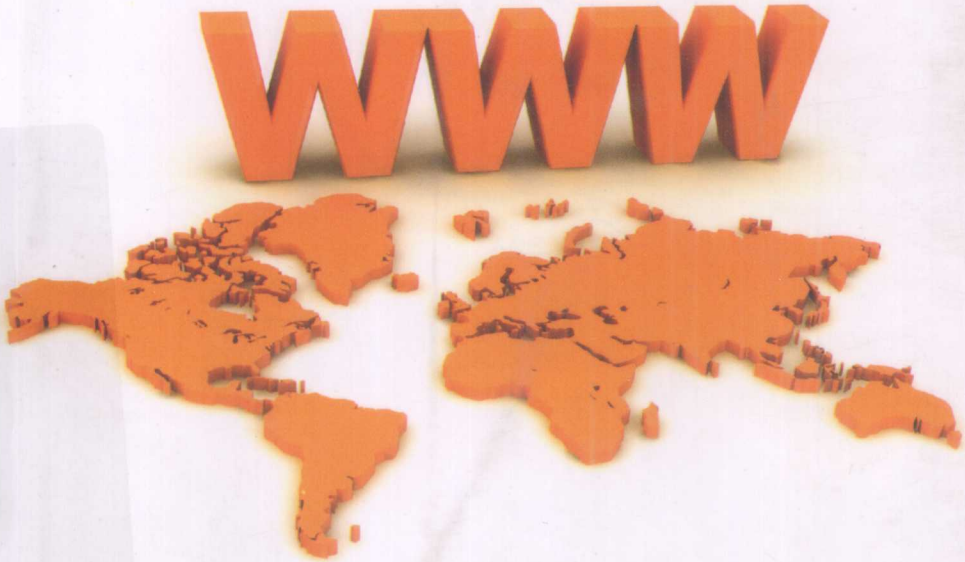
信息检索与信息服务系统方向

跨语言信息检索

刘伟成 著



海南出版社

信息检索与信息服务系统方向

跨语言信息检索

刘伟成 著

海南出版社

图书在版编目(CIP)数据

跨语言信息检索/刘伟成著.—海口：海南出版社，2009.10
ISBN 978-7-5443-3134-0

Ⅰ.跨… Ⅱ.刘… Ⅲ.情报检索—研究 Ⅳ.G252.7

中国版本图书馆CIP数据核字（2009）第177620号

信息检索与信息服务系统方向

跨语言信息检索

刘伟成 著

责任编辑 万 胜 封面设计 夏 仁

※

海南出版社出版发行

(570216 海口市金盘开发区建设三横路2号)

湖北科学技术出版社黄冈印刷厂印刷

2010年1月第1版 2010年2月第1次印刷

开本:787×1092毫米 1/32 总印张:7.5 字数:25万

书号:ISBN 978-7-5443-3134-0

定价:35.00元

前　言

多语性是网络社会交流的重要特征之一，如何将网络资源，用不同语言提供给需要的用户，用可互换语言形式描述相同及类似内容的信息资源，是跨国界跨专业领域信息交流的重要课题。跨语言信息检索（cross language information retrieval，CLIR），就是解决以某种语言检索另外一种语言表达的文献资源，为近年来国内外很活跃的研究课题。在互联网的发展和经济技术全球化的趋势下，这一领域将获得更多的关注和更长远的发展，跨语言信息检索必将成为网络信息检索中的重要组成部分。

依据2006年ETHNOLOGUE目录上的统计，全世界语言数目高达6，912种（http://www. ethnologue. com/，2006）。目前，中国的人口总数大约占世界人口总数的22%，上网人口总数却仅占世界的10%左右，中文网络文献量更少，中文网页仅占世界网页总数的2%。英语等西方语言仍然是网络世界中的常用语言。研究跨语言信息检索的理论与方法，对拥有众多人口的中国网上用户了解世界，学习与从事科学研究有着十分重要的应用价值。

本书的研究，就是试图从宏观上揭示跨语言信息检索的基本原理和模式，并以基于查询翻译的跨语言信息检索为研究对象，深入分析和研究了语言转换前的文字信息预处理、语言转换中的语言歧义问题。本研究还提出了跨语言本体的概念，跨语言本体的构建及其在查询翻译消歧中的应用，并以典型案例对基于查询翻译的跨语言信息检索的几种主要模式进行了剖析和对比，为跨语言信息检索的进一步深入研究提供了现实依据。

跨语言信息检索的研究涉及语言学、情报学、计算机科学等多

门学科知识，本书的研究着重于情报学和信息检索领域，在一定程度上丰富和拓展了网络环境下信息检索的理论体系。

本书共分8章，主要内容如下：

1. 引言

本章主要论述跨语言信息检索研究的背景和意义，并对国内外的研究现状进行了详细的分析。

2. 跨语言信息检索概述

跨语言信息检索（CLIR）涉及查询语种和检索语种两个基本的概念，查询语种是用户查询请求所属语种，检索语种是检索目标对象所属语种，如何能够在这两者之间建立起沟通的桥梁是目前跨语言信息检索技术研究最核心和关键的问题。在深入研究基于查询的跨语言信息检索之前，本章将首先探讨跨语言信息检索的一般模型，对跨语言信息检索的类型、应用领域进行了细致的归纳与详细的阐述，并分析了跨语言信息检索与机器翻译的联系与区别，从而在总体上概括和总结跨语言信息检索近十几年来的研究与发展。

3. 跨语言信息检索模型

只要谈到信息检索就必然涉及检索模型，这里以跨语言信息检索为标题，还有另外一层含义，即这些模型在查询语言转换以及查询翻译消歧中还将发挥重要作用，比如国外的许多学者已经探讨了扩展布尔模型（weighted Boolean models）、语言模型、本体模型等在跨语言信息检索和翻译消歧中的应用，Dumais 等人将潜语义模型（LSI）、Carbonell 等人将广义向量空间模型（GVSM）分别应用于跨语言信息检索，在不需要翻译的条件下实现了跨语言信息检索，为跨语言信息检索的研究开辟了一条新路。本文将其归纳为5种主要模型：布尔模型，向量空间模型，概率模型，统计语言模型，本体模型。并分别从模型概念与相关文档判定方法、相关性计算两方面对经典模型进行介绍，对它们各自主要派生模型作简单

概述，并针对不同模型之特征，论述其在跨语言信息检索中的应用。

4. 语言转换前的文字信息预处理

当前，基于查询翻译实现 CLLR（query translation based CLIR）已成为一种最为流行的技术，如果结合实现过程中所利用的简单语言学处理，如词性标注或者短语索引等等，根据国外学者的试验证明，该方法一般能够达到相应单语检索效率的 50% 到 75%。在进行查询语言转换之前还要进行一些前期的文本预处理，比如语言识别（Language Identification）、信息抽取（Information Extraction）、分词（segmentation）、信息标引（Information Indexing）等。对于不同的语言来说，预处理的内容也是不尽相同的。如英语、德语、俄语等，词与词之间有明确的界限，基本不存在分词的问题，但同时这些语言属于曲折语，即存在词的形态变化，需要形态还原；汉语属于孤立语，没有词的形态变化，但却存在着书写时不区分词而连写的形式，所以分词也就成为中文信息处理最基本也是最首要的任务；日语也存在分词问题，但由于它属于有构形附加语素的黏着语，其分词处理要比汉语简单，词形变化分析比英语简单。信息预处理是每一个跨语言信息检索系统必备的功能和模块，本章以汉语、英语为例分析文本信息预处理的方法、成果及其在跨语言信息检索中的应用。

5. 语言转换中的语言歧义问题

语言歧义问题影响到基于查询翻译的跨语言信息检索效率。本章首先总结了语言转换中存在的 4 个主要问题，即语言的歧义性、未登录词、短语的识别和翻译、翻译资源中的错误。并提出了几种解决的方法：基于词性标注的词义消歧，基于主题词表的词义消歧，虚拟语境消除目标查询的多义性，基于互信息的词义消歧，基于浅层句法分析的短语识别，基于查询扩展的词义消歧，基于语料库的词义消歧，专有名词的音译，基于机读词典的词义消歧等。

6. 跨语言本体的构建及其在语言转换中的应用

本章阐述了本体的定义、组成、分类，以及在查询翻译消歧中的应用。并首次提出了跨语言本体的概念，以 EuroWordNet（欧洲词网）、CCD（中文概念词典）、Sinica BOW（中英双语知识本体词网）为例对跨语言本体的构建方法进行了探索性研究，最后以一个中英双语知识本体与领域检索接口雏型为例对其在跨语言信息检索语言转换中的应用进行了分析。

7. 跨语言信息检索系统评价

检索系统的性能评测对于检索系统的研制和发展是至关重要的。对跨语言信息检索（CLIR）系统的评价基本上采用了与一般信息检索系统评价相同的方法和步骤。在单语言信息检索（MIR）试验中，研究人员通常保持测试主题和文献集合不变，而改变检索系统以比较不同检索系统之间的性能。然而，CLIR 评价通常改变测试主题而不是检索系统，以比较相同系统下 MIR 和 CLIR 的检索性能（采用相同的检索系统和检索条件），作为评价 CLIR 系统性能的最重要的指标。CLIR 系统评价需要覆盖多种语言的评价测试集。CLIR 评价测试集通常包含三个部分：包含至少两种语言以上的测试文档集合；包含与文档集不同语言的检索问题集合；检索问题的正确答案集合。本章对检索性能的各种评测方法以及现有的测试文档集作了较详细的介绍。在检索性能评测方法中主要介绍了查准率和查全率以及由此派生出的调和中数和 E 测度方法。在对现有测试文档集介绍部分中重点介绍了影响比较大的 TREC，NTCIR 以及 CLEF。这些都为下一章的案例分析与比较提供依据。

8. 基于查询翻译的跨语言信息检索系统案例

有关跨语言信息检索效率的最早试验结果，是在 1969 年由 Cornell 大学的 Salton 所记录，Salton 利用手工编制的叙词表实现了受控语言的跨语言信息检索。但那时的研究主要是针对国际联机检索进行的，而当时联检系统并不普及，国际互联网尚不为人们

所知，因而人们对网络信息的需求并不强烈。跨语言信息检索研究的真正活跃期，是在 Internet 迅猛发展的 90 年代，国际上先后有许多相关论文发表，一些试验性跨语言信息检索系统相继问世。

本章将基于查询翻译的跨语言信息检索归纳为 4 种主要模式（机制）：基于机器翻译系统的方法、基于语料库的方法、基于双语词典的方法和基于本体的方法。并选择 5 个典型案例进行了深入剖析，在分析的基础上，对各种查询翻译机制的原理、特点、应用等进行了比较，通过分析可以看出这些试验系统或商业系统基本上都采用了混合的方法（以某一方法为主），有效结合了不同语言转换机制的优点，这也可以看作是一个发展的趋势和研究的方向。

相对于单语种信息检索，跨语言信息检索涉及的学科门类广，综合性强，是一个富有挑战性的研究领域。目前国外对跨语言信息检索的研究方兴未艾，但在国内这方面的研究相对薄弱，检索准确率还比较低。由于时间仓促和作者水平的限制，错误和缺点在所难免，希望能抛砖引玉，以推动跨语言信息检索理论和技术的发展成熟，实现“跨越时间、空间、语言障碍”的信息世界。错误、疏漏之处，肯请读者批评指正。

本书的研究和撰写得到来自多方面的帮助和支持，感谢恩师焦玉英教授对本书篇章结构及内容观点的辛勤指导与认真审核，感谢武汉科技大学管理学院的领导所给予的多方面关心与鼓励，感谢海南出版社万胜编审的热情帮助，他们为本书的出版付出了辛勤劳动，在此表示诚挚的谢意。同时，本书的出版受到武汉科技大学优秀青年教师专项资助项目（06007901）和湖北省教育厅科学研究计划优秀中青年人才项目（Q20081108）以及教育部 2009 年度人文社会科学研究一般项目“基于多语言本体的跨语言信息检索系统及其应用研究”的资助，在此一并表示感谢。

作　　者

2009 年 12 月于武汉科技大学

目　录

1 引　言

1.1 跨语言信息检索研究的背景和意义

1.1.1 现实的背景

多语性是网络社会交流的重要特征之一，如何将网络资源，用不同语言提供给需要的用户，用可互换语言形式描述相同及类似内容的信息资源，是跨国界跨专业领域信息交流的重要课题。跨语言信息检索（cross language information retrieval，CLIR），就是解决以某种语言检索另外一种语言表达的文献资源，为近年来国内外很活跃的研究课题。

新一代信息传播的特色是：国际互联网突破空间距离，打造一个不分国界的信息地球村。尤其透过全球信息网，各地的信息皆唾手可得，不但丰富且实时。在网际网络上流通的信息除了数量非常庞大之外，所使用的语言种类也非常多。依据 2006 年 ETHNOLOGUE 目录上的统计，全世界语言数目高达 6,912 种（http://www.ethnologue.com/，2006）。在现实世界中，语言按使用人口数排名，前几名依次为中文、英文、印度文、西班牙文、葡萄牙文、孟加拉文、俄文、阿拉伯文、日文。但根据 2001 年 3 月份的统计估算，在网络世界中，语言使用的人口数，前几名依次为英文（47.5%）、中文（9.0%）、日文（8.6%）、德文（6.1%）、西班牙文（4.5%）、韩文（4.4%）、法文（3.7%）、意大利文（3.1%）、葡萄牙文（2.5%）、俄文（2.1%）。而国际互联网中信息内容所使用的语言比例，前几名依次为英文（68.39%）、日文（5.85%）、德文（5.77%）、中文

(3.87%)、法文(2.96%)、西班牙文(2.42%)、俄文(1.88%)、意大利文(1.56%)、葡萄牙文(1.37%)、韩文(1.29%)。

另据2002年10月的统计,在使用英文搜索引擎中提出语言翻译请求的几种主要语种有:西班牙文47.2%、法文17%、拉丁文7.8%、德文6.2%、日文4.7%、意大利文3.2%、俄文2.4%、中文2%①。据网站 http://www.vilaweb.com 报道:当前全世界3130亿网页内容所使用的语言依次为英文68.4%、日文5.9%、德文5.8%、中文3.9%、法文3.0%、西班牙文2.4%、俄文1.9%、意大利文1.6%、葡萄牙文1.4%、韩文1.3%、其他文种4.4%。因此,人们要想广泛地实现网上信息资源共享,对语言自动翻译的需求越发迫切。为了消除网络资源利用中的语言障碍,跨语言信息检索技术(Cross - Language Information Retrieval,CLIR)成为当前信息检索领域中重要的研究课题。越来越多的研究团体深入研究跨语言信息检索问题,并研制开发跨语言信息检索的不同方法。

跨语言信息检索是指用户以一种语言提问,检出另一种语言或多种语言描述的相关信息的方法。过去这项研究,英文使用的名称非常混乱,直至1996年在ACM SIGIR Workshop for Multilingual Information Retrieval会议上,经与会人员讨论,将其定名为Cross - Language Information Retrieval。几乎在同一时间,美国Defense Advanced Research Project Agency (DARPA),也将这项研究给予另一种称呼:Translingual Information Retrieval。不管是哪种称呼,其研究目标一致,都是希望在多语性的信息时代,提供跨语言的信息检索服务。

在跨语言信息检索的研究上,近几年有多项国际会议举办专题讲座(Chen, 1997, 1998; Hovy and Idel, 1998; Grefenstette,

① http://www.translate-to-success.com/language-translation.html(Accessed sep. 2009)

1998)、甚至举办专题国际会议(Grefenstette, 1996; Oard, 1997; Vossen, 1997)、数字元图书馆系列论文(Borgman, 1997; Oard, 1997; Powell and Fox, 1998)。著名的计算语言学和信息检索领域国际会议,如 ACL Annual Meeting, ACM SIGIR99(SIGIR00)等,都有特别的议程探讨跨语言信息检索的发展。ACM SIGIR02 有一个研讨会,由三个主要跨语言信息检索评比组织(TREC、CLEF、与 NTCIR)共同规划,讨论了未来几年的评比重点。

文本检索会议(Text Retrieval Conference, TREC),是由美国国家技术标准局组织召开的国际会议,其旨在促进大规模文本检索领域的研究,加速研究成果向商业应用的转化,促进学术研究机构、商业团体和政府部门之间的交流与合作①。跨语言信息检索是在第六届文本检索会议(TREC - 6)评价中建立的一项新任务。与以往 TREC 评价中关于西班牙语、汉语等信息检索的多语种任务比较,CLIR 具有一种优势——对于多语种文本集合,只要求用户提交一种语言的查询,而不必提交以每一种语言书写的各种查询。在 2000 年举行的第九届文本检索会议(TREC - 9)的 CLIR 任务评价中,第一次引入汉语作为文本描述语言。

1.1.2 研究意义

(一)理论意义

信息检索是一个发展非常迅速的研究领域。自从 20 世纪中期情报检索的概念被提出,距今已有半个多世纪的历程,期间情报检索的理论不断得到拓展、延伸和丰富,已不再是仅仅以布尔模型、概率模型等为理论基础,而是逐渐引入遗传算法、并行算法、粗糙集理论等构建智能检索模型;检索手段也不断更新换代,从传统手工检索发展到单机检索、联机检索,现在则是网络检索的时代;检索对象已从文本扩大到多媒体信息;单语检索已发展到多语言

① http://trec.nist.gov(Accessed sep. 2009)

检索。多媒体、多语性、多文化性是互联网的三个基本特色，如何跨越媒体、语言、文化的障碍，实现跨国界的交流，既是一个非常重要的技术问题，也需要理论研究来指导检索实践。

一方面跨语言信息检索的研究集语言学、情报学、计算机科学等多门学科知识，是一个综合性强、富有挑战性，且长期以来被国内外学术界十分关注但有诸多难题的研究领域，通过对这一领域中涉及的文本语言及语音识别技术、汉语言分词、信息标引与提取技术、查询翻译中的语言歧义及其消歧等的研究，对构建网络环境下信息检索理论体系，进一步丰富与拓展其研究内容，促进网络环境下情报学理论创新发展有重要的理论意义；另一方面，通过全面、系统地总结十余年来国内外在跨语言检索研究中的理论和实践成果，对网络环境中跨语言信息检索系统的设计与推广应用具有启示和现实指导作用。

（二）实践意义

2009 年 1 月，中国互联网络信息中心（CNNIC）发布第 23 次《中国互联网络发展状况统计报告》，声称中国的网民总数已经达到 2.98 亿人，网民总数全球第一，普及率达到 22.6%，超过全球平均水平。另据国家信信息化办公室统计，截止到 2004 年底，网上共有中文网页 6.5 亿，全世界为 300 亿，中文仅占 2%。而英语仍然是网络世界中的常用语言。研究中英文跨语言信息检索的理论与方法，对拥有众多人口的中国网上用户了解世界，学习与从事科学研究带来便捷，有着更为重要的应用价值。

主要体现在下列几个方面：

（1）跨语言的信息存取已经成为人们工作生活的一部分

随着经济全球化的快速发展，以及电子商务（EC）、网上交易、虚拟企业等商务形式的产生，许多国际性组织、跨国企业集团都需要进行跨多种语言的信息检索和组织，以用户的母语输入并以用户熟悉的语言呈现给大家，使来自不同国家、使用不同语言的人可