Project和Go Guide等，其实质属于目录式搜索引擎：一种按目录分类的网站链接列表，用户可以按照分类目录或关键字找到所需要的站点或栏目（即网页类信息源）。目录搜索引擎以人工方式或半自动方式搜集信息并整理分类。例如Open Directory Project的编辑工作目前共有近9万人参与，搜集了400万个站点信息，拥有100多万个分类[3]。该类信息源服务的缺点是需要人工介入、维护量大、信息量少、信息更新不及时。

本文研究科技领域相关的互联网信息源自动发现技术，以网站/网页类（以下简称Web类）和Twitter类信息源作为主要研究对象，提出并实现了互联网信息源自动发现技术，并对信息源的应用要求和服务形式进行了研究。

2 相关工作

2.1 问题描述

科技情报人员关注的互联网信息源可分为传统的Web信息源和社交媒体信息源两大类。其中，Web信息源主要包括领域相关的新闻聚合页或者重要机构的新闻发布页等。而社交媒体主要包括Twitter、Facebook、BBS、博客或者微信等，本文选取Twitter作为研究对象。

在信息源发现的需求建模中，科技情报人员往往无法使用有限的关键词对其关注的信息源进行描述。但是对于具有一定

1 http://www.internetlivestats.com/total-number-of-websites/

2 https://en.wikipedia.org/wiki/Twitter

3 http://www.dmoz.org

工作经历的科技情报人员来说，他们已经掌握了有限数量的领域内信息源，因此本文信息源自动发现技术的思路是：以已知信息源为种子，通过算法发现更多未知的信息源。如图1所示，首先给定一定数量的已有信息源作为种子，根据网页/Twitter所具有的网络关联特性或内容相关性，自动发现与种子领域相关且重要的新信息源，这个过程可以转化为挖掘与种子网页和Twitter账户相关度高的其他网页和账户的过程。

2.2 相似网页自动发现相关工作

相似网页/网站发现的相关工作可以简单分为基于内容的方法和基于链接关系的方法。基于内容的方法完全根据网页的内容来计算网页间的关联度。参考文献[2]从网页的各种标签内容中提取特征，提出了一种模糊内容分析方法来探索网页间的相关度。参考文献[3]首先用元搜索方法得到潜在相关的网页集合，然后抽取网页关键词进行相关性分析。SimilarSiteSearch[4]基于网页内容，使用机器学习方法对主题相近的网页进行识别，并在互联网上提供有限的服务和相关API。基于链接关系的算法将全部网页视为一个有向图，并利用图的连通性和加权信息来计算网页间的关联度。PageRank[4]算法和HITS[5]算法可以在一定程度上对相关网页进行排序，但是PageRank算法过分关注权威性而忽视相关性，HITS算法中可能出现主题漂移现象。参考文献[6]使用Companion和Co-Citation的两种算法来度量网页间的相关度。Companion算法将利用给定网页的出链接与入链接及其邻近网页构建一个有权图，并用一种HITS变种算法来挖掘给定网页的相关网页。Co-Citation即共引算法，通过检查网页的共引关系强度来挖掘给定网页的相关网页。参考文献[7]将网页分块算法引入共引过程中，并综合了链接锚文字的相似性和网页模板块过滤等方法，提高了关联网页的挖掘精度。

4 http://www.similarsitesearch.com/about.html

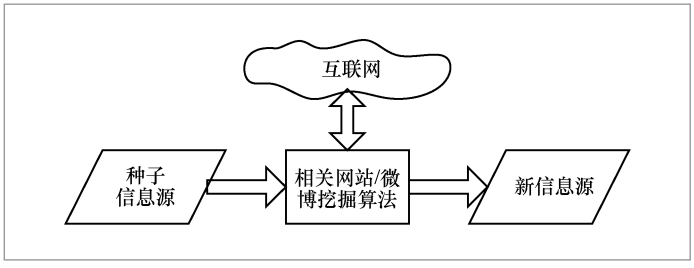


图1 信息源自动发现流程

2.3 相似微博用户自动发现相关工作

社交媒体用户之间通过关注、交互等行为形成了巨大的网络，微博相似用户发现方法首先将分析对象定位为网络的拓扑结构，相关的研究集中在：团体挖掘（发现用户的社交圈）[8]、人物影响力计算[9,10]、信息传播[11]等问题。参考文献[12]和参考文献[13]提出了两种基于标签信息进行用户推荐的方法。参考文献[14]提出在社交网络的历史数据可以获取的情况下，使用基于内容的方法进行用户推荐是有效的。参考文献[15,16]对LDA模型进行改进后，将其应用于微博主题挖掘，得到了较好的效果，能够进一步用于相似主题用户的发现。

3 互联网信息源自动发现技术

本文中Web类信息源的自动发现将完全依赖于链接关系而不考虑文本内容，这是因为Web类信息源的所有者一般都是机构组织等，网页内链接需经过审查才得以上线，因此比较能够代表相关性和权威性。同时网页内正文内容难以获得（各网站页面结构差异较大），噪声较多，基于文本内容进行相关性与权威性度量并

不理想。Twitter信息源则采用基于关联关系和内容相结合的自动发现方式，主要因为Twitter用户多为个人，用户之间的关注关系比较随意和多样化，无法真正反映出领域相关性。同时由于字数限制，推文（Tweet）内容比较精辟，在遣词造句上多选择具有实际意义的词。推文内容能够批量获得，且结构性比较好，因此本文同时基于关联关系和内容对Twitter信息源进行自动发现。

3.1 网页类信息源自动发现技术及实现

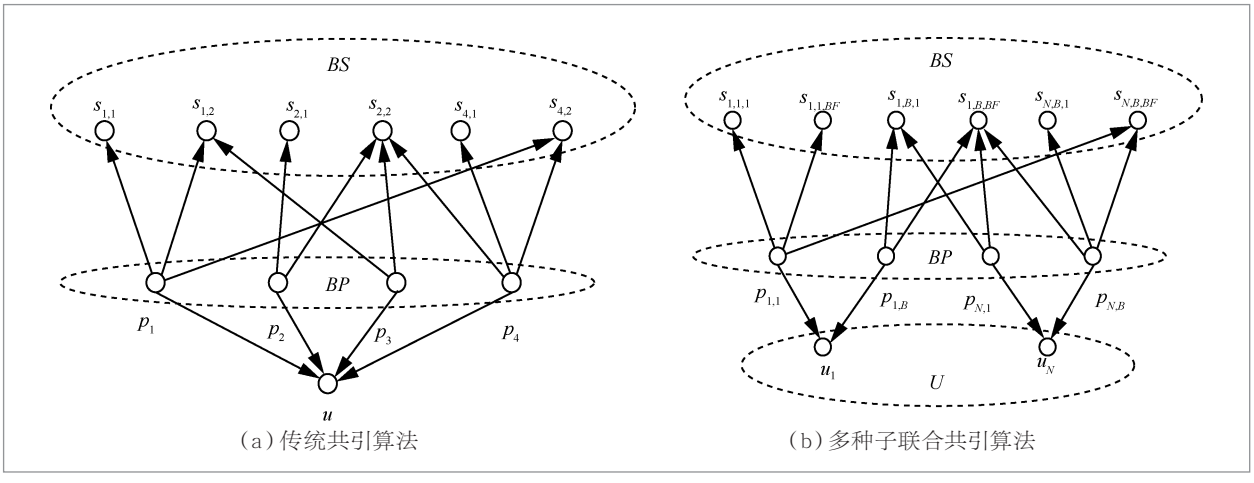
3.1.1 网页类信息源自动发现技术

针对网页类信息源，主要基于共引思想来自动发现与信息源相关的新信息源。给定一个网页u，含有指向u的链接的网页v称为u的父亲网页，也称v引用了u；u内部的链接指向的网页w称为u的儿子网页，也称w被u引用。如果网页p_1和p_2具有相同的父亲网页，则p_1和p_2称为共引关系。

共引分析最早出现在学术文献的分析中，共引是指两篇文献同时被其他文献引用。同被引用的文献在主题上具有或多或少的相似性，因此同被引用的次数可以预测文献在内容方面的相关性。在互联网中同样存在上述特性，一般认为具有共引关系的网页在所属领域上具有或多或少的相似性，因此共引次数可以预测网页在内容方面的相关性[17]。给定种子信息源，本文通过挖掘互联网中与其具有共引关系的网站来构建候选信息源。

共引算法一般过程是[6]：设u为种子信息源，首先找到引用它的父亲网页集合BP，再抽取BP中每一个父亲网页所引用的其他网页，组成兄弟网页集合BS。计算BS中网页与u出现共引的次数，共引次数越多说明与u的相关性越高。以图2（a）为例，可以直接看出BS中的共引次数，其中$s_{2,2}$与u的共引为3次。如果把阈值设为2次，则可以认为$s_{1,2}$、$s_{2,2}$、$s_{4,2}$与u相关，它们是由种子u得到的新信息源。

在传统共引算法基础上，前期研究[18]中提出了基于多种子联合共引的信息源发现算法，与传统算法不同，该算法选择N个已有信息源（种子集合U）作为输入，同时考虑了父亲网站的质量对最终结果的影响。为了对父亲网页的质量进行度量，引入了引用度的概念。如图2（b）所示，BP中父亲网页$p_{i,j}$（$i\in[1,N]$，$j\in[1,B]$，其中N为种子信息源总个数，B为每个种子信息源父亲



（a）传统共引算法　　（b）多种子联合共引算法

图 2　共引算法示意

网页的总个数）引用U中所有种子网页的总次数，称为$p_{i,j}$的引用度，表示为$C(p_{i,j})$，对种子集合引用次数越多，其引用度就越高，代表与种子之间的相关性（质量）越高。假设在图2（b）中$p_{1,B}$和$p_{N,1}$为同一个网页，即$p_{1,B}=p_{N,1}$，以图2（b）的引用关系为例，BP中节点的引用度见表1。相应地，BS中兄弟网页$s_{i,j,k}$（$k\in[1,BF]$，其中BF是每个父亲网页除种子信息源外其他儿子网页的总个数）的共引度则定义为$s_{i,j,k}$所有父亲的引用度之和。以图2（b）的引用关系为例，BS中节点的共引度见表2。

与传统共引思想相同，本文得到的共引度同样代表了BS中网页与种子网页之间的相关性。同时，与HITS算法[5]类似，BP对种子节点的引用度代表了Hub值，而BS中兄弟节点被BP引用的次数则代表了Authority值，因此本文共引度在一定程度上也代表了网页的重要度。

3.1.2 网页类信息源自动发现技术实现

在对Web信息源自动发现技术的实现中，首先对已掌握的信息源按照相关度进行人工分组（每组平均10个），每个组作为输入的种子信息源集合。令父亲网页数B=200，兄弟网页数BF=40。父亲网页的自动抓取使用Google公司或者AOL公司的Link搜索功能，当查找http://news.sciencemag.org/的父亲页面时，只要输入“link：http://news.sciencemag.org/”，便会返回众多父亲页面，本文通过编程实现了父亲网页的自动获取。目前以现有的200个信息源作为种子，利用本文技术获得6 200个质量较高的新信息源。参考文献[18]对采用多种子联合共引算法与普通共引算法的实验结果进行了对比，指出准确度能够提高50%以上。

表1 BP节点的引用度

BP节点	引用度
$C(p_{1,1})$	1
$C(p_{1,B})$	2
$C(p_{N,1})$	2
$C(p_{N,B})$	1

表2 BS节点的共引度

BS节点	共引度计算式	共引度
$D(s_{1,1,1})$	$C(p_{1,1})$	1
$D(s_{1,BF})$	$C(p_{1,1})$	1
$D(s_{1,B,1})$	$C(p_{1,B})+C(p_{N,1})$	4
$D(s_{1,B,BF})$	$C(p_{1,B})+C(p_{N,1})+C(p_{N,B})$	5
$D(s_{N,B,1})$	$C(p_{N,B})$	1
$D(s_{N,B,BF})$	$C(p_{1,1})+C(p_{N,B})$	2

3.2 微博类信息源自动发现技术及实现

3.2.1 微博类信息源自动发现技术

Twitter用户之间通过关注、被关注、消息转发等行为构成复杂的社会网络，本文基于社会网络分析法研究Twitter信息源自动发现技术。在Twitter使用实践中，用户积极选择并参与构建个性化关系，与一些具有相似特征和爱好的用户自发地聚集到一起形成社区[19]，因此可以把与种子信息源处于相同社区的其他用户作为领域相关的候选新信息源，可以基于推文内容对相关性进一步度量。社会网络中中心度的概念往往代表着节点的重要性，因此可以通过研究社区内节点的中心度来衡量新信息源的权威性。Twitter信息源自动发现主要分为候选集构建、用户重要度评估和领域相关性度量3个步骤，具体流程如图3所示。

候选集构建。首先选择种子用户作为起点，抽取其所有粉丝（关注者）作为第二轮样本，继续选择每个粉丝的粉丝作为第

三轮样本，依次进行抽取，直到达到终止条件。同时将种子用户自己关注的其他用户加入用户样本。本质上，该滚雪球样本一般是围绕着种子用户的关系而组织的[20]，构成的网络关联是紧密的，可以认为该样本与种子用户之间已经具备一定的领域相关性。在此基础上，利用基于图分割的社区挖掘方法获得种子所属的社区，进一步剔除无关用户。

用户重要度评估。中心性分析以社会网络节点的度数衡量节点中心性特征，以反映出节点在网络中的中心性地位差异，如果节点具有较高的度数，则它可能拥有更大的影响力。本项目用点度中心度来评测社区中的重要人物，点度中心度值高表示该用户受到较多人的关注，他发表的言论能够迅速被他人接收并对他人产生影响，该用户具有信息源的潜质。

领域相关性度量。领域相关性是评价信息源质量的重要指标，通过社团发现算法得到的候选集仍存在大量相关性不高的用户，因此本文引入了基于主题模型的推文内容相关性度量方法。LDA（latent dirichlet allocation）是一种重要的主题模型，本文使用LDA对候选集中用户的推文进行话题聚类，如果某个用户与种子用户在某一段时间内所发推文属于同一主题，则认为该用户与种子用户具有领域相关性。

3.2.2 Twitter信息源自动发现技术实现

Web类种子信息源大都对应Twitter官方账号，本节以Web类信息源对应的Twitter账号作为Twitter种子信息源。编程实现了Google搜索和Twitter API用户搜索相互补充的Twitter账户的自动获取，由200个Web类种子信息源得到了134个Twitter种子信息源。

Twitter信息的获取主要基于Twitter API实现，首先抽取种子用户的关注用户以及种子用户的粉丝、粉丝的粉丝，从而获得用户样本；采用Pajek[21]对该样本组成的网络进行可视化分析，计算种子所在社区以及各节点点度中心度；抓取每个用户最新的200条推文组成该账户的文档，使用的JGibbLda工具包对用户文档进行聚类，预先设置主题数量为4；在聚类结果中，如果与种子文档归属相同的主题，则说明该文档对应的账户与种子具有内容上的紧密相关性，通过该过程过滤掉不相关账户；最后结合各用户的中心度得到最终相关度高、重要性高的新信息源。

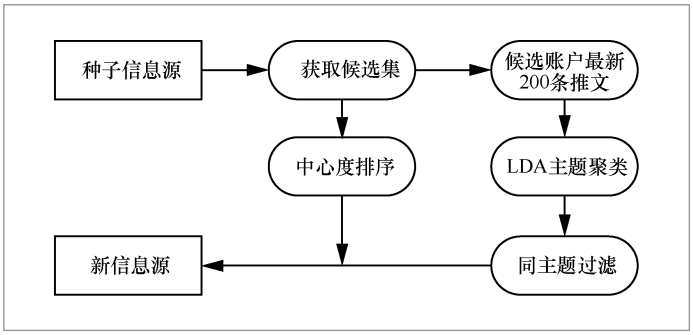


图 3 Twitter 信息源自动发现流程

4 互联网信息源自动发现技术应用

相比人工搜集方法，本文提出的信息源自动发现方法具有自动高效、覆盖面全、对新产生信息源反应快等明显优势。为了提供完善的应用服务，需要进一步对信息源进行标注和分类，建立国别地区、技术领域、应用范围、所有者性质（如个人、政府机构、大学院所）等维度的分类体系。然后根据信息源对应网站和微博的标题、关键词、摘要等描述信息，利用机器学习方法实现对信息源的分类和组织，最后形成完备的信息源库。结合大数据时代对科技情报工作提出的新要求，信息源自动发现技术具有如下应用场景。

（1）构建信息源地图，系统掌控全球科技信息资源

信息源地图指的是用可视化手段对信息源的综合展示，利用地图、热图和网络图等多种形式来展现科技领域信息源的地理位置、活跃度、统计分布、类别、信息源间关联交互等情况。通过一个全面、准确、动态的互联网信息源地图，决策人员和情报研究人员可以对科技信息资源进行全局把控和分析，从更高层次上挖掘发现其特点和规律，预测其变化趋势，具有重要战略意义。

（2）实现信息源检索服务，为情报研究工作提供保障

提供完善、灵活的信息源检索服务，为情报研究人员实现对科技领域互联网信息的持续跟踪和完成各项应急任务提供有力保障。其检索形式主要有以下3种。

- 目录式检索：用户通过分类层次目录方式检索库中已存在的信息源。
- 关键字检索：通过匹配信息源对应的描述性信息，检索库中已存在的信息源。
- 种子检索：当利用以上两种方式无法检索到所需要的信息源时，说明库中可能不存在该类信息源，这时用户可以输入已有信息源作为种子，通过服务系统在线挖掘获得新信息源。

除提供以上3种基本检索服务外，还可以开发个性订制和相关推荐等多种形式的智能服务。

（3）全面、深度挖掘科技信息源，为科技情报大数据提供数据来源

自动、高效、全面发现科技领域的信息源，建立标准的访问接口，实现与互联网海量信息采集平台无缝连接，为成规模的互联网信息资源获取提供必要前提。依据具有高度领域相关性的信息源采集数据，能够提高互联网数据采集的精准性和针对性，减少噪声数据的干扰，降低带宽、存储和计算成本。

5 结束语

互联网信息源自动发现技术能够高效发现大量新信息源，但较大的数量可能会使科技情报人员应接不暇，同时无法保证每个新信息源都是真正需要的，对新信息源的二次甄别也会影响其有效利用。值得庆幸的是，大数据相关技术已经广泛用于互联网信息的海量采集、处理和分析，大大提高了科技情报工作的效率，本文技术的直接用户更倾向于机器，而非情报人员本身。

下一步工作需要充分考虑从不同类型数据中发现信息，更全面地发现新信息源。因此，Web信息源自动发现和Twitter信息源自动发现两个过程不应孤立串行执行，应充分利用两类信息之间的互相映射、互相引用等关联关系，使两个过程紧密结合起来。再进一步，互联网资源采集系统对信息源采集到的网页和推文中包含的大量外链信息或者Twitter用户信息进行相关度和权威度的评估，选择优质信息源入库，实现信息源库的自我扩展。

参考文献

[1] 中国计算机学会大数据专家委员会. 中国大数据技术与产业发展白皮书（2013），2013
CCF Task Force on Big Data. White Paper on Big Data Technology and Industry Development in China (2013), 2013

[2] Loia V, Senatore S, Sessa M I. Discovering related web pages through fuzzy-context reasoning. Proceedings of the 2002 IEEE International Conference on Plasma Science, Banff, Alberta, Canada, 2002: 150~155

[3] Jaskirat S, Mukesh K. A meta search approach to find similarity between web

pages using different similarity measures. Proceedings of ICAC3 2011, Mumbai, India, 2011: 150~160

[4] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 1998, 30(98): 107~117

[5] Kleinberg J M. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999, 46(5): 604~632

[6] Dean J, Monika R. Finding related pages in the world wide web. Computer Networks, 1999, 31(11): 1467~1479

[7] 沈筱彦. Web信息检索若干关联挖掘问题的研究(博士学位论文). 北京: 北京邮电大学, 2009
She X Y. Research on several association rule mining problems for web information retrieval system (doctor dissertation). Beijing: Beijing University of Posts and Telecommunications, 2009

[8] Clauset A, Newman M E J, Moore C. Finding community structure in very large networks. Physical Review E: Statistical Nonlinear & Soft Matter Physics, 2004, 70(6): 264~277

[9] Crandall D, Cosley D, Huttenlocher D, *et al*. Feedback effects between similarity and social influence in online communities. Proceedings of the KDD'08, Las Vegas, Nevada, USA, 2008: 160~168

[10] Weng J, Lim E P, Jiang J, *et al*. Twitterrank: findingtopic-sensitive influential twitterers. Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, New York, USA, 2010: 261~270

[11] Wang D S, Wen Z, Tong H H, *et al*. Information spreading in context. Proceedings of the WWW 2011, Hyderabad, India, 2011: 735~744

[12] Yan Z, Zhou J. User recommendation with tensor factorization in social networks. Proceedings of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 2012: 3853~3856

[13] Guy I, Zwerdling N, Ronen I, *et al*. Social media recommendation based on people and tags. Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Geneva, Switzerland, 2010: 194~201

[14] Chen J, Geyer W, Dugan C, *et al*. Make new friends, but keep the old: recommending people on social networking sites. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, USA, 2009: 201~210

[15] 张晨逸, 孙建伶, 丁轶群. 基于MB-LDA 模型的微博主题挖掘. 计算机研究与发展, 2011, 48(10): 1795~1802
Zhang C Y, Sun J L, Ding Y Q. Topic mining for microblog based on MB-LDA model. Journal of Computer Research and Development, 2011, 48(10): 1795~1802

[16] 张晓艳, 王挺, 梁晓波. LDA模型在话题追踪中的应用. 计算机科学, 2011, 38(10A): 136~139
Zhang X Y, Wang T, Liang X B. Use of LDA model in topic tracking. Computer Science, 2011, 38(10A): 136~139

[17] Larson R. Bibliometrics of the world wide web: an exploratory analysis of the intellectual structure of cyberspace. Proceedings of Ann Meeting Am Soc Information Sciences, Medford, USA, 1996

[18] Gao H, Niu H B, Luo W. Internet information source discovery based on multi-seeds cocitation. Proceedings of International Conference on Security, Pattern Analysis, and Cybernetics (SPAC) 2014, Wuhan, China, 2014

[19] 王连喜, 蒋盛益, 庞观松等. 微博用户关系挖掘研究综述. 情报杂志, 2012, 31(12): 91~97
Wang L X, Jiang S Y, Pang G S, *et al*. A literature review of user relationship mining on microblog. Journal of Interlligence, 2012, 31(12): 91~97

[20] Scott J. 社会网络分析法(第二版). 刘军

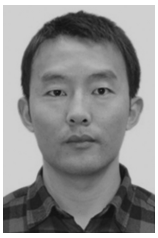
（译）. 重庆: 重庆大学出版社, 2007
Scott J. Social Network Analysis (Second Edition). Translated by Liu J. Chongqing: Chongqing University Press, 2007

[21] Wouter D N, Andrej M, Vladimir B. Exploratory Social Network Analysis with Pajek (Second Edition). Cambridge: Cambridge University Press, 2011

作者简介

高辉, 男, 博士, 中国国防科技信息中心工程师, 主要研究方向为互联网信息获取、信息抽取、知识库构建和信息可视化。

陈钧, 男, 中国国防科技信息中心高级工程师、研究室主任, 中国计算机学会大数据专家委员会委员, 中国科学技术情报学会信息技术专业委员会委员, 主要研究方向为科技信息大数据、网络工程等。



牛海波, 男, 中国国防科技信息中心工程师, 主要研究方向为大规模互联网信息资源获取、信息重构与融合、知识库构建等。

罗威, 男, 中国国防科技信息中心副研究员, 主要研究方向为信息抽取、大规模文本挖掘。

收稿日期: 2015-10-24
基金项目: 国家社会科学基金资助项目（No.4CTQ012）
Foundation Item: The National Social Science Foundation of China (No.4CTQ012)
论文引用格式: 高辉, 陈钧, 牛海波等.面向科技情报的互联网信息源自动发现技术. 大数据, 2015040
Gao H, Chen J, Niu H B, *et al*. Internet information sources automatic discovery technology for scientific and technological intelligence. Big Data Research, 2015040