计算机科学2002Vol. 29№. 8

一种新型的跨语言信息检索技术*)

A New Type of Cross-Language Information Retrieval Technique

张玥杰　连　理　吴立德

(复旦大学计算机科学与工程系　上海200433)

Abstract The traditional Information Retrieval(IR)system is mainly implemented for monolingual document collections. Generally, information retrieval systems use the most familiar language for users as query language. However, with the ever growing amount of information available to us all, situations when a user of an information retrieval system is faced with the task of querying a multilingual document collection are becoming increasingly common. Thus a very important problem is occured, which is the matching of user queries specified in one language against documents written in a different language-crossing the language boundary, namely Cross-Language Information Retrieval (CLIR). In current information society, cross-language information retrieval has become a pivotal question, which needs to be solved quickly in word. It is however, the global information infrastructure of the Internet that has been largely responsible for the growing awareness of a need for cross-language information retrieval systems. In this paper, we will present the cross-language information retrieval technique in more detail. We hope the introduction will be helpful to the other researchers in the field of Chinese and foreign language information proessing.

Keywords Information retrieval, Cross-language information retrieval, Machine translation, Corpus, Linguistics

1. 前言

随着科学技术迅猛发展，信息交流量与日俱增，人们开始广泛应用高速度、大容量的现代化工具——计算机进行信息处理。为使计算机能够应用于更广泛的用户，利用计算机高效率地进行各种语言信息处理已成为一个迫切需要研究的课题。由此，语言信息处理应运而生，成为一门新兴学科，其相关理论和方法研究在计算机科学与人工智能领域也显得尤其重要。自动的信息检索(Information Retrieval, IR)也作为语言信息处理研究领域的重要课题，越来越引起人们的兴趣与重视。人们希望用机器来实现信息自动检索，以解决人工方式带来的困难与复杂，如今正随着人们在语言信息处理领域里所取得的成果而变为现实。

信息检索泛指用户从包含各种信息的文档集中找到所需要的信息或知识的过程。传统的信息检索系统主要是针对单一语种的文档集实现，一般是使用用户最为熟悉的语种作为查询语言。而随着日益增长的大量信息成为可利用的，用户面对查询一个多语种文本集合的情形，变得越来越普遍。这就产生一个非常重要的问题—以一种语言描述的用户查询与以不同语言书写的文本之间的匹配问题，也就是一种如何跨越语言界限的问题，即跨语言信息检索(Cross-Language Information Retrieval, CLIR)。这些多语种文本集合通常是由来自多国公司本地办公室文本组成，或者是由来自多语种国家不同区域的文本组成(例如，瑞典、加拿大等国)，或者是由来自大型国际组织的文本组成(例如，联合国、欧共体等)。当然，环球网也是这种文本集合的一个样本。

对于利用这类多语种数据资源的大量用户来说，他们必须具备一些外语知识。但是，这些用户对于某些外语的精通程度并不足以明确地叙述查询，以正确地表达他们自身的信息需要。如果用户能够以本国语言输入查询，那么他们将受益匪浅。因为，用户能够对文本进行检查，即使这些文本未加以翻译。另一方面，单一语种的用户可使用人工或者完全自动的辅助翻译，来帮助他们访问到搜索结果。可以看出，在当今信息社会中，跨语言信息检索已成为世界范围内一个亟待解决的关键问题。然而，在很大程度上，Internet 的全球信息基本结构造成针对跨语言信息检索系统的迫切需要。这就导致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。

本文将对跨语言信息检索技术作较为详细的介绍，希望通过这些介绍，使我国进行汉语及外国语言信息处理的研究者有所借鉴。

2. 信息检索的一般模型

信息检索系统的目标在于向用户提供一系列满足其信息需求的项。其中，信息需求即为“查询”，而对于所选择的项即为“文档”。因此，可将信息检索任务看作为，在给定用户查询之后，从文档集中识别出最为匹配的文档。一般来说，每一种信息检索方法具有两个组成部分：1)表示文本的某种技术(即，查询与文档)；以及2)比较这些表示的某种途径。其最终目的就是，通过计算查询表示与文档表示之间的比较，而自动完成检查文档的过程。当通过上述过程所产生的结果，类似于通过人工比较查询与文档所产生的结果时，则该自动过程，即“信息检索过程”成功实现。

对于基本的信息检索模型，经常加以扩展操作，以解决在查询与文档特征之中所存在的差别。例如，查询经常非常短(可能只具有一个或者两个单词的长度)，而文档很可能为几

*)本文受国家863基金(编号：2001AA114120)和复旦大学青年科学基金(2001年)资助。张玥杰　博士，主要研究领域为机器翻译、中文信息处理及相关技术。连　理　硕士研究生，主研究领域为中文信息处理及相关技术。吴立德　教授，博导，主要研究领域为计算语言学、计算机图形学及相关技术。

百页长。另一个问题是，用户经常采用显著不同于文档的词汇，而这些文档中包含着用户所寻求的信息，即为“释义问题”。信息检索系统调和这些差别的一种途径，就是通过构造表示函数来解决，表示函数以不同方式处理查询与文档，从而达到一致表示的目的。跨语言信息检索作为“释义问题”的一种特殊情形，上述方法也为其奠定了解决问题的基础。

信息检索的过程及组成部分如图1所示。

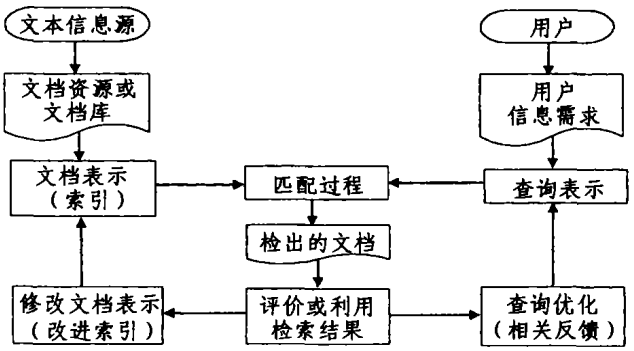

图1　信息检索的一般模型及检索过程

图中的模型主要包括：

·文档模型——即文档的索引，也就是文档内容的识别和表示，包括语义内容和上下文属性（如作者、编者等）。

·查询模型——即用户需求信息的获取与表示。

·匹配函数——即在文档表示和查询表示的基础上，定义查询和文档的相关程度函数。

·性能评价——一般采用精度（Precision）和检索率（Recall）对检索出的文本进行评价，处理速度有时也用于评价系统的效率。

·反馈修正过程——根据检索出的结果对查询表示（少数情况下也对文档表示）进行扩充与参数优化，以提高系统性能。

如果采用一种形式化描述，则上述模型可表示如下：

·给定查询集合 Q，查询表示函数为 q，其范围为 R，即统一的文本（查询与文本）表示空间；

·给定文档集合 D，文档表示函数为 d，其范围为 R；

·比较函数（匹配函数）为 c，其定义域为 R×R，其范围为 [0,1]，即0与1之间的实数集合。

在一个比较理想的信息检索系统中，

$c(q(query),d(doc))=j(query,doc),\forall\ query\in Q,\forall\ doc\in D$，

其中，$j:Q\times Da[0,1]$表示用户对于两种文本之间所存在的某种关系的判断，例如，基于内容相似性或者风格相似性进行测量。

任何一种信息检索模型都有其理论基础和一组假设，检索模型的一些普遍性假设包括：

·被检索对象主要为文档对象。

·对象的检索与其它对象是否被检索出无关，具有独立性（按类别检索时显然不合适）。

·检索是根据文档内容的表示及所需信息的表示进行的。

·文档内容和所需信息的表示都是非精确的。

60年代中期以来，人们提出了大量检索模型。自最初的为一些较小的和较为结构化的文档所设计的特殊模型（如文献记录，包括题目、作者和主题码等），发展到现在具有较强理论基础和能处理多种文档格式的模型。当前的模型能够处理具有复杂内部结构的文档，并且一般都具有学习和利用相关反馈进行查询优化等功能，使得系统性能大大提高。当前应用中最主要的三个模型是：

·严格匹配模型——是许多商业信息检索系统的理论基础。

·概率模型——把检索看作是文档表示和查询之间匹配程度的概率估计问题。

·向量空间模型——把文档和查询看作是多维向量空间中的向量，用距离作为相似度的一种度量方式。

实验表明，后两种模型的许多性能优于严格匹配模型，但应用到商业产品上只是近几年的事情。不同的模型有不同的理论基础和性能特性，在检索效率和计算复杂性上也有所区别，但所有的模型都要计算相似性。

3. 跨语言信息检索的方法

所谓跨语言信息检索，其关键之处就是通过使用以单一语言书写的查询，如何从利用不同语言书写的多语种文本集合中检索相关文本。这也许可以通过翻译用户查询、翻译文本、或者将用户查询与文本两者翻译为某种媒介物或者中间语言表示来解决。例如，一种可选择的解决方法就是，将用户查询与文本通过人工或者自动方法翻译为一种通用的受控索引词汇表。然而，对于更为一般的方法，其中利用整个文本与用户查询相匹配。

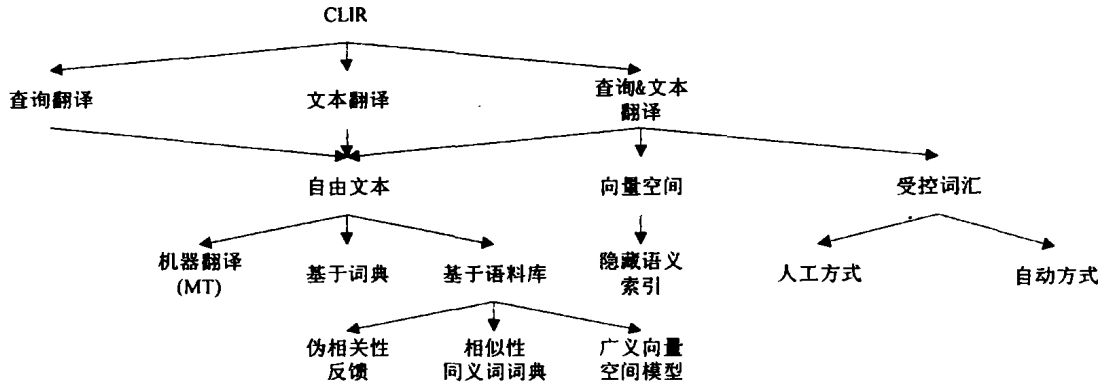

图2　跨语言信息检索方法的分类

根据在跨语言信息检索过程中所使用的资源，可将跨语言信息检索方法分为以下三类：

·基于机器翻译（Machine Translation，MT）的 CLIR 方法；

·基于机器可读词典（Machine Readable Dictionary，MRD）的 CLIR 方法；

·基于语料库(Corpus)资源的CLIR方法。

根据在跨语言信息检索过程中所处理的对象,可将跨语言信息检索方法分为以下三类:

·基于查询翻译(Query Translation,QT)的CLIR方法;

·基于文本翻译(Document Translation,DT)的CLIR方法;

·基于查询与文本翻译(QT&DT)的CLIR方法。

图2描述了跨语言信息检索方法的分类。

3.1 基于机器翻译的CLIR方法

对于CLIR,MT技术是一种显然的选择。基于MT的CLIR方法是跨语言信息检索问题的最为直接的一种解决方式。最近几年,已多次提出该方法[1~4],即通过使用完全自动的机器翻译系统对查询或者文档进行处理,从而基于单语表示来表达查询与文档。

对于目前所存在的完全自动的机器翻译系统,通常都具有一个弱点,就是仅在受限领域中产生较高质量的翻译。Fluhr观察到,同语义上的翻译错误相比,信息检索系统更加能够容忍句法上的翻译错误[2]。但是,对于一个翻译系统,在编码领域知识不充足的情况下,其语义准确率将会受到影响。因此,从Fluhr的观察中可以看出,由于编码领域知识代价非常高,对于基于机器翻译的CLIR方法,其效率将会受到限制。

当前,基于查询翻译实现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,该方法一般能够达到相应单语检索效率的50%到75%。当处理短查询时,查询翻译策略相对较为有效。但是,由于在仅包含少量单词的查询中缺乏足够的上下文信息,从而限制了系统为查询项选择最为适合翻译的能力。机器翻译系统试图利用整个文档之中的上下文线索,来产生最可能的翻译。由此,通过翻译文档而不是查询,可能会在部分上缓和该问题。因为文档长度通常远远大于查询长度,则嵌入于文档处理过程中的机器翻译系统,同嵌入至查询处理过程中的机器翻译系统相比,将会具有更多的上下文信息作为进行语义选择的基础。而且,如果语义选择的控制模式适合,则信息检索系统一般可以容忍偶尔发生的语义错误。较长的文档通常包含一个较大的词汇表,则该词汇表可以用于改进选择正确语义的控制模式。

然而,由于文档处理过程必须应用至大量的文档中,当将翻译系统嵌入至文档处理过程中时,机器翻译的效率就会成为问题所在。而且,通过机器翻译系统所完成的一些工作,并未提高检索效率。例如,翻译过程要求选择单词顺序,并在目标语中加入一些禁用词。但是,已在文档与查询处理过程中将这两种特征删除。

事实上,通过机器翻译系统所完成的一些工作,可能会降低有关检索效率的一些度量指标。因为在不同语言中,不能利用相同的方法分组词义,机器翻译方法将试图为所使用的多义词确定其最可能的词义。按照这种分析,为每个多义词选择唯一一种词义。然而,在信息检索系统中,可设计文档与查询表示函数,以保存关于不确定性的信息,同时设计匹配函数以提高效率。作为这种情形的一个实例,即为严格匹配信息检索系统,其中用多义词的每种可能翻译替换单词本身,从而提高查全率(以查准率为代价)。而对于一些基于概率的信息检索系统,能够表示并利用关于多义词的每种词义为正确的概率信息。如果从机器翻译系统中能够抽取以上这种概率信息,就可以通过提高查全率来改进平均查准率,而限制了对于查准率所造成的副作用。

3.2 基于机器可读词典的CLIR方法

对于CLIR中所使用的机器可读词典,也可将"同义词词典"作为其广泛采纳的一种定义。同义词词典是一种组织语项的工具,通过其可以编码领域知识,以适用于具体应用中。多语同义词词典是指从多种语言中组织语项。由该定义可以知道,双语词典一般定义与一种语言的语项相对应的另一种语言的语项。当然,也包含计算语言学中编码有关语项的句法和语义信息的词典。在自动的信息检索系统中,作为概念索引的复杂同义词词典,也属于同义词词典定义的范畴之内。甚至一个简单的技术语项双语列表,其中每个语项对应一种翻译,也可称之为同义词词典。表1描述在CLIR中所使用的同义词词典类型。

表1 多语同义词词典的类型及其特征

同义词词典类型	特征
主题同义词词典	分层与关联关系; 每个节点分配唯一的一个语项。
概念列表	划分为概念类别的语项列表。
语项列表	跨语言同义词列表。
词典	机器可读的句法与/或者语义。

基于同义词词典的技术存在一些优势,但同时也包含一些局限性。因为同义词词典可通过一种人类可理解的方式表示语项和概念之间的关系,从而基于同义词词典的信息检索允许用户利用在搜索过程中所获取的信息,来更好地重新阐述查询。而且,因为在同义词词典中编码了大量的领域知识,对于一个有经验的用户来说,基于同义词词典的信息检索系统将会成为一个强有力的工具。另一方面,使用同义词词典可能会对用户所利用的词汇表,以及应用信息检索系统的领域,事先加以一些限制。目前,存在多种同义词词典的构建与维护技术,需要做出努力以有效地使用复杂同义词词典中所包含的概念关系。在同义词词典中,可编码领域知识的几个方面。每一种多语同义词词典的关键特征在于,跨语言同义词的规范。在较为复杂的同义词词典中,一般包含分层概念关系以及关联关系。

对于同义词词典,可通过人工方式或者自动方式加以使用。在所谓的"受控词汇"系统中,利用唯一一个描述语项标记每个概念,以便用户可人工规定其查询中适合的概念。当自动使用同义词词典中所编码的概念关系时,即为"概念检索"技术。在简单的概念检索系统中,可通过使用一个概念列表,以概念类别替换每一个语项,从而提高查全率(以查准率为代价)。一种更为复杂的方法——"查询扩展",使用同义词词典中所编码的概念关系来选择语项,以达到同时提高查全率与查准率的目的。概念替换与查询扩展试图通过减少释义问题的影响,而提高查全率。通过在同义词词典中包含句法或者语义信息,来减轻歧义的影响,从而提高查准率。例如,在受控词汇系统中,同义词词典经常提供语义信息,以帮助用户人工选择正确的语项。而对于概念检索,通过利用一些单词的词性进行标注,从而选择适合于该词性的翻译。

3.2.1 早期的研究工作 Pigur描述了一种英语、法语和德语的多语受控词汇同义词词典[5]。但是,有关跨语言信息检索效率的最早实验结果,是在1969年由Cornell大学的Salton所记录[6]。Salton通过翻译已有英语概念列表中的一

些单词为德语，来构建一个多语概念列表，而后利用该表扩充其SMART信息检索系统。针对图书馆科学摘要集的48个英语查询，通过人工方式翻译为德语，并且评价四种可能的语言对。在468个德语摘要中，使用英语查询而不是德语查询，使得平均查准率从0.35降至0.34(3%)；而在1095个英语摘要中，使用德语查询而不是英语查询，使得平均查准率从0.33降至0.31 (6%)。由此，Salton得出结论——尽管检索效率在文档集之间变化，跨语言检索几乎与单语处理同样有效。在更为细致地检查检索失败之后，Salton认为，在未来的实验环境中，应该使用一部更为完整的同义词词典。

在1973年的论文中，Salton实现了一个英-法多语概念列表，并通过在建立一个共同的概念集之后，单独开发针对每种语言的相应部分，来达到更为完整的覆盖范围[7]。同时，其中并未编码或者使用有关概念之间关系的任何信息。在此项研究中，Salton采取包含52篇摘要的法-英并行语料库，并使用包含16个已翻译查询的集合。Salton观察到，在法语摘要中，使用英语查询而不是法语查询，使得平均查准率从0.43升至0.45 (5%)；而在英语文档中，使用法语查询而不是英语查询，使得平均查准率从0.43降至0.38 (12%)。上述情形，也许可以通过在这类小集合中，对于分配给单独摘要的排序，平均查准率的度量标准存在敏感性来解释。

同时，Pevzner通过使用俄语PNP-2严格匹配受控信息检索系统，完成一种类似的实验[8]。PNP-2的复杂俄语词典包含数千个单词、数千个概念以及这些概念之间所存在的600种以上的关系，Pevzner将该词典扩展至英语[9]。然后，基于包含103个短俄语查询的相同集合，使用PNP-2分别检索俄语和英语文档。据Pevzner记录，对于有关电子工程的4000篇俄语和4400篇英语文档，测试结果在统计上并未显示出显著差别。

3.2.2 *受控词汇系统* 到1973年，利用多语同义词词典的受控词汇与概念检索系统都已成功建立，而其所达到的跨语言性能与利用相同技术的单语性能等同。由此，商业上的认可随之而来，到1977年，Iljon记录在欧洲已存在四种跨语言信息检索系统[10]。从上述早期工作开始，逐渐形成有关多语同义词词典的六条主要研究线路：

·*设计标准* 在1970年，各研究团体已达成共识——标准化同义词词典开发，以防止创建存在大量分歧与不一致的主题索引词汇表。在1971年，联合国教科文组织(UNESCO)提出有关多语同义词词典创建的标准[11]。在1973年，国际标准组织(ISO)开始从事该项任务，到1976年，已在很大程度上对规范草案加以扩充[12]。在1978年，通过该草案，而将其作为ISO 5964，并在1985年进行修正。在标准中，不仅描述了如何将领域知识合并于多语同义词词典中，而且确定了针对多语同义词词典开发可选择的技术。在1982年，前苏联采纳了一种类似的标准，即GOST 7.24-80[13]。

·*开发与维护工具* 随着大型多语同义词词典逐渐增多，其相应的设计与维护工具也变得越来越重要。在1970年，Neville描述了合并同义词词典的过程，该过程可用于合并单语同义词词典，以生成多语同义词词典。在1975年，Neville对上述方法与合并同义词词典的其它方法加以比较[14]。同年，Bollmann与Konard介绍了一种合并单语和双语同义词词典的技术。而在1977年，Iljon调查了可用的同义词词典设计和维护工具，并具体描述了一些系统[15]。

·*专用目的的硬件* 在1988年，来自NEC东京软件工程开发实验室的Kitano，描述了用于支持跨语言信息检索而设计的一种硬件工具的研制[16]。Kitano通过使用一种NEC集成电路——智能字符串搜索处理器，实现了一部日-英同义词词典。然而此时，ISSP同义词词典实现并未与信息检索系统集成在一起，因此未记录实验结果。

·*新的语言对和新领域* 有关跨语言信息检索的研究著作，已提供了实现新的语言对[17,18]以及新领域[19~21]的几个系统实例。由于这类报告描述了在同义词词典设计中所涉及到的以前未见过的语言现象，以及信息检索系统的其它方面(例如，抽取词干与复合词识别)，因此案例研究将会非常有益于实现ISO 5694以及类似的国家标准。

·*用户接口* 在1970年中期，IBM荷兰科学与交叉工业中心的Semturs，提供了从现代商业角度开发有关跨语言信息检索系统的用户接口的一些见解[22,23]。Semturs描述了一个商业产品的性能，即STAIRS-TLS严格匹配信息检索系统，该系统能够适用于德语、英语以及法语的查询和文档。STAIRS最初只是一个单语全文检索系统，而STAIRS-TLS加入了多语同义词词典。在STAIRS-TLS中，包含一个与基于同义词词典的工具的交互式接口，用于形成受控词汇查询的形式化描述。在Semturs的论文中，并未记录有关性能的具体说明，而是提供了有关跨语言信息检索的市场需求的一些观点。

·*用户需求评估* 受控词汇信息检索系统已广泛应用于图书馆，图书馆以及信息科学研究者非常关注于用户需求评估。在1974年，Rolling描述了为欧共体所进行的一次用户需求评估[24]。而TRANSLIB项目——针对图书馆程序的欧洲委员会I*欧洲远程信息处理的一部分，其中提供了最近有关用户需求评估的一些实例[25]。TRANSLIB的目标在于，为一个在线图书馆目录，开发三语种主题搜索(希腊语、西班牙语以及英语)。Chachra讨论了针对在线多语图书馆目录的用户需求评估，并提供了来自VLIS在线图书馆目录系统的一些实例[26]。除单语全文信息检索之外，VLIS通过使用一部多语同义词词典，来表示第二种语言的受控词汇语项。Rolland-Thomas描述了在加拿大DOBIS双语在线图书馆目录中所存在的类似特征，并从用户需求方面对自动技术的利用加以讨论[27]。

现今，跨语言信息检索系统已广泛应用，但几乎每一种商业系统均使用严格匹配方法。迄今为止，对于许多领域和语言，都已开发出复杂的多语同义词词典，而且也已了解加入新的领域和语言的过程。目前，对于受控词汇信息检索技术，主要在三方面存在局限性：

·*代价* 同义词词典构建是一项代价高昂的活动。但是，使用同义词词典的代价可能会更高，这是因为在受控词汇系统中，必须将反映每篇文档所包含概念的语项分配给文档。尽管通过使用自动的处理工具可提高人类生产率，只要需要，人类智能活动能识别并组织信息，但代价仍然是一个实质性问题。事实上，随着计算机硬件的价格不断下降，人类活动如同义词词典维护与受控词汇索引，已成为系统代价的主导因素。由于同义词词典构建与/或者使用在经济上不可行，这将限制已有的基于同义词词典的系统对于快速增长的电子可读文本的适用性，以及对于新领域的通用性。

·*未经训练用户的可用性* 这种限制也存在于全文严格匹配技术中，未经训练用户难以利用赋予他们的各项能力。对于选择语项，使用同义词词典中所编码的语项关系，以及为构建查询使用操作符如and,or与not，在熟练用户与未经训练

用户之间存在显著差别。在许多情形中已证明，提供已经训练的媒介比对用户进行充分训练更为经济。先进的用户接口如MenUSE为缓和该问题提供了一定的可能性，同时研究了在单语环境中从自然语言中构造布尔查询的专家系统[28]。

·效率　语言使用是一种具有创造性的活动，对于人类语言，每年都需要加入新词。因为同义词词典构建非常耗时，具体应用中的同义词词典在某种程度上落后于供一般使用的语项。而且，同义词词典的设计者难以预见哪些概念与关系将对系统用户有用。因为基于语料库的技术是基于所观察到的双语使用的统计实现，则通过该技术可认识到语项使用的重要性并对其加以利用，由此为提高效率而对基于语料库的跨语言信息检索技术进行深入研究。

3.2.3　概念检索　Salton的早期实验提供了有关概念检索的一个实例[29]。Slaton表示概念的一种方法，即为表示语项，并通过使用多语同义词词典来指导语项的选择过程。实质上，这是针对单语信息检索所研究的查询扩展技术的一种变体。查询扩展的基本思想是，通过利用相关项扩展查询中的语项，从而适应于语项使用的变化。但是，查询扩展一般是以查准率为代价来提高查全率，选择不适合的语项将会降低整体性能指标，如平均查准率。因此，在跨语言信息检索的环境中，查询扩展的目标在于，适应于跨语言语项使用的变化，而最小化对效率所造成的副作用。

最近，新墨西哥州立大学的Davis与Dunning已对几种跨语言信息检索技术进行评价，其中之一是基于查询扩展实现。对于第四届文本检索会议(TREC-4)有关西班牙语信息检索的评价，Davis与Dunning将25个西班牙语查询人工翻译为英语，然后基于这些查询，通过使用INQUERY信息检索系统，从包含来自墨西哥报纸"EI Norte"的58,000篇报道的文档集中选择文档。之后，通过从一个简单的双语语项列表中为查询中的每一个单词选择每种英语翻译，来形成西班牙语查询。该方法获得0.04的平均查准率，而参加TREC-4西班牙语信息检索评价的10个研究组中的5个，通过直接使用西班牙语查询，在相同集合中获得超过0.21的平均查准率。因此，Davis与Dunning的结果表明，对于跨语言信息检索，未受限的查询扩展技术，其性能在有限范围之内。

基于上述研究，法国Rank-Xerox的Hull与Grefenstette已对更为复杂的查询扩展方法进行评价[30]。他们将TREC的50个短查询人工翻译为法语，并创建了一个包含每一个法语单词的每一种可能翻译的双语语项列表。然后，使用未受限的跨语言查询扩展，通过SMART向量空间信息检索系统，从大约500,000篇新闻报道中进行选择，其中使用了相关性判断。Hull与Grefenstette发现，在双语语项列表中加入短语，使其效率指标从0.27升至0.36(33%)。而利用原始的英语查询，效率指标为0.39。由此，Hull与Grefenstette得出结论，在双语语项列表中包含短语，使得跨语言环境中利用查询扩展技术所达到的性能，与单语环境中利用传统统计技术的性能同样好。

3.3　基于语料库的CLIR方法

使用同义词词典的另一种方式是，直接利用从并行语料库中所获取的有关语项使用的统计信息。这种较为直接的方法非常适合于同基于语项使用统计的信息检索技术集成在一起。统计技术一般利用有关语项使用的两种重要观察。第一种是，用户判断为相似的文档一般使用相似的语项。另外一种观察是，对于文档区分来说，出现次数最少，即最罕见的语项，其可用性最大，而出现次数最多，即最常见的语项，其可用性最小。与相应内容几乎没有关系的常见语项，一般通过"禁用词表"将其删除。而对于剩余的语项，经常使用"倒排文档频率"进行加权，计算过程如下所示：

$$idf_i=\log_2\left(\frac{\text{文档数目}}{\text{具有术语 } i \text{ 文档数目}}\right)$$

组合"tfidf"(语项频率与倒排文档频率)中的两组结果

$$tfidf_{i,j}=tf_{i,j}*idf_i$$

其中，tf_{ij}为语项i出现于文档j中的次数。在信息检索系统中，经常使用有关语项和文档频率的更为复杂的tfidf函数。

3.3.1　自动构建同义词词典　在一定意义上，可以将基于语料库的技术看作为同义词词典自动构建技术中的一种，其中有关语项之间关系的信息，是从所观察到的语项使用的统计中获得。而该技术与其它技术之间的差别在于，不需要人工构建"同义词词典"。因为存在其它许多跨语言信息检索技术，所以在单语环境中，自动构建同义词词典具有非常重要的研究价值。目前，对于同义词词典的自动构建存在大量研究，从信息检索的角度出发，主要存在以下两种技术：

·基于名词短语抽取的方法　该方法是荷兰数字设备公司的van der Eijk开发的[31]。基于包含技术文档中大约1000个荷兰语和英语句子对并行语料库，Eijk从中抽取1,100个名词短语，并以此为基础进行测试。通过使用一种基于统计的词性标注器以及一个简单的句法分析器，标识每一个句子对中的名词短语。通过比较在包含名词短语的句子对的英语部分中所出现的每一个荷兰语名词短语的频率，以及出现于整个集合中的英语语项的频率，来为每一个荷兰语名词短语构建候选翻译。实验结果表明，唯一正确翻译的标识占45%，而产生包含唯一正确翻译的候选翻译列表的情况占66%。另外，句子对齐、词性标注以及句法分析错误占全部错误的85%。因此，Eijk推测到，利用该技术所能达到的性能上限应该是，唯一正确翻译的标识占60%左右，而包含正确翻译的列表占95%。因为所使用的并行语料库较小，当在语料库中引入相同语项的多种翻译时，不可能确定该技术的性能如何。最终所生成的双语词典并未用于信息检索中，因此不能确定是否造成翻译错误的因素也将影响到检索效率。

·基于机器学习的方法　最近，Arizona大学的Lin与Chen将机器学习应用于多语同义词词典构建之中[32]。基于有关语项聚类的早期工作，Lin与Chen通过使用一个包含1052个标题的集合，研制了一个汉-英概念表，而这些标题主要取自中国技术论文，多数标题中同时存在汉语和英语单词。通过使用基于相同标题中相邻语项的共现频率所获得的权重，Lin与Chen构造了一个Hopfield神经元网络，来生成语项的聚类。该系统通过语项聚类，将68%的文档转换为36个概念(不存在重复)。Lin与Chen说明，通过人工观测可以看到，语项与所有相关且准确的概念描述相关联，并且一些聚类同时包含汉语和英语语项。然而，他们并未对实验中的检索结果进行描述。

3.3.2　隐藏语义索引　应用于跨语言信息检索的一种统计技术，即为隐藏语义索引(Latent Semantic Indexing，LSI)[33]。其基本思想为，通过使用一种矩阵分解来标识由文档集所定义的向量空间的主要构成，然后将向量投射至这些主要构成所属的空间。在LSI中，主要构成用于表示重要的概念差别，而次要构成用于表示语项使用的变化。因此，LSI着重于tfidf的重要方面，而排除语项使用变化所造成的影响。

然后，使用余弦相似性测量比较文档，并按照一般方法进行排序。

对于在跨语言信息检索中所使用的LSI技术，其基本方法已由Landauer与Littman进行描述[34]。基于Handsards集合——有关加拿大议会会议的一个并行语料库，从中随机选择900个训练段以及1，582个测试段。首先，应用LSI来标识训练集合的主要构成。当将LSI应用于一个并行语料库时，通过矩阵分解标识与每一种语言相关联的向量空间中的主要构成，同时生成从每个主要构成至共同的表示空间的一种映射。然后，在共同的表示空间中，为测试集合中的每个段落选择tfidf向量的主要构成，而忽略了语言的不同。但是，由于缺乏具有相关性判断的双语语料库，不能进行传统的查全率-查准率评价。

Berry与Young利用英语和希腊语《圣经》中的段落章节，继续进行上述的研究工作。他们证明，利用较为细致的训练数据——仅使用每一节中的第一句，来标识主要构成，提高了检索效率，并超过Landauer与Littman所采用的粗略方法。使用16个查询，对于其中每个查询，在所构建的包含734小节的集合中，具有2-6个相关小节。Berry与Young观察到，当在集合的每一节中分布相同数目的句子，而不是聚集于少量章节中时，相关文档的平均排序从第六降至第四。

卡内基-梅隆大学的Evans与其他研究者，组合基于语料库与基于同义词词典两种技术。他们基于以西班牙语表达的自然语言查询，通过使用LSI，从包含125个英语医学语项的受控词汇表中选择语项。利用来自英语和西班牙语的相关词汇，对三部英语医学同义词词典中的定义进行扩充，从而获得一个包含3，084个单词的训练集合。

4 有关跨语言信息检索技术的一些讨论

对于基于机器翻译的CLIR方法，Fluhr与Radwan已证明，利用自动的机器翻译对查询进行预处理的方法效率，会对跨语言信息检索的整体效率造成一定影响。实现完全自动的机器翻译系统所需要的各类资源，可能会限制采取该方法进行小规模研究。另外，查询翻译存在一个缺陷，即为通过限制有关词义的上下文线索，而加重由歧义所带来的副作用。为处理这种影响，Hull与Grefenstette已提出，通过使用来自文档空间的结构化信息，以加强有关查询的具体解释。而且，Fluhr与Radwan已经实现了利用上述方法的一个简单版本。

基于机器可读词典的CLIR方法正逐步向前发展，并得到较好的应用。但是，有关同义词词典的完全自动的构建技术，仍然处于进展阶段。而对于多语概念检索技术如查询扩展，其中利用在同义词词典中所编码的信息，而在索引或者检索过程中没有人工干预，其性能仍然局限于与相同领域中利用相同技术的单语效率接近。如果不存在有效的同义词词典自动构建，则针对受限领域的概念检索技术将会存在很大的局限性。

目前，基于语料库的CLIR方法正逐渐趋向成熟，这就意味着在其成熟过程中，同义词词典仍然是CLIR系统的重要组成部分，而并未考虑所使用的检索模型。在计算语言学的研究领域中正在积极探索的一个方向，即为集成基于同义词词典的技术与基于语料库的统计技术。当将这两种技术组合在一起时，就可以利用每一种技术中所存在的优势。

在相同实验条件下，单语技术的性能成为针对检索效率上限的一个非常好的基准。一般来说，跨语言性能不可能超过单语技术的性能。实际上，单语检索与跨语言检索之间的一个重要差别，就是一个非常关键的约束因子——歧义。随着领域范围不断扩大，与单语检索相比，越来越显示出在跨语言检索中需要解决歧义。三个研究者，利用不同的实验设计，已发现可通过使用句法和语义信息减少歧义，其中最简单的一种方法就是短语识别。这表明，在单语环境中限制使用的词义消歧，将会成为未来研究的一个热点方向。

结束语 自从1965年开始，受控词汇的CLIR系统在一些领域中得到应用，并达到令人满意的效果。之后，研究者不断寻求适合于更为广泛领域而且有效的其它技术。在本文，我们对跨语言信息检索技术的发展及其应用进行一些介绍，并对其实现的具体方法作较细致的分析。CLIR方法主要划分为基于机器翻译、基于同义词词典以及基于语料库三类。利用机器翻译的两种主要方法为查询翻译与文档翻译，而利用同义词词典的两种主要方法包括受控词汇与概念检索。在少数一些情形中，已应用较为深层的语义处理。通过同义词词典的自动构建，将基于同义词词典与基于语料库两种方法联系起来，成为一种新型的组合方法。实验表明，该组合方法能够达到较好性能。目前，各研究团体较为关注的两个重要方面为：

- 缺乏大规模的多语训练语料库；
- 减轻歧义对跨语言信息检索效率所造成的副作用。

随着通讯技术的日益发展，国家之间的交往与联系逐渐增多，跨语言信息检索已成为一种非常重要的技术，而已应用于现有系统中的技术，无疑将会继续发挥其效用。但是，基于已有的CLIR研究工作，仍须努力开发出更为适合的新技术。

参考文献

1 Davis M W，Dunning T E. A TREC evaluation of query translation methods for multi-lingual text retrieval. In：D. K. Harman，eds. The Fourth Text Retrieval Conference (TREC-4). NIST，Nov. 1995

2 Fluhr C . Multilingual Information Retrieval. In：Ronald A Cole，Joseph Mariani，Hans Uszkoreit，Annie Zaenen，and Victor Joe Zue，eds. Survey of the State of the Art in Human Language Technology，Center for Spoken Language Understanding，Oregon Graduate Institute，1995. 291～305

3 Oard D W，et al. On automatic filtering of multilingual texts. In：Conf. Proc. 1994 IEEE Intl. Conf. on Systems，Man，and Cybernetics，volume 2，Oct. 1994. 1645～1650

4 Tallving M，Nelson P. Japanese databases and machine translation：A question of international accessibility to Japanese databases. In：Raitt D I，ed. 14th Intl. Online Information Meeting Proc. Oxford，Learned Information，Dec. 1990. 423～437

5 Pigur V A. Multilanguage information-retrieval systems：Integration levels and language support. Automatic Documentation and Mathematical Linguistics，1979，13(1)：36～46

6 Salton G. Automatic processing of foreign language documents. Journal of the American Society for Information Science，1970，21(3)：187～194

7 Salton G . Experiments in multi-lingual information retrieval. Information Processing Letters，1973，2(1)：6～11

8 Pevzner B R . Comparative evaluation of the operation of the Russian and English variants of the "Pusto-Nepusto-2" system. Automatic Documentation and Mathematical Linguistics，1972，6(2)：71～74

9 Pevzner B R . Automatic translation of English text to the language of the Pusto-Nepusto-2 system. Automatic Documentation and Mathematical Linguistics，1969，3(4)：40～48

10 Iljon A . Scientific and technical data bases in a multilingual society. On-Line Review，1977，1(2)：133～136

11 Educational U N. Scientific and Culture Organization

(UNESCO). Guidelines for establishment and development of multilingual scientific and technical thesauri for information retrieval. Palce de Fontenoy, Paris 7e, Dec. 1971
12 Ausin D. Progress towards standard guidelines for the construction of multilingual thesauri. In: Commission on the European Communities, ed. Third European Congress on Information Systems and Networks, volume 1, Verlag Dokumentation, May 1977. 341～402
13 Pashcenko N A, et al. Basic principles for creating multilanguage information retrieval thesauri. Automatic Documentation and Mathematical Linguistics, 1982, 16(3): 30～36
14 Neville H H. Alternatives to conventional multilingual thesauri. In: Verina Horsnell, ed. Report of a Workshop on Multilingual Systems, 1975. 10～12
15 Iljon A. Creation of thesauri for EURONET. In: Commission of the European Communities, ed. Third European Congress on Information Systems and Networks, volume 1, Verlag Dokumentation, May 1977. 417～437
16 Kitano H. Multilingual information retrieval mechanism using VLSI. In: A. Lichnerowicz, ed. RIAO 88 Program: User-Oriented Content-Based Text and Image Handling, volume 2, March 1988. 1044～1059
17 Ata B M A, et al. Sisdom: a multilingual document retrieval system. Asian Libraries, 1995, 493: 37～46
18 Cacares C. Russian-Spanish multisubject computer dictionary. Automatic Documentation and Mathematical Linguistics, 1986, 20 (2): 122～125
19 Benking H, Kampffmeyer U. Harmonization of environmental metainformation with a thesaurus-based multi-lingual and multimedial information system. In: Arthur Zygielbaum, ed. AIP Conf. Proc. 283, Earth and Space Science Information Systems, American Institute of Physics, 1992. 688～695
20 Lebowitz A I, et al. Multilingual indexing and retrieval in bibliographc systems: The AGRIS experience. Quarterly Bulletin of the International Association of Agricultural Libraries and Documentalists, 1991, 36(3): 187～192
21 Volodin K I, et al. Bilingual indexing of geological documents. Automatic Documentation and Mathematical Linguistics, 1991, 25 (6): 43～45
22 Semturs F. Information retrieval from documents in multilingual textual data banks. In: Third European Congress on Information Systems and Networks, Munich, May 1977. 463～467
23 Semturs F. STAIRS/TLS-a system for "free text" and "descriptor" searching. In: Everett H. Brenner, ed. Proc. of the ASIS Annual Meeting, volume 15, American Society for Information Science, Nov. 1978. 295～298
24 Rolling L. Multilingual systems: survey of the European scene. In: Verina Horsnell, ed. Report of a Workshop on Multilingual Systems, Oct. 1975. 4～5
25 Synellis C. TRANSLIN user survey report: [TRANSLIB technical report]. University of Patras Central Library, Rio 261 00 Patras, Greece, May 1995
26 Chachra V. Subject access in an automated multithesaurus and multilingual environment. In: Sally Mc Callum and Monica Ertel, eds. Automated Systems for Access to Multilingual and Multiscript Library Materials, International Federation of Library Associations and Institutions (IFLA), K. G. Saur, Aug. 1993. 63～76
27 Rolland-Thomas P, Mercure G. Subject access in a bilingual online catalog. Cataloging and Classification Quarterly, 1989, 10(1/2): 141～163
28 Marcus R S. Intelligent assistance for document retrieval based on contextual, structural, interactive Boolean models. In: RIAO 94 Conf. Proc. ings, Intelligent Multimedia Information Retrieval Systems and Management, Volume 2, Paris, Oct. 1994. 27～43
29 Buckley C, et al. Automatic query expansion using SMART: TREC 3. In: D. K. Harman, ed. Overview of the Third Text Retrieval Conference (TREC-3), NIST, Nov. 1994. 69～80
30 Hull D A, Grefenstette G. Experiments in multilingual information retrieval. In: Proc. of the 19th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, 1996
31 van der Eijk P. Automating the acquisition of bilingual terminology. In: Sixth Conf. of the European Chapter of the Association for Computational Linguistics, April 1993. 113～119
32 Lin Chunghsin, Che Hsinchun. An automatic indexing and neural network approach to concept retrieval and classification of multilingual (Chinese-English) documents. IEEE Transaction on Systems, Man and Cybernetics, 1996, 26(1): 75～88
33 Deerwester S, et al. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 1990, 41(6): 391～407
34 Landauer T K, Littman M L. A statistical method for language-independent representation of the topical content of text segments. In: Proc. of the Eleventh Intl. Conf. Expert Systems and Their Applications, volume 8, Avignon France, May 1991. 77～85

(上接第58页)

数据，使用(3)中的算法加密后传送给 Web 用户，后者解密后得到操作结果。

通过以上过程，在 Web 用户与后台数据库之间建立了一条从前端 Web 用户到后端数据库服务器的安全通道，远程用户可以使用该通道进行多次安全的数据交换，直到事务完成，然后数据库服务器询问是否拆除安全通道，经 Web 端用户同意后拆除该安全通道，废除 X.509证书，同时后台 DBMS 从 Role DB 的"用户授权表"删除用户角色授权信息，回收角色功能。

结束语 随着 Internet 技术的发展，基于网络的大规模应用系统面临着日益复杂的数据资源安全管理以及大量的访问权限管理，基于扩展 X.509认证的角色访问控制提供了一种灵活、有效、易于扩展的安全管理策略。

由于 X.509认证的安全性主要取决于公钥证书的生成算法，进一步的工作将是在此基础上研究更合适的加密算法，同时对 X.509证书进一步改进以减少证书管理的负担，增加用户对证书主体的信任程度。

参 考 文 献

1 Ahn G, Sandhu R. Role-Based Authorization Constraints Specification. ACM Trans. on Information and System Security, 2000, 3 (4)
2 Herzberg A, et al. Access Control Meets Public Key Infrastructure, Or: Assigning Roles to Strangers. IEEE Symposium on Security and Privacy, Oakland, May 2000
3 Housely R. Internet X.509 Public Key Infrastructure: Certificate and CRL Profile(RFC 2459), Jan. 1999
4 Housely R. Internet X.509 Public Key Infrastructure Operational Protocols: FTP and HTTP. RFC2585, May 1999
5 Kent S, Atkinson R. Security Architecture for the Internet Protocol Nov. RFC 2401 1998