情报学报 ISSN 1000-0135
第27卷第5期657-663，2008年10月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol.27 No.5, 657-663 October 2008

组织专家的检索系统设计与实现[1)]

陆 伟 韩曙光
（武汉大学信息资源研究中心，武汉 430072）

摘要 组织专家的检索是当前垂直信息检索研究的热门领域，也是组织信息检索研究的重要内容。本文总结了目前国内外组织专家的检索研究现状，分析了构建组织专家的检索系统的需求和挑战，采用基于相关文档集的归并排序法，利用组织内外的网页和期刊数据库等信息智能识别专家的专长，并基于此设计了从数据资源采集、规整、索引、检索到可视化等整个过程的组织专家的检索系统模型及以武汉大学为例的检索系统平台。

关键词 专家检索 专长识别 组织检索 专家聚类

Design and Implementation of Organization Expert Search System

Lu Wei and Han Shuguang
(*Center for Studies of Information Resources, Wuhan University, Wuhan* 430072)

Abstract As an important part of Organization Information Retrieval, Organization Expert Search is the current hot area in Vertical Information Retrieval research. Based on the analysis of the requirement and challenges, this paper summarizes the current development of the expert search, and proposes a general architecture of the organization expert search system, which contains data collections, sorting, indexing, retrieving, visualizing and so on, by using the relevant web pages and academic database as the data collections. At last, we construct an expert search system taking Wuhan University as an example.

Keywords expert search, expertise recognition, organization search, expert clustering

1 引 言

组织的竞争优势源于其自身知识的集合及学习能力[1]。根据 Delphi Group 的调查，组织中最大部分(42%)的知识是存在于员工头脑中的隐性知识[2]。这使得越来越多的组织意识到对自身知识，尤其是员工头脑中的隐性知识进行有效管理的必要性。然而识别这些知识并加以直接利用却非易事，但随着 Internet 的发展，企业、科研教学机构等纷纷构建起自己的网站，员工的专长信息及员工头脑中的隐性知识可以通过组织的相关网页(如组织官方网页、相关项目网页、员工主页等)、内部交流和共享的邮件记录等系列相关文档逐渐间接显化。因此如何从这些文档中识别出员工的专长进而辅助发现特定专长的专家，促进组织内外部人员的协作，为项目或团队挑选合适的人选以及选择项目评价所需的人员等就成为一个现实需要解决的研究课题。在国际上，该研究属于专家的识别与检索研究范畴。

知识便检索 Ssity E 与人工智能领域的专家系统不同，本文所谓组织专家的检索，是指利用组织内外能够表征专家专长的各种文档和资源，识别专家在某给定查询主题(领域)的专长(相关性)程度，并按程度高低排序显示专家结果列表的过程。早期组织专家的检索方法主要是通过建立描述组织内人员

收稿日期：2008年3月26日

作者简介：**陆伟**，男，1974年生，博士，副教授，主要研究领域：信息检索与智能挖掘，数字图书馆，知识管理等。E-mail: reedwhu@gmail.com。**韩曙光**，1987年生，硕士研究生，主要研究领域：信息检索。

1) 本文为国家社科基金项目(编号：06CTQ006)成果之一。

专长信息的数据库[3]，然而该方法不仅耗费人力财力，而且由于专家的技能和知识存在着分布性、难以量化、难以分级、不断变化的特点[4]，使得专家的描述信息具有很强的动态性和模糊性，导致数据库方法明显缺乏灵活性。

为了满足日益增加的检索专家的需求，解决专家专长信息的动态变化问题，本文借鉴TREC(文本检索国际会议)专家检索的基本方法，构建了一个通用的组织专家的检索系统框架模型。该框架模型可以定义组织内外表征专家信息的资源列表，设定资源动态更新周期，实现信息的动态采集，并结合组织内部专家列表，智能识别与检索组织专家。在下文章节2中，将介绍国内外组织专家的检索系统研究现状，章节3将详细介绍通用组织专家的检索系统框架模型的设计，在章节4中，将以武汉大学为例构建一个实际的组织专家的检索平台——WHU-ES，并给出初步的系统评价，文章最后对未来的工作做了简单的介绍。

2 国内外研究现状

为了动态挖掘组织内部相关资源和专家专长信息，国内外展开了一系列相关研究。TREC企业检索任务中的专家检索(Expert Search)子任务在一定程度上代表了当前组织专家的检索研究进展，当然目前也有如People Finder，MITRE's Expert Finder等在实际中使用的组织专家的检索系统，下文将分别予以介绍。

2.1 TREC会议专家检索子任务

作为Web Track的后继项目，TREC[5]于2005年起增加了企业检索(Enterprise Search)任务，并设立专家检索(Enterprise Expert Search)子任务[6]。该任务利用企业内部的网站网页、共享文档、电子邮件、数据库文件以及访问日志等作为企业数据集，对于给定的查询主题，参与者构建各自的专家识别与检索模型，并将得到排序后的相关专家列表等结果返回到TREC组办者进行测评。迄今为止，企业专家的检索已经举办了三届，在专家实体识别、专家专长表征信息提取、专家排序检索模型构建等方面取得了一系列的研究成果。笔者等自2006年以来也连续两年参加了专家检索子任务，对企业专家的检索进行了深入研究。

TREC专家检索的核心过程主要包括专家的实体识别和专家的检索排序两个部分。专家的实体识别作为实体识别的一种特殊情况，本文尚未考虑，笔者在具体实现时采用手工方式构建组织内的候选专家列表；关于专家的检索排序，目前主要有两种典型方法[7]，即：基于专家档案的方法和基于相关文档集归并排序的方法。前者利用组织内的各种信息资源，如网页、报告、邮件信息等，根据专家特征信息(主要是利用专家的姓名和电子邮件)在其中出现的情况，自动构建针对每个专家的个人描述(Profile)，然后将这些专家的描述作为文档，利用常用的检索技术对这些个人描述进行索引和检索，实现对专家的检索。笔者等[8]利用该方法采用窗口技术参加了TREC专家检索2006年的年度活动；后者首先利用传统信息检索方法检索出与主题相关的文档集合，然后进一步利用专家在文档中的特征信息对文档集合与得分进行归并，最终得到专家相对于查询主题的得分并排序。笔者等[9]采用该方法利用文档过滤模型参加了TREC专家检索2007年的年度活动，取得了良好的效果。关于这两种方法的具体实现思路及流程框架，可参见"基于文档权重归并法的企业专家检索"一文[10]。总体说来，这两种方法各有特点，各有优劣，目前到底该采用何种方法尚无权威的结论，在具体实现中可以根据需要而定。

2.2 已有专家的检索系统介绍

除了TREC会议参与者所采用的实验系统外，目前也出现了一些较典型的组织专家的检索系统，如MITRE公司构建的MITRE's Expert Finder系统、CSIRO(澳大利亚联邦科学与工业研究组织)构建的People Finder系统等。

2.2.1 MITRE's Expert Finder

MITRE's Expert Finder[4]系统的建立主要是为方便用户快速查找所需要的专家。该系统预收集了组织中表征员工基本信息的所有数据，包括员工之间的交流文档、员工的简历、网站网页及其他组织内部相关文档，并与组织员工数据库加以整合，构建组织专家的检索数据集。对于给定的查询主题，通过与该员工紧密相关的关键词和短语在数据集出现的频次等特征，计算员工与该查询主题的相关度并加以排序，同时提供相关的支撑文档。评价结果表明该系统平均可达到40%的查准率和30%的查全率。

2.2.2 People Finder

People Finder[11]是CSIRO在P@NOPTIC

Expert[12]基础上构建的基于 Web 的组织专家的检索系统,它主要利用发布在组织内部网上的所有文档及部分组织自身的其他数据,自动识别某个领域的专家。该系统的基本形式类似于搜索引擎,所不同之处在于,针对特定查询主题,其返回的不再是相关文档,而是一系列与该主题相关的经过排序后的专家列表,并辅助提供专家的详细联系方式和相关的支撑文档。该系统的效果受包含项目描述信息、企业员工简历及内部交流文档等信息的组织数据集的影响。

除上面介绍的两个专家的检索系统外,国外还出现了一些商用的检索专家系统,如 TACIT Active-Net(tm)、AskMe、Recommind 等。国内目前尚无采用类似机制检索专家的系统,值得一提的是,重庆维普资讯有限公司利用自身数据库资源(主要是期刊论文、学位论文以及学科分类体系)的优势,构建了中国科学家门户[13],为使用者提供了按照作者姓名、作者学科以及作者单位等检索专家的功能。而鉴于学科分类体系尚不完善,该系统缺乏对具体领域或专业查询主题的灵活支持;对自然语言查询的处理方面亦存在明显的不足;再者,检索专家所采用数据集资源的单一性,往往使系统不能全面反映专家在各个层面的专长。

3 组织专家的检索系统设计

综观 TREC 检索专家的实验系统及上文所述之应用系统,尽管都提供了根据特定查询主题生成经过排序的专家列表的功能,但在专家之间关联特性的挖掘上却都有欠缺。笔者认为,如能根据专家之间的共现规律,利用社会网络分析等方法可视化呈现专家之间的关联和聚类关系,对组织和访问者准确把握员工(专家)的专长有着重要意义。同时,专家专长信息动态变化的特性,也要求专家数据集能够动态更新,而以上各个系统对此都未予以重视,不利于组织动态把握员工的专长。再者,专家识别与检索数据集的单一性往往无法反映专家各个层面的专长,而上述系统也未考虑采用不同类型的数据集。基于此,笔者认为组织专家的检索系统主要应提供以下几个方面的功能:定义表征专家专长的数据集类型;动态构建及更新专家数据集;专家专长(领域)的动态识别;检索针对特定查询主题的相关专家;专家共现和聚类关系的可视化呈现等。围绕着这些功能目标,笔者提出并构建了一个通用的组织专家的检索系统框架模型,下文将就构建思路和系统体系结构作详细介绍。

3.1 系统总体思路

借鉴 TREC 专家检索的两种基本方法,笔者认为组织专家的检索系统构建的主要思路是:首先,通过 spider(信息采集蜘蛛程序)采集已定义的表征专家信息的组织内外部数据资源,获取专家数据集,并生成专家数据集索引文件;然后,提取组织内所有专家列表,利用专家数据集索引文件,根据专家表征信息(如专家姓名、电子邮件等,本文主要采用专家姓名)在数据集文档中的出现情况,生成专家-文档映射文件(Expert-Docs,该文档记录了数据集中每个专家对应出现的文档列表);则针对特定查询主题,用户在检索专家时将首先返回与该主题相关的文档,然后对于每位专家根据其在相关文档和 Expert-Docs 中的共现情况,过滤产生每个专家针对该主题的相关文档,然后归并各文档得分作为专家得分(目前的归并方法是简单的线性相加),最后根据专家得分的高低排序显示。该检索过程的基本思路可用下面算法(JAVA 风格)实现:

```
0:Float ExpertScore[ ];
  String ExperDocs[ ]Expert.readExpertDocs( );
  //读取专家文档映射文件、初始专家得分(初始化为 0);
1:query = Query term;
  //获取用户查询主题:
2:hits = searcher.search(query);
  //根据查询主题检索并返回相关文档集(用 hits 结构表示,该结构包含文档号,得分等详细信息);
3:for (int i = 0;i < hits.length;i + + )
  {
  //遍历相关文档集(即 hits 结构)中的所有文档;
4:for(int j = 0;j < ExperNames.length;j + + )
  {
5:if(ExpertDocs[j].contains(hits.id(i)));
6:ExpertScore[j + ] = hits.score(i);
  }
  //遍历专家,若文档中含有该专家信息,则依据文档得分模型(如下公式(1)增加专家得分。
  }
```

对于每篇文档相关性得分的计算采用的是向量空间模型(VSM),如公式(1)所示,其采用的是开源

软件 Lucene 的评分机制，详见[14]。

$$w_i = \sum_{t \text{ in } q} tf(t \text{ in } d) * idf(t) * boost(t.filed \text{ in } d) * lengthNormal(t.filed \text{ in } d) \quad (1)$$

3.2 系统体系结构

根据上文所述，本文所构建的检索系统主要包括下面四个模块，即 Spider 模块、Indexer 模块、Assistant 模块以及 Searcher 模块，各模块间关系参见系统整体框架图(图 1)。具体如下：

(1) Spider 模块，即信息采集模块，主要功能是根据系统管理者定义的表征组织专家专长的不同信息资源构建相应的采集策略，并设定资源采集周期，定期采集和更新专家数据集。笔者认为可以参考使用的专家数据集主要有组织内部表征专家专长的网页和文档库、利用专家名称和单位构建检索主题从搜索引擎获取的表征专家的信息库、相关学术数据库以及专利数据库等。

(2) Indexer 模块，该模块为数据集索引模块，其主要功能是对 Spider 模块采集到的不同专家数据集加以解析、整合，并根据不同索引策略，建立不同数据集的索引并根据需要归并各个索引文件。

(3) Assistant 模块，即辅助文档构建模块，该模块的主要功能是构建系统运行所必须的一系列辅助文档，主要包括组织数据源列表文件(该文件是 Spider 模块采集数据资源的依据)、专家列表文件(含专家姓名和机构名称，主要为从搜索引擎和学术数据库中检索数据资源构建检索表达式等)、主题词列表文件(为动态智能识别和呈现专家专长领域提供主题词)、专家—文档映射文件(见 3.1)等。

(4) Searcher 模块，该模块根据用户提交的查询主题，利用传统信息检索的模型和方法得到排序的检索结果，并使用 Assistant 模块生成的专家—文档映射文件和专家列表文件对相关文档加以分析，生成针对特定主题的专家得分，继而依据得分高低显示专家排序。该模块是用户接口模块，需考虑用户不同层面的需求，由于递减排序的专家列表无法直接反映各专家之间的聚类关联关系，本模块还包含了利用可视化的方法呈现专家共现及聚类关系的功能。

此外，为进一步了解用户的需求，不断提高专家专长识别的准确度，本系统还设计了用户反馈模块，并与检索过程加以整合，这里不再详述。下文将以该系统整体框架为基础，以武汉大学为例，介绍武汉大学专家的检索系统——WHU-ES 的具体实现及初步评价。

4 WHU-ES 的实现与评价

4.1 WHU-ES 的实现

(1)初始辅助文档库的定义：初始辅助文档库包

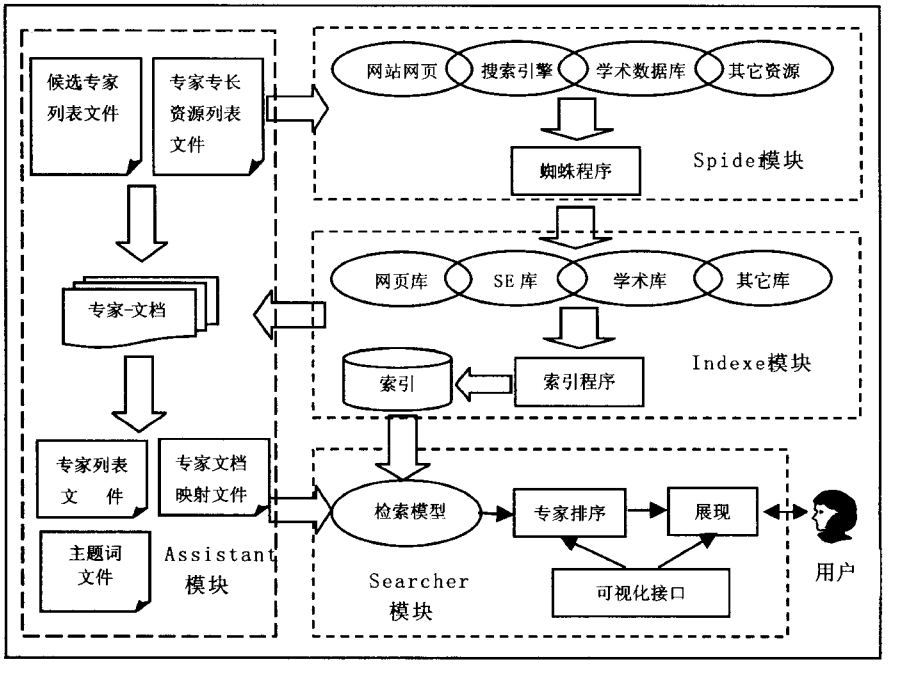

图 1 系统整体框架图

括武汉大学专家列表文件，主题词列表文件，专家数据集资源列表文件等。WHU-ES 中专家列表文件利用各院系主页的教师列表信息加以构建；主题词列表文件参考现有的主题词表等构建，并支持动态添加和更改；专家数据集资源列表文件所确定要采集的内容包括武汉大学所有以 whu.edu.cn 子域名结尾的网站的网页信息、万方数据库收录的武汉大学近十年来发表的论文信息（包括标题、作者、摘要等）以及搜索引擎中检索武汉大学专家得到的结果数据等。

（2）专家数据集的动态采集和更新：在预定义采集参数（主要有采集范围，采集频率，采集的广度和深度等）的基础上，根据专家数据集资源类型的不同，需构建不同的采集策略，并实现动态采集。在 WHU-ES 中，笔者利用自己开发的蜘蛛程序采用广度优先策略扫描网页（为限定网页采集的数量，笔者设置了相关的过滤策略，如屏蔽特定的网站、限定网页的深度、剔除特殊的文件格式等），构建组织内部网页数据集（采集时间：2007 年 12 月，记录 23 万余条）；利用特定采集程序在搜索引擎 Baidu 和 Google 中抓取通过提交专家名结合专家机构的检索式（如："马费成" and "武汉大学"）所得到的结果集，构建搜索引擎数据集（采集时间：2007 年 12 月，记录 21 万余条）；学术数据集的构建则是通过程序自动抓取万方数据库中限定作者单位是"武汉大学"的所有数据（采集时间：2007 年 10 月，数据记录 5 万余条）。

（3）数据集的规整及索引：采集来的不同数据集的数据格式不尽相同，在建立索引之前需对数据集做剔除噪音信息等预处理，WHU-ES 中所有获取的文档集合都按照特定策略被规整成网页形式。建立索引的程序是在全文检索开源软件 Lucene 以及开源的 HTML 解析器 Html Parser 的基础上二次开发的，Lucene 的详细信息可参见文献[15]，Html Parser 的详细信息可参见文献[16]。通过 Html Parser 解析规整后的 Html 网页，将得到的结果写入索引的不同域（索引文件存储标题、文档全文、URL、修改时间四个域）中。

（4）用户检索接口的开发：为方便用户的使用，WHU-ES 提供了基于 Web 的检索接口，可实现从组织内部网页数据集、搜索引擎数据集、学术数据集等多个数据集角度分别检索识别专家，也可利用多个数据集归并后的综合数据集实现检索功能。当用户提交特定查询主题后，系统根据专家的检索模型返回专家列表，并根据共现关系提供可视化的展示。图 2 是用户利用搜索引擎数据集，检索"知识产权"所获得的检索结果，图中左侧展现的是排序后的专

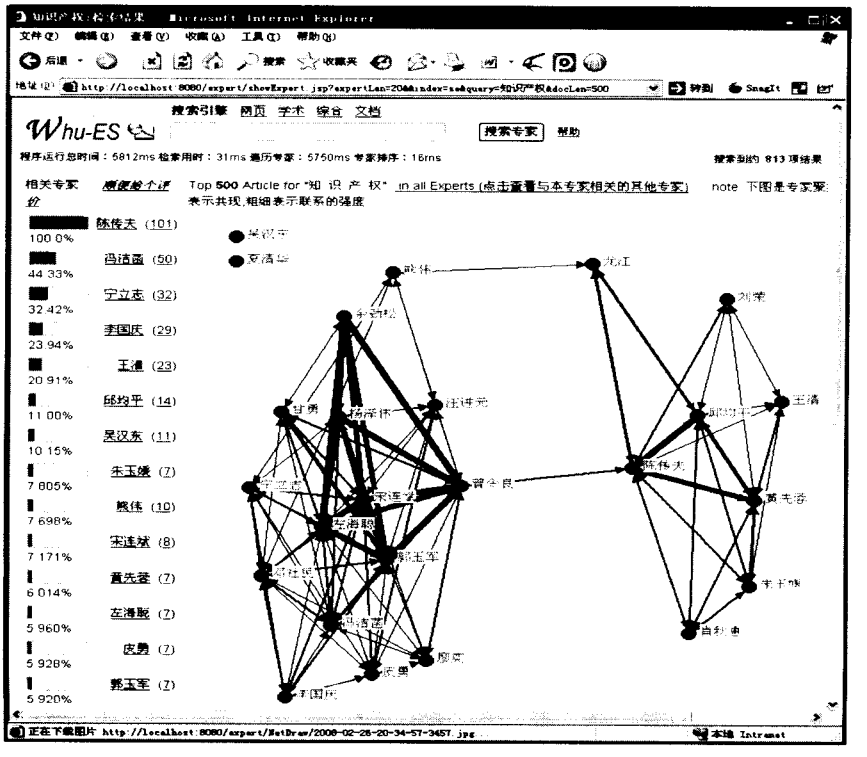

图 2　检索"知识产权"返回的专家排序列表和专家共现网络关系图

家列表,右侧展现的是专家共现聚类关系图,该关系图使用社会网络软件 NetDraw[17] 的批处理接口,根据检索得到的专家姓名列表实时动态生成。关系图中的连线表示专家之间存在共现关系,连线的粗细表示专家关联(共现)强度,从图中可以明显看出,武汉大学"知识产权"研究团体分为两类,这也与实际的情况极其相符,一类是法学院的研究团队,另一类是信息管理学院的研究团队。此外,通过点击专家姓名右侧的数字可以显示其专长的支撑文档。

4.2 WHU-ES 的初步评价

信息检索系统的评价是系统性能的量化反映,一般来说,可以从性能和效果两个角度去考察。本文采用了最常用的检索评价指标之一的 P@N,对返回专家的检索效果作了初步评价。P@N 是系统对于查询主题返回前 N 个结果的准确率,常常可以比较有效地反映系统在真实应用环境下所表现的性能。为全面考虑到各领域的专家,笔者设计了武汉大学社会科学部、理学部等六大学部共 50 个查询主题,并将 WHU-ES 检索返回的专家列表(利用从搜索引擎上采集的专家数据集)连同查询主题做成调查问卷,送给相关专业人员进行评价,获得各查询主题的 P@5 和 P@10 得分,取平均值,得到的评测结果见表 1。

从表中可以看出,系统平均的 P@5 和 P@10 值为 0.7640,0.6060,即前 5 个结果中一般有 3 ~ 4 个相关专家,前 10 个结果中也一般有 6 个相关结果,能够基本满足用户的应用需求。

表 1 武汉大学专家的检索系统测评值

所属学部	查询主题	P @ 5	P @ 10
社会科学部	14	0.8714	0.7786
人文科学部	9	0.7111	0.4889
理学部	9	0.7111	0.5222
信息科学部	7	0.7429	0.5714
工学部	6	0.8333	0.6333
医学部	5	0.6000	0.5000
总体	50	0.7640	0.6060

5 结 语

本文采用基于相关文档集的归并排序法,构建了一个通用的组织专家的检索系统框架模型,同时开发了武汉大学专家的检索系统 WHU-ES,并对该系统专家识别的效果做了初步的评价。通过该系统,可初步识别出具有特定领域专长的专家,并能对专家之间的共现关系进行可视化呈现。当然,该系统的设计只是笔者专家识别与检索研究工作的起点,在未来的研究中,将对系统的功能进一步予以完善,不仅要考虑专家与文档级的映射关系,还要考虑其与具体的章节甚至段落等片断信息的映射,以提高专家识别的准确度;同时还将在系统中引入实体识别技术及本体技术等,不断提高其自动化程度;关于专家识别效果的深入评价是未来研究工作的一个重点,笔者将进一步考虑引入能表征专家专长的其他数据集(如专利数据集等),并对所采用的各种数据集的效果进行全面评价。

参 考 文 献

[1] Chris Argyris, Donald Schon. Organizational Learning: a theory of action perspectives[M]. Reading, Massachusetts: Addison-Wesley, 1978:346-348.

[2] 显性知识和隐性知识相互转换的过程[EB/OL]. [2008-05-20]. http://www.360doc.com/showweb/0/0/555973.aspx.

[3] Dawit Yimam-Seid, Alfred Kobsa. Expert finding systems for organizations: Problem and domain analysis and the demoir approach[J]. Journal of Organizational Computing and Electronic Commerce, 2003,13(1):1-24.

[4] Mark Maybury. Discovering Distributed Expertise[C/OL]// AAAI Fall Symposium Regarding the "Intelligence" in Distributed Intelligent Systems (RIDIS), 2007. [2008-5-30]. http://www.mitre-corporation.org/work/tech_papers/tech_papers_07/07_0730/07_0730.pdf.

[5] TREC Home Page[EB/OL]. [2008-05-20]. http://trec.nist.gov/.

[6] Nick Craswell, Arjen P. de Vries, Ian Soboroff. Overview of the TREC-2005[C/OL]// Proceedings of the 14th Text Retrieval Conference, 2005. [2008-5-30]. http://trec.nist.gov/pubs/trec14/papers/ENTERPRISE.OVERVIEW.pdf.

[7] Thijs Westerveld. Correlating Topic Rankings and Person Rankings to Find Experts[C/OL]. //Proceedings of the 15th Text Retrieval Conference, 2006. [2008-05-30]. http://trec.nist.gov/pubs/trec15/papers/cwi.ent.final.pdf.

[8] Lu Wei, Stephen Robertson, Andrew Macfarlane, etc. Window-based Enterprise Expert Search [C/OL]// Proceedings of the 15th Text REtrieval Conference, 2006.

[2008-05-30]. http://trec.nist.gov/pubs/trec15/papers/cityu-london.ent.pdf.

[9] Jiang Jiepu, Lu Wei, Liu Dan. CSIR at TREC 2007 Expert Search Task[C/OL]//Proceedings of the 16th Text REtrieval Conference, 2007. [2008-5-30]. http://trec.nist.gov/pubs/trec16/papers/wuhanu.ent.final.pdf.

[10] 陆伟，赵浩镇. 基于文档权重归并法的企业专家检索[J]. 现代图书情报技术，2008(7):38-42.

[11] Alistair McLean, Anne-Marie Vercoustre, MingFang Wu. Enterprise People Finder: Combining Evidences from Web Pages and Corporate Data[C/OL]//Proceedings of the 8th Australian Document Computing Symposium. Canberra, Australia, 2003. [2008-05-19]. http://hal.archives-ouvertes.fr/docs/00/03/54/04/PDF/ADCS-03.pdf.

[12] Nick Craswell, David Hawking. P @ NOPTIC Expert: Searching for Experts not just for Documents[C/OL]// The Ausweb Poster Proceedings, 2001. [2008-05-19]. http://research.microsoft.com/users/nickcr/pubs/craswell_ausweb01.pdf.

[13] 维普数据库中国科学家门户[EB/OL]. [2008-05-20]. http://www.cqvip.com/zuozhekj/.

[14] Otis Gospodnetic, Erik Hatcher. Lucene IN Action 中文版[M]. 谭鸿，等译. 北京：电子工业出版社，2007，1:72-73.

[15] Apache software Foundation, Apache Lucene-Overview[EB/OL]. [2008-05-21]. http://lucene.apache.org/java/docs/index.html.

[16] HTML Parser[EB/OL]. [2008-05-21]. http://htmlparser.sourceforge.net/.

[17] A Brief Guide to Using NetDraw[EB/OL]. [2008-05-20]. http://www.analytictech.com/Netdraw/NetdrawGuide.doc.

（责任编辑　王建平）