●金 燕 张玉峰

知识检索中自然语言控制机制研究*

摘 要 情报检索过程中,对自然语言进行词汇控制是可行方法。借助各种技术和措施揭示词间的语义关系,词汇控制至少可以实现查询词的自动转换、一定程度的查询扩展、关联检索、排歧检索等,提高检索的语义性、知识性、智能性,提高查全率和查准率。参考文献 12。

关键词 知识检索 自然语言 词汇控制 语言学
分类号 G354

ABSTRACT In information retrieval, the vocabulary control over natural languages is feasible. By using various techniques and measures, we can reveal the semantic relationships of words, realize automatic search term conversion and search extension, and improve the coverage and accuracy of retrieval. 12 refs.

KEY WORDS Knowledge retrieval. Natural language. Vocabulary control. Linguistics.
CLASS NUMBER G354

1 知识检索与自然语言控制

知识检索是针对信息检索中存在的语义性较差、智能性低、知识性较弱等现状提出的一种基于语义和知识关联,运用知识处理技术和知识组织技术,实现信息查询语义化、智能化的一种高级信息检索方式。与传统信息检索基于字、词的简单机械匹配不同,知识检索具有较高的分析和理解自然语言的能力,能对自然语言提问和文档信息内容进行语义理解和分析,支持语词、语义的排歧和扩展,实现同义词扩展检索和语义蕴涵检索;依据概念间的语义关系,能够实现基于概念、基于语义的关联检索、推理检索,克服纯粹字词匹配的不足。

知识检索大多支持自然语言检索。在信息检索领域,自然语言指的是出现在文献(包括文献题名、摘要、正文或参考文献)中的具有一定实质意义的词。自然语言由于易用性等特点,在知识检索领域得到了广泛应用。这是因为,对一般用户而言,使用受控语言构造复杂的检索表达式太困难。而自然语言选词的自由性、直接性减轻了他们的认知负担,符合他们的使用习惯;同时自然语言词汇有足够的专指度,而且一旦文献中有新词出现,检索时马上可以使用。但另一方面,自然语言的自由度,以及自然语言词义模糊、词间关系不清、词的变体形式多等特点也给检索带来了问题——如果不加以控制,系统对自然语言提问及文档内容的语义理解存在不精确、有歧义的可能。如果不能排除歧义,不能明晰词间关系,就不能精确地理解语义,就谈不上基于语义、基于知识的检索,因为知识检索的核心就是查询请求和文档内容的语义理解及语义匹配问题。因此,自然语言词汇的控制问题,直接影响着词汇间语义关系的揭示程度,影响检索语言表达的准确性以及概念的灵活伸缩性,从而影响整个检索系统的效率和性能。

2 自然语言词汇控制的语言学层次

对自然语言进行词汇控制,要遵循两个原则:尽可能地保留和发挥自然语言的优势,如易用性好,选词灵活,易增加新概念等;尽可能地弥补和克服自然语言天然的缺陷,如词的自由度过大,存在词义模糊,一词多义、一义多词,词间关系不清等。知识检索的语义性、知识性,要求人们在遵循这两个原则的同时,还要遵从这样一个规则,即"对词汇实施控制,必须以概念为中心,通过识别概念间的关系,建立一个与概念体系相对应的具有层次结构的术语体系"[1]。而从语言学的角度来看,对自然语言的规范和控制,可以归结为如下几个方面。

(1)形态学层次。主要是从外在形态上识别和处理各种词汇。自然语言中有很多同根词,对西文而言,指的是有共同词干的词汇,如英文词的原形与其单复数形式、各种时态的变形,以及前后辍形式等;

* 本文为教育部人文社会科学研究重大资助项目"信息可视化与知识检索"(项目编号 02JAZJD870004)研究成果之一。

对汉语而言,指的是有共同词素的词汇。除同根词,自然语言中还存在着大量的"一词多形"的状况,如同一个概念的全称和缩略语,同一词在不同时代的不同拼写形式等。这些词汇现象的大量存在,迫使我们必须对自然语言语词的形态进行规范,包括对同根词作识别和处理;对一般词汇进行拼写识别,自动更正拼写错误;对同一词汇的缩写形式和不同书写形式进行识别和规范等。

(2)语汇学层次。这个层次,主要是对不同词性的词进行识别和控制。自然语言词类众多,每种词类对文献内容的揭示深度和针对性有所不同[2];而且,同一个词可能具有不同的词类,词类不同时,语义也可能有所不同。因此,对检索语言进行语汇学上的控制必不可少。控制方法一般是建立词典,从词典中得知每个词的详细特征,从而判断词性。其应用体现在对文档或查询请求进行语义理解时可以标注词性、识别特殊词汇如专有词;还可以据此编制停用词、禁用词词典,辅助自动分词。

(3)句法层次。句法控制指根据事先确定的句法关系,把若干词组组合起来表示特定的论题,以避免产生语义含混现象。控制方法包括规定引用次序、倒置、插入和轮排等[3]。在知识检索中的应用主要体现在识别短语和分句,支持语义理解和控制,自动摘要等。

(4)语义学层次。自然语言中存在大量的一义多词现象。如果不对此进行全面揭示,必然会影响检全率。自然语言的自由性也使得语词间的关系非常复杂,词间关系不清。自然语言的这种天然缺陷迫使知识检索系统必须采取一定的控制措施,才能有助于实现查询扩展、聚类检索、关联检索等。解决方法包括建立入口词表、后控词表、语义网、概念网络、本体论等,形成一个可靠的术语体系,用以揭示词间的等同关系以及各种关联关系,包括上下位类、相关关系等。

(5)语用学层次。要考虑特定语境对语义的影响。比如"黄鹤楼"一词,既可指三大名楼之一的黄鹤楼,也可指湖北产的一种名烟,具体取何意,则要视语境而定。采用加限定(如加含义限定、范围限定)方法建立排歧词典是常用的控制方法之一。

3 知识检索中自然语言控制的措施

3.1 传统词表辅助方法

(1)入口词表控制法。入口词表是一种先控方法。所谓先控,指在标引阶段进行控制。较常用的入口词表有两类:一类是在现有的受控词表的基础上增加自然语言入口词形成,一类是专用的入口词表。这种方法可以保证用户检索时仍使用自然语言,但在系统内部,自然语言可自动转换为相应的受控语言。

(2)后控词表控制法。与先控方法相反,后控制词表方法是采用一定的人工干预手段,对自然语言用词进行形态学、词汇学、语义学等方面的规范,根据所揭示的词间关系及其他关联,在检索阶段对自然语言进行控制。采用这种方法,标引时仍使用自然语言,在检索阶段才使用词表。后控制词表方法在实践中用的较多,一般而言,后控制词表的控制程度和效果取决于如下因素:词表中自然语言的收录率;对自然语言概念关联的揭示程度;能否持续增加新概念词。

3.2 词汇控制的新技术

采用入口词表、后控制词表对自然语言进行控制,可以向用户提供与检索主题相关的同义词、近义词和相关词,在一定程度上解决了"词汇差异"问题,弥补了自然语言的不足,提高了检索效率。但在揭示语义和词间关系上,仍停留在较低水平,并没有克服"知识获取瓶颈"[4],对标引者和词表编制者有较高的要求,难以完全适应高水平的知识检索的需要。因而,词表在知识检索中是一种辅助手段,单纯的词表控制方法难以完全实现"概念匹配、知识检索"。要实现高水平的知识检索,除了借助辅助词表,还必须采取适应知识检索推理机制的新型控制手段和方法,非线性地处理、揭示自然语言信息源和查询请求。概念空间、本体论、语料库等属于这类方法。

(1)概念空间方法。这是指利用计算机对目标文档进行分析,自动构造概念语义网络并以此为基础进行概念检索的一种方法。其实质是一种借助语词共现分析、自动标引等技术自动生成词表的一种方法。概念空间方法适应知识检索的需要,通过文本挖掘技术,建立一个知识丰富的、可理解的概念空间,揭示文档中的概念及概念间的关联,是词表控制方法在新环境的发展,不仅保留了传统词表对语词控制的优势,也减轻了人工编制词表的负担。一旦概念空间生成,用户既可以自主浏览词表,也可以使用系统提供的多词语多链接的交互式词语建议,实现联想检索。有文献介绍了一种概念空间方法,包括4个步骤:文档和对象列表收集(识别特定领

域的文档集合，作为文本处理中词汇、术语来源）；对象过滤和自动标引（抽词和标引）；共现分析（借助一定的算法，语词间的关联系数，确定关联）；联想检索[5]。

概念空间方法为解决词汇控制问题提供了一种新的理念。但也有人对这种方式提出质疑，认为概念空间方法应该与人工生成词表进行整合，才能提供高质量的控制和检索。

(2)本体论方法。本体论(ontology)原本是一个哲学概念，与认识论相对，研究事物客观存在的本质。后来被计算机科学借用，用于描述共享词汇。本体论是对概念化对象的明确表示和描述[6]。就其实质而言，本体论是特定领域内的概念集，描述了该领域内公认的对象以及对象间的关系。我们知道，一般的主题词表用BT，NT，RT，UF等揭示词间关系，使用相对简单，但词间关系相对也比较模糊，不足以覆盖所有的词间关联。本体论有类似于主题词表的方面，但与主题词表相比，不仅对术语（概念）做了精确定义，还囊括了更为丰富的词间关系如kind of，part of，instance of，attribute of等，以及诸如caused by，used by，written by等众多的语义体系[7,8]。与一般的主题词表相比，本体论能够更精确地表达概念，更全面地揭示概念间的关联，而且，本体论描述语言还具有一定的推理功能，更适用于智能化的知识检索。

实质上，本体论提供了一个特定领域的可控概念词典，每个概念都被清晰定义并拥有可机器处理的语义，从而对自然语言的控制更为详细和准确，支持自然语言与检索系统的语义交流[9]。本体论的术语库和关系库的建立非常严格，需要特定领域的专家和语言学家的共同参与。目前广泛使用的本体论有Wordnet（基于心理语言规则的英文词典），Frameriet（一个提供很强的语义分析能力的英文词典）等[10]。

(3)语料库方法。语料库是一个由大量真实文本经过词法、句法、语义等多层次加工形成的语言材料库。加工方法包括标注词的词性、语义项、短语结构、句型、句间关系等。语料库本身不能直接用于自然语言处理，但由于包含了文本的词汇、语法、语义和语用信息，可以辅助计算机对自然语言新文本进行标注和控制，对信息源进行语义分析[11]。

4 结束语

张琪玉先生曾经说过，情报检索的过程绝对不能没有控制，而如何克服自然语言由于不规范和缺乏语义关联性而对检索不利的问题，是自然语言在情报检索中的应用所面临的主要难题之一[12]。对自然语言进行词汇控制是目前可行的方法之一。借助各种技术和措施揭示词间的语义关系，词汇控制至少可以实现查询词（包括同义词、近义词）的自动转换、一定程度的查询扩展、关联检索、排歧检索等，增加检索的语义性、知识性、智能性，提高查全率和查准率。随着信息检索向知识检索过渡，如何自动化地构建准确的、专业性的、智能性的动态语义网络知识库，对信息源进行标引，在检索阶段自动识别用户请求，理解其查询的“概念”并将其与标引进行匹配，使基于自然语言的知识检索系统既能符合用户的使用习惯，又能深入到语义、知识层面，提高检全率和检准率，是将来需要研究的核心问题之一。

参考文献

1 吕娟，袁湘琴.论第四种情报检索语言系统.中国图书馆学报，2002(1)

2,3 柴省三.情报检索语言的词汇控制问题.情报科学，2000，18(7)

4,5 朱晓华.基于概念空间方法的信息检索技术研究.大学图书馆学报，2003(2)

6,8,10 Ontology理论研究和应用建模——《Ontology研究综述》. http://gis.pku.edu.cn/Resources/TR/Ontology_Study_application.doc.2003-07-19

7 张晓林.Semantic Web与基于语义的网络信息检索.情报学报，2002，21(4)

9 万捷，腾至阳.本体论在基于内容信息检索中的应用.计算机工程，2003，29(4)

11 何儒云，汤艳莉.智能化信息检索研究.图书馆，2003(3)

12 张琪玉.网络信息检索工具增强关键词检索功能的措施.图书馆杂志，2001(3)

金燕 武汉大学信息管理系2002级博士生。通讯地址：武汉大学信息管理学院2002博。邮编430072。

张玉峰 武汉大学信息管理学院教授、博士生导师。通讯地址：武汉大学信息管理学院。邮编430072。

（来稿时间：2004-04-05）