

国家科技文献中心多语言信息服务研究及其应用*

□ 张满年 高影繁 徐红姣 张均胜 王惠临 / 中国科学技术信息研究所 北京 100038

摘要: 如何提高多语言信息服务质量已成为数字图书馆等科技信息服务领域的重要研究问题。文章首先介绍了国内外多语言信息服务相关研究,然后具体从跨语言信息检索和机器翻译两个方面介绍了国家科技文献中心多语言信息服务研究成果在国家科技文献在线服务系统中的应用。将跨语言信息检索功能和文摘翻译服务功能引入数字图书馆在线查询系统,在国内数字图书馆信息服务领域尚属探索性尝试,可以为进一步提高数字图书馆多语言信息服务质量提供经验。

关键词: 多语言信息服务, 跨语言信息检索, 数字图书馆, 机器翻译

DOI: 10.3772/j.issn.1673—2286.2011.12.003

1 引言

因特网和万维网的普及与迅猛发展使得信息获取变得更加快捷与方便。物理空间距离不再是信息交流与传递的主要障碍,而多语言信息交流问题已逐步成为阻碍人类交流的主要障碍。

数字图书馆作为数字信息的收藏者和信息交流港口,在信息与知识传播中起到越来越重要的作用。数字图书馆领域中同样存在多语言文献信息服务问题。数字图书馆中的文献信息表示语言可以是多种自然语言,如汉语、英语、法语、德语、西班牙语、俄语、日语、葡萄牙语、韩语等等。虽然有些学者与科学家能够掌握多种外语,并且可以熟练地在多语文献中获取信息,但是大多数科学家仅能够熟练使用一到三种外语,难以有效地了解其他语种的文献信息,难以避免信息调查片面,对科技信息决策与研究产生不利影响。为此,急需在科技文献数字图书馆领域开展多语言信

息服务研究,并在数字图书馆网络服务系统中加以应用,满足广大科技工作者的多语言信息获取需求。

目前国内主要数字图书馆网络服务系统中,主要通过分语种检索的方式提供信息检索服务。在国家科技图书文献中心(National Science and Technology Library, NSTL)网络信息服务系统中首次加入跨语言信息检索词推荐及英文文献标题及文摘信息翻译服务。本文在调研分析国内外多语言信息服务研究的基础上,重点介绍NSTL网络系统中的多语言信息服务应用及相关的技术研究,并对国内数字图书馆领域多语言信息服务未来发展方向进行探讨。

2 相关工作

跨语言信息检索服务研究起源于多语言信息检索。1996年,在ACM SIGIR Workshop for Multilingual Information Retrieval会议上,与会人员经讨论提

* 本文受中国科学技术信息研究所学科建设“自然语言处理”课题(XK2011-6)、中国科学技术信息研究所重点工作“多语言信息获取关键技术研究与应用示范”课题(ZD2011-3-3)、中国科学技术信息研究所科研项目预研资金(YY-201121)和国家社科基金项目“基于本体的跨语言信息检索理论与实验研究”(06BTQ030)支持。

出跨语言信息检索 (Cross-Language Information Retrieval, CLIR)。Jian-Yun Nie^[1]将多语言信息检索 (Multilingual Information Retrieval) 和跨语言信息检索 (Cross-Language Information Retrieval) 分别定义为 “retrieval of documents in several languages from a query” 和 “retrieval of documents in a language different from that of a query”。在利用一种语言检索另外一种语言的文献时, 后者的描述更为精准。查询所用语言称为源语言 (Source Language), 文档所用语言称为目标语言 (Target Language)。如何将繁多的网络资源用不同的语言提供给用户, 如何用可互换的语言形式描述相同或相似内容的信息资源, 成为近年来信息检索研究领域中的前沿课题。

从时间上划分, CLIR的研究历程主要分为三个阶段^[2-5]:

(1) 20世纪60年代后期~20世纪80年代中期

这一时期普遍采用受控词表方法, 利用多语种叙词表来使查询语言和文献表示语言达到一致。该时期的研究表现出明显的区域特性, 主要是美国和欧洲的学者参与, 所开展的研究主要针对欧洲语言。最早有关CLIR研究的报道来自于1964年, 研究者将一部兼具英法德三种语言的受控叙词表应用于国际公路研究文献系统的开发^[6]。Pevzner分别于1969^[7]和1972年^[8]发文阐述如何运用翻译为英语的俄语叙词表来建立布尔精确匹配的文本检索系统。1970年康奈尔大学的Salton^[9]利用手工编制的英德双语词表, 对其SMART检索系统进行扩充, 最先对跨语言检索的效率进行了评价。1973年, 针对国际联机检索, Salton又利用一个英法多语种词表进行实验^[10], 在建立一个共同的概念集之后, 单独对每种语言的相应部分进行开发, 以使词表有更大的覆盖范围。CLIR的商务应用紧随其后, 到1977年, 欧洲已有能够识别四种语言的多语文本检索系统^[11]。1978年国际标准化组织ISO颁布了关于多语言叙词表 (Multilingual Thesauri) 的国际标准ISO5964, 并于1985年进行了修订^[12]。

但是, 以叙词表为代表的受控词表方法并不足以解决CLIR中存在的问题。Dubios在1987年指出: ①索引和维护成本限制了基于叙词表的CLIR系统的扩展性; ②叙词表的更新速度总是落后于新词的产生速度; ③未经过训练的用户很难驾驭基于叙词表的技术。因此, 基于受控词表的CLIR需要变革, 面向自由文本的技术整装待发。

(2) 20世纪80年代后期~20世纪90年代后期

这一阶段跨语言检索面临着从基于受控词表的CLIR到面向自由文本CLIR的转变, 与前者相比, 后者直接使用出现在文献中的词。Landauer和Littman最早研究基于自由文本的CLIR, 并利用一种扩展了的自动技术来降低词汇差异对检索效果的影响^[13,14]。Radwan和Fluhr在1991年提出一种使用了人工编码的翻译知识, 通过查询翻译策略实现CLIR的技术^[15]。尽管此后又取得了很多进展, 但上述两项研究所确定的两种主要方法——基于语料库的方法和基于知识的方法仍然占主导地位。此后应用于CLIR的理论方法不断创新, CLIR研究活动空前活跃。1994年诞生了第一篇关于跨语言信息检索的博士论文。1996年, 相似性叙词表 (Similarity Thesaurus) 被应用于跨语言检索^[16]。1997年卡内基梅隆大学将广义向量空间模型 (Generalized Vector Space Model, GVSM) 应用于跨语言检索。1998年, Pirkola提出用结构化查询的方法来消除查询翻译歧义^[17]。马萨诸塞大学和施乐公司采用基于词典的跨语言信息检索方法^[18]。

这一时期CLIR的研究队伍不断壮大, 除欧美地区外, 亚洲学者也开始表现出对这个领域的关注。汉语、日语、韩语、越南语、印尼语等语言进入跨语言检索的研究视野之内。9.11事件后, 美国开始重视研究用英语查询阿拉伯语信息。特别值得注意的是, 国际重量级的学术会议开始关注CLIR。1996年8月, 美国计算机学会 (ACM) 信息检索特别兴趣小组 (Special Interest Group on Information Retrieval, SIGIR)^[19]召开的跨语言信息检索专题研讨会是较早的一次关于跨语言检索的专题会议。从1997年开始, 文本检索会议 (Text Retrieval Conference, TREC)^[20]将跨语言检索正式接纳为一个专题。1997年3月, 美国人工智能学会 (AAAI)^[21]在斯坦福大学举办了一次跨语言文本和语音检索高级会议。1999年, 由美国国防部高级研究计划署 (DARPA) 资助的第一个大型跨语言信息获取研究项目TIEDS (Translingual Information Detection, Extraction and Summarization)^[22]正式进入研究阶段。1999年8月, 在日本东京举行了第一届日本国家科学信息系统中心跨语言信息检索测试会议 (NII-NACSIS Text Collection for IR Systems, NTCIR)^[23], 侧重于亚洲语言 (如汉语、日语、俄语) 的CLIR问题研究。2000年9月, 第一届跨语言评价论坛 (Cross-Language Evaluation Forum, CLEF)^[24]在葡

萄牙召开,主要内容是欧洲语言之间的CLIR评价。

(3) 20世纪90年代后期至今

20世纪90年代后期以来,CLIR技术更是取得了突飞猛进的发展。1997年由德国语言技术研究室人工智能研发中心研究开发的Mulinex系统^[25,26]是世界上第一个成功地运用跨语言自动翻译技术,使人们利用本国语言就能有效获取网上其他语言信息的CLIR系统,系统涉及英、德、法三语种。根据Oard D W的不完全统计,到2002年4月共有17个CLIR系统问世。到目前为止,CLIR系统大致分为两类:示范性系统和商业性系

统^[27]。前者比较著名的有Aport系统、Arctos系统、Eric系统、Mulinex系统、Mudial系统等;后者更具商业实用性,如Cindor系统^[28]、Rotondo系统、TextFinder系统等。近年来,许多研究机构致力于开发交互式跨语言检索系统,如美国马里兰大学的MIRACLE系统^[29]、西班牙的UNED系统^[30]、荷兰Twente大学的Twenty-One系统、英国谢菲尔德大学的Clarity系统等。这些交互式系统因用户的参与提高了CLIR的系统性能且检索界面更加友好,是CLIR系统中颇具潜力的一种。

本文以国家科技图书文献中心国家科技数字图书

表1 1998—2009年西文跨语言检索文献的年发文章量

论文来源	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
西文期刊	0	4	1	2	3	8	16	22	22	10	11	5
外文会议	8	8	9	80	113	109	30	57	28	48	40	9
学位论文	1	0	0	2	0	1	0	1	0	0	1	0
科技报告	1	0	4	5	1	3	0	2	3	0	1	0
总计	10	12	14	89	117	121	48	82	53	58	53	14

表2 1998—2009年中文跨语言检索文献的年发文章量

论文来源	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
中文期刊	0	0	0	7	1	3	11	15	28	4	8	2
中文会议	0	0	0	1	1	3	1	1	3	2	4	0
学位论文	0	0	0	1	0	5	0	4	5	5	8	6
总计	0	0	0	9	2	11	12	20	36	11	20	8

馆^[31]的西文库和中文库作为数据源,对1998年至2009年间的学术论文进行检索(截至2010年3月21日),西文库采用“cross language information retrieval”或“bilingual information retrieval”作为检索式,中文库采用“跨语言信息检索”或“跨语言检索”作为检索式。西文、中文的检索结果分别见表1和表2。由于数据统计发生在2010年初,所以2009年数据可能存在部分论文未入库的情况。从表1中的数据可以看出,进入21世纪后,CLIR的研究成果急剧增多,并在2001~2003年形成了一个高潮;此后的研究热度虽有所降低,但基本上呈现了一个稳定的发展态势。与表1相比,表2的数据

表明中文CLIR的研究明显滞后于西文,21世纪初是它的发展期,2005~2006年形成了一个小的研究高潮,随后的研究状态也基本稳定。

从本世纪初开始,CLEF、NTCIR等主要的CLIR会议和论坛都将研究方向从即兴跨语言信息检索(ad-hoc CLIR)转向更明确的、更特定的检索任务的研究,如跨语言问答(cross-language question answering)、交互式跨语言检索(interactive CLIR)、跨语言图像检索(cross-language image retrieval)、医学信息跨语言检索(medical CLIR)、多语言语音检索(multi-lingual speech retrieval)等。

表3 各种多语言搜索引擎

语种	地区	搜索引擎
英语	欧洲	www.google.com、www.yahoo.com、
	美洲	www.aol.com、www.lycos.com、www.altavista.com、www.google.co.uk、
	非洲	www.webcrawler.com、www.dogpile.com、www.google.ca、www.google.com.au、uk.yahoo.com、ca.yahoo.com
	亚洲	
	大洋洲	
西班牙语	欧洲	www.lycos.es、www.google.es、www.yahoo.es、www.google.com.ar、www.terra.com.mx、www.google.com.mx、aryahoo.com、www.google.cl、mx.yahoo.com
	南美	
阿拉伯语	非洲	www.google.ae、www.google.com.ly、www.google.com.eg、www.google.com.sa、www.google.co.ma、
	中东	
法语	西欧	www.lycos.fr、www.google.fr、www.yahoo.fr、www.seek.fr、www.hotbot.fr、www.netscape.fr、www.voila.fr、www.google.co.za、www.orange.fr、www.google.gr、cf.yahoo.com
	美洲	
	非洲	
	大洋洲	
俄语	东欧	www.yandex.ru、www.mail.ru、www.Rambler.ru、www.google.ru、www.yahoo.ru、www.Nigma.ru
	中亚	
葡萄牙语	欧洲	www.google.pt、pt.altavista.com、www.clix.pt、www.aonde.com、www.google.com.br、busca.uol.com.br、br.altavista.com、buscador.terra.com.br
	非洲	
	南美	
德语	欧洲	www.google.de、www.yahoo.de、www.altavista.de、www.lycos.de、www.aladin.de、www.acoon.de、www.allesklar.de、www.fireball.de、www.web.de、at.altavista.com
意大利语	欧洲	www.it.altavista.com、www.google.it、www.it.yahoo.com、www.lycos.it
土耳其语	亚洲	Tr.abacho.com、www.google.com.tr、www.arabul.com、www.altavista.com
波兰语	欧洲	www.wp.pl、www.onet.pl
波斯语	中东	www.google.com/intl/fa/、www.iranmania.com、www.google.co.in
马来语	东南亚	www.google.com.my、www.sajasearch.com、Malaysia.yahoo.com、searching.com.my
泰语	东南亚	www.google.co.th、www.yahoo.co.th
越南语	东南亚	www.google.com.vn、vn.yahoo.com
日语	东亚	www.yahoo.co.jp、www.infoseek.co.jp、www.google.co.jp、www.excite.co.jp
韩语	东亚	www.naver.com、www.korea.com
汉语	东亚	cn.search.yahoo.com、www.google.com.hk (www.google.com.cn)

在理论方面,语言模型(language model)^[32]、短语翻译(phrase translation)^[33]、语义匹配(meaning matching)^[34]、概率结构化查询(probabilistic structured query)^[35]等新兴技术被应用于跨语言信息检索研究,并取得了丰硕的成果。在应用方面,这一时期的特点是大型搜索引擎通过“语言工具”实现了跨语言信息检索。大多数搜索引擎都很重视对多语种的支持,纷纷推出适合不同语言的本地化搜索引擎,或者通过在线翻译机制对网页进行语言转换。Yahoo!曾于2005年7月发布跨语言搜索引擎Yahoo! Germany^[36]和Yahoo! France^[37],只限于德文和法文。全球最大的搜索引擎公司Google于2007年5月24日发布了其跨语言搜索引擎Google Translated Search^[38],支持包括英文、中文(简体、繁体)、阿拉伯文、法文、德文、西班牙文、俄文在内的多种语言,标志着跨语言信息检索技术正式走向实用化。世界各国各地区依据其所使用语言的不同,多种多语言搜索引擎各领风骚^[39],具体内容见表3。

作为最成功的跨语言搜索引擎,Google采用查询翻译的方法,将用户输入的查询由源语言翻译成目标语言后进行检索,最后将检索结果再翻译成源语言提供给用户浏览。可以说,Google跨语言引擎将多语言检索技术和机器翻译技术很好地结合在一起,这种信息服务方式一经推出便受到好评。遗憾的是,Google已于2010年3月23日凌晨退出中国大陆市场,其目前的中文服务改由香港站支持(www.google.com.hk),并将逐步实现完全退出。中国拥有世界第一的网络使用人口,Google约占中国搜索引擎市场35%,百度占有率为60%,Google退出中国内地市场,百度版图将持续扩张。另外,腾讯的搜搜、微软的Bing等引擎也在积极准备填补Google的中国市场份额,但无论是百度还是Bing,其跨语言信息服务能力都与Google相距甚远,中国巨大的跨语言信息服务市场枕戈待旦。

3 国家科技图书文献中心多语言信息服务

信息技术迅速发展促使多语种网络信息资源量以惊人的速度增长。中国国家科技图书文献中心(National Science and Technology Library, NSTL)收藏的多语科技资源也日益增长,截至2010年11月18日,英文库期刊、会议、学位论文、科技报告等的馆藏数量达到19,344,948篇,俄日文库的期刊论文数分别为

486,648 (俄文) 篇和1,191,046 (日文) 篇, 中文库期刊、会议、学位论文高达31,314,947篇, 这些多语种资源的数量将会持续快速地增长。为保证多语种信息资源的有效利用, 为NSTL用户提供有价值的多语言信息服务, 急需引入多语言信息检索与计算机辅助翻译系统, 以方便我国科研工作者使用NSTL的多语言文献资源。

在十一五科技支撑计划“科技文献信息服务系统应用示范”中示范应用汉英双向跨语言检索查询以及计算机辅助翻译技术, 其目标是实现基于NSTL、万方数据等海量科技文献信息资源的汉英双语检索服务, 满足广大科研人员对科技文献信息服务的需求, 为多语言信息服务技术的全面推广应用以及形成产业化产品生产奠定重要的基础。

为方便中国广大科技工作者获取英文科技信息, NSTL网络系统中目前多语言信息服务主要包括两部分: (1) 汉英双语跨语言检索词翻译推荐; (2) 英文文献标题和文摘的机器翻译辅助服务。随着研究与开发工作的不断开展, 更多外语跟汉语之间的跨语言检索词翻译推荐以及更多外文文献的标题和文摘翻译功能将加入到目前的应用系统中。

3.1 英汉双语跨语言检索

英汉双语跨语言检索词推荐通常有两种方法: 机器翻译的方法和机遇多语词典的查询翻译。与基于机器翻译的跨语言检索查询翻译方法相比, 基于词典的方法不依赖语言模型, 不依赖大规模并行语料, 具有简单易行、更新灵活、复杂度低等优点。项目的应用背景是NSTL科技文献检索平台, 高质量科技词典的作用更具优势。项目采用基于多策略的查询翻译接口设计方法, 主体结构如图1所示。

跨语言检索查询翻译和机器翻译紧密相关, 但又有所不同。汉英双向跨语言信息检索查询接口系统独立于检索平台, 并且基于多语言主题词表和双语言翻译频率词典。机器翻译要求翻译结果符合目标语言的语法规则, 而跨语言检索对翻译结果的顺序没有要求, 更注重对查询中关键术语和词汇的识别和准确翻译, 因为这关系到跨语言检索系统的查全和查准率。NSTL系统中的数字资源主要是科技文献, 因此对科技词汇的准确识别和翻译更加重要。

在跨语言信息检索过程中组块识别方法包括基于停用词的组块识别方法和中英文混杂查询识别方法。

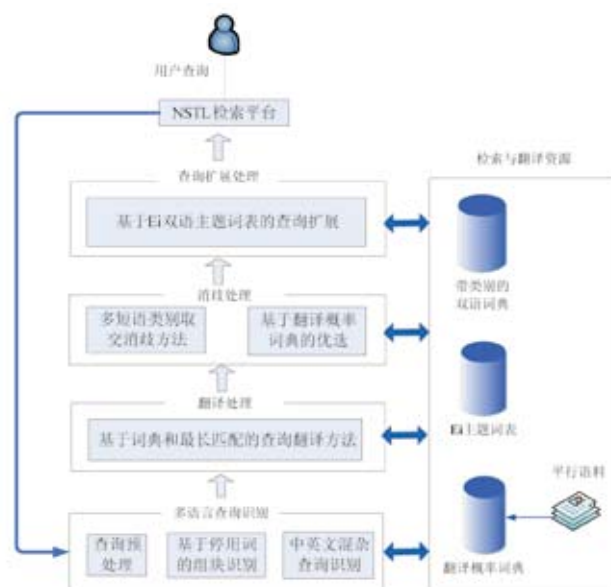


图1 基于多策略的查询翻译接口系统框架

(1) 基于停用词的组块识别方法将源语言查询中用停用词分隔的多个检索组块识别出来, 并分别进行翻译消歧处理, 然后取并集, 得到跨语言检索所需的目标语言检索词。对于新闻领域, 该方法不具一般性, 因为新闻类场景的停用词范围不固定, 在一个场合可以认为没有作用的词在另一个场合可能比较重要; 对科技领域而言, 停用词相对比较固定, 因此本方法比较适用。

(2) 中英文混杂查询识别方法适用于用户查询出现中英文混杂的情况。如果只是简单地将源语言查询中的目标语言词汇删除后进行查询翻译, 则会损失很多重要信息。因此, 我们使用异种语言词汇位置标记方法来解决该问题。

科技文献的术语通常由多个词汇组成, 而且相当一部分是非组合型的, 即不能够由其组成词推断出复合词的意义。科技术语的准确识别和翻译将直接关系到跨语言检索系统的准确率。在查询翻译中, 我们采用基于词典的最长短语识别和翻译方法。该方法与自然语言处理分词方法中的最大正向匹配算法颇为类似, 以查询中对应于词典的最长短语识别及翻译作为基础来实现全部用户查询的翻译处理。

在查询翻译时, 复合词语的翻译常常伴随多个候选翻译结果, 如何消除查询翻译歧义是关系到跨语言检索准确率的一个重要问题。我们采用多短语类别交集消歧和基于翻译概率词典的翻译优选两种策略。

(3) 多短语类别交集消歧利用带类别信息的英汉双语词典, 经过翻译模块获得查询中包含词和词组的译项及译项对应的类别信息后, 以最长短语为中心, 以距离最长短语最远的词和词组为半径, 取词和词组译项中类别信息一致的部分, 从而进一步消除查询翻译的歧义。

(4) 基于翻译概率词典的翻译优选策略依据翻译概率词典进行翻译优选。翻译概率词典的生成以并行语料为基础, 因此翻译词典中词汇的多个翻译结果的概率值在一定程度上能够反映某一领域中该词汇的使用习惯, 概率值越大则选择该翻译结果的倾向性越大。在原始查询由多个词或短语组成时, 分别计算多个词和短语的多个翻译结果的概率值的和, 取和最大的一个翻译组合作为该查询最终的翻译结果。

跨语言检索的查询扩展分为翻译前扩展和翻译后扩展两种方式。在NSTL跨语言信息检索接口中实现的是基于Ei双语主题词表的翻译后扩展。Ei主题词表是高质量的主题词表资源, 我们对Ei英文主题词表进行了汉化后, 选择有代表性的中文词汇与英文Ei术语对应。目前跨语言信息检索服务能够提供的扩展方式包括基于Ei结构的英汉和汉英查询翻译结果的上位词扩展、下位词扩展和相关词扩展。

为建立独立于搜索引擎的查询翻译接口, 首先实现了同NSTL检索平台的接口, 其工作流程如图2所示, 接口参数包括用户查询语言和查询本身, 查询翻译接口获取方式为页面直接抓取。



图2 嵌入NSTL的查询翻译接口

为满足跨语言信息检索服务在NSTL检索平台上的稳定运行, 查询翻译接口需要满足NSTL的检索压力需求。为此, 我们对提供查询翻译服务的服务器进行了压力位测试, 相关硬件配置及软件环境设置见表4, 测试结果如图3、图4和图5所示。

服务器实时动态监测的结果显示: 查询翻译测试

进行时, 服务器PE使用率约1.97G (为了提高查询翻译速度, 各种翻译词典在Tomcat启动时自动载入内存, 因此占据了一定的内存空间), CPU最大使用率接近100%。约40分钟的压力测试结果表明: 系统在100个并发用户的压力位下运行平稳, 基本满足NSTL检索平台的用户检索需求, 在应用层面说明了本项目查询翻译系统的可行性。

表4 压力测试环境

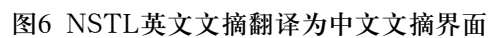
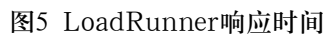
服务器硬件配置	CPU 2.33G双核 MEM 3.37G 硬盘500G
服务器操作系统	Windows 2003 Server
Tomcat服务设置	最大线程数调整为200, 将Tomcat JVM内存优化为1400M, 服务端口号由8080改为80
压力测试软件	LoadRunner9.0, 设置最大并发用户数为100, 采取的加压方式为每秒加载10个用户, 测试持续时间为40分钟

Period: 07/09/2010 17:49:07 - 07/09/2010 18:29:02	
Analysis Summary	
Scenario Name:	Scenario1
Duration:	39 minutes and 55 seconds.
Statistics Summary	
Maximum Running Users:	100
Total Throughput (bytes):	81,842,294
Average Throughput (bytes/second):	34,158
Total Hits:	28,386
Average Hits per Second:	11.847

图3 LoadRunner测试结果统计信息



图4 LoadRunner用户数信息



4 总结与展望

本项目课题的研究所形成的技术思路和技术积累（工具和数据）为进一步发展面向海量的科技信息资源的多语言信息服务技术奠定了重要的基础，多语言信息服务技术不仅是网络信息环境以及经济和科学技术全球化背景下不可或缺的数字化智能信息技术，同时多语言技术的发展也是通向语义技术必须的技术步骤。

- [1] NIE JIAN-YUSN. Cross-lingual and Multi-lingual IR [EB/OL]. [2010-03-18]. <http://www.iro.umontreal.ca/~nie/>.
- [2] OARD D W, DIEKEMA A R. Cross-language information retrieval [J]. *Annual Review of Information Science and Technology*, 1998(33): 223-256.
- [3] 刘伟成. 基于查询翻译的跨语言信息检索研究[D]. 武汉: 武汉大学, 2006.
- [4] 吴丹. 基于相关反馈的英汉跨语言信息检索查询翻译优化技术研究[D]. 北京: 北京大学, 2008.
- [5] GEY F C, KANDO N, LIN CHIN-YEW, et al. New Directions in Multilingual Information Access [C/OL]// SIGIR 2006. [2010-06-05]. http://www.sigir.org/forum/2006D/2006d_sigirforum_gey.pdf.
- [6] PIGUR V A. Multilanguage Information-Retrieval Systems: Integration levels and language Support [J]. *Automatic Documentation and Mathematical Linguistics* (English translation from Russian), 1979, 13(1): 36-46.
- [7] PEVENER B R. Automatic Translation of English Text to the language of the Pusto-Nepusto-2 System [J]. *Automatic Documentation and Mathematical Linguistics* (English translation from Russian), 1969, 3(4): 40-48.
- [8] PEVENER B R. Comparative Evaluation of the Operation of the Russian and English Variants of the "Pusto-Nepusto-2" System [J]. *Automatic Documentation and Mathematical Linguistics* (English translation from Russian), 1972, 6(2): 71-74.
- [9] SALTON G. Automatic processing of foreign language documents [J]. *Journal of the American Society for Information Science*, 1970, 21(3): 187-194.
- [10] SALTON G. Experiments in multi-lingual information retrieval [J]. *Information Processing Letters*, 1973, 2(1): 6-11.
- [11] ILJON A. Scientific and Technical Data Bases in a Multilingual Society [J]. On-line Review, 1977, 1(2): 133-136.
- [12] International Organization for Standardization (ISO). Guidelines for the Establishment and Development of Multilingual Thesauri [S]. Geneva, Switzerland: ISO, 1985.

- [13] LANDAUER T, LITTMAN M. Fully Automatic Cross-language Document Retrieval Using Latent Semantic Indexing [C]// Proceedings of the 6th Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research. Waterloo, Ontario, 1990: 31-38.
- [14] LANDAUER T, LITTMAN M. A Statistical Method for Language-Independent Representation of the Topical Content of Text Segments [C]// Proceedings of the 11th International Conference: Expert Systems and Their Applications. Avignon, France, Nanterre, France, 1991: 77-85.
- [15] RADWAN K, FOUSSIER F, FLUHR C. Multilingual access to textual database [C]// Proceedings of a Conference on Intelligent Text and Image Handling. Barcelona, Spain, 1991: 475-489.
- [16] MATEEV B, MUNTEANU E, SHERIDAN P, et al. ETH TREC-6: routing, Chinese, cross-language and spoken document retrieval [C]// Proceedings of Text Retrieval Conference. Gaithersburg, USA, 1996: 623-635.
- [17] PIRKOLA A. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval [C]// Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia, 1998: 55-63.
- [18] SCHAUBLE P. CLEF 2000 state-of-the art multilingual information access [C]// Proceedings of a conference on Cross-Language Evaluation Forum. Lisbon, Portugal, 2000: 1-21.
- [19] [OL]. [2010-03-21]. <http://www.sigir.org/>.
- [20] [OL]. [2010-03-21]. www.trec.nist.gov/.
- [21] [OL]. [2010-03-21]. www.aaai.org/home.html.
- [22] [OL]. [2010-03-21]. www.nlpir.nist.gov/tides/index.html.
- [23] [OL]. [2010-03-21]. <http://research.nii.ac.jp/ntcir/>.
- [24] [OL]. [2010-03-21]. www.clef-campaign.org/.
- [25] CAPSTICK J, DIAGNE A K, et al. MULINDEX: Multilingual Web Search and Navigation [C]// Proceedings of Natural Language Processing and Industrial Applications. Moncton, Canada, 1998.
- [26] [OL]. [2010-03-21]. <http://mulindex.dfki.de/>.
- [27] 王昊. 跨语言信息检索实现方法与关键技术探讨[J]. 情报检索, 2005(7): 46-49.
- [28] [OL]. [2010-03-22]. <http://cindorsearch.com/>.
- [29] HE D Q, OARD D W, WANG J Q, et al. Making MIRACLES: interactive translanguar search for Cebuano and Hindi [J]. ACM Transactions on Asian Language Information Processing, 2003, 2(3): 219-244.
- [30] OSTENERO F L, GONZALO J, VERDEJO F. UNED at iCLEF 2003: searching cross-language summaries [C]// Proceedings of the 3th Cross-language Evaluation Forum. Trondheim, Norway, 2003: 450-461.
- [31] [OL]. [2010-03-24]. <http://www.nstl.gov.cn>.
- [32] LARKEY L S, CONNELL M E. Structured queries, language modeling, and relevance modeling in cross-language information retrieval [J]. Information Processing and Management, 2005, 41(3): 457-473.
- [33] OSTENERO F L, GONZALO J, PENAS A, et al. Interactive cross-language searching: phrases are better than terms for query formulation and refinement [C]// Proceedings of the 2nd Cross-Language Evaluation Forum (CLEF'02). Rome, Italy, 2002: 416-429.
- [34] WANG J Q. Matching meaning for cross-language information retrieval [D]. Park: University of Maryland, 2005.
- [35] Darwish K, Oard D W. Probabilistic structured query methods [C]// Proceedings of 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Toronto, Canada, 2003: 338-344.
- [36] [OL]. [2010-03-24]. <http://de.docs.yahoo.com/translator/>.
- [37] [OL]. [2010-03-24]. <http://fr.docs.yahoo.com/translator/>.
- [38] [OL]. [2010-03-24]. http://www.google.com.hk/language_tools?hl=zh-CN/.
- [39] [OL]. [2010-03-21]. <http://www.forexporter.com/news309.htm/>.

作者简介

张满年, 中国科学技术信息研究所副所长, 从事信息资源管理研究。E-mail: zhangmn@istic.ac.cn

高影繁, 博士, 计算机应用专业, 主要研究方向: 跨语言信息处理。E-mail: gaoyingf@istic.ac.cn

徐红姣, 硕士, 信息管理专业, 主要研究方向: 跨语言信息检索。E-mail: xuhj@istic.ac.cn

张均胜, 博士, 计算机软件与理论专业, 主要研究方向: 多语言信息服务, 语义计算。E-mail: zhangjs@istic.ac.cn

王惠临, 研究员, 博士生导师, 主要研究方向: 多语言信息服务, 机器翻译, 自然语言处理。E-mail: wanghl@istic.ac.cn

Multilingual Information Service in National Science and Technology Digital Library of China

Zhang Mannian, Gao Yingfan, Xu Hongjiao, Zhang Junsheng, Wang Huilin / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: How to provide multilingual information service effectively has become an important research issue in scientific and technical information service domain. This paper introduces the development of multilingual information service research, and then focuses on introducing the cross-language information retrieval and machine translation services in national science and technology digital library of China. Cross-language information retrieval and online machine translation for abstracts of literature are firstly used in science and technology digital library in China, and this could help promote the applications of multilingual information service in digital libraries in China.

Keywords: Multilingual information service, Cross-language information retrieval, Digital library, Machine translation

(收稿日期: 2011-11-03)