●龚芳　耿骞　王洋

自然语言检索中的概念控制

摘　要　自然语言检索的本质是概念检索,在自然语言检索系统中采用概念控制可以优化检索效果。概念控制的内容有:概念抽取、概念扩展、概念组配。概念控制离不开知识体系的支持。图 2。参考文献 8。

关键词　自然语言检索　概念控制　语义网络

分类号　G254.0

ABSTRACT　The nature of natural language retrieval is concept retrieval. We can optimize retrieval by using concept control in natural language retrieval system. In this paper, the authors summarize the contents of concept control and introduce the knowledge system to support it. 2 figs. 8 refs.

KEY WORDS　Natural language retrieval. Concept control. Semantic web.

CLASS NUMBER　G254.0

自然语言检索是信息检索中的一个类型。随着互联网的普及和发展,信息检索的最终用户日趋扩大,自然语言检索成为重要的发展趋势。

目前,自然语言检索系统采用的是模式匹配技术。所谓模式指的是关键词或索引词汇。模式匹配技术处理速度快,简单易行,但也有缺点。自然语言检索系统对同义词、近义词、多义词和其他一些与其相关的词语没有进行规范和统一,词间缺乏有机的联系。当用户提问的检索概念具有多种表达形式时,采用单一的关键词或自然语言索引词匹配方式势必会影响检全率。自然语言检索系统的选词没有严格限制,词量过多过杂,这样会影响检准率,并且会过多地占用磁盘的存储空间,影响查询匹配的速度。要想解决这些问题,必须对自然语言查询做进一步的处理,也就是进行概念控制。

1　概念控制及其实现基础

1.1　概念控制的目的

概念并不是孤立存在的,一个概念总是与其他概念之间存在着各种各样的关系,如上下位关系、同义关系、反义关系等。关键词也会出现一词多义、一义多词以及同一事物多种表述的情形。根据概念之间的相互联系,在词的概念含义层次上建立联系,为检索用户提供相关的结果分析是概念控制的一个应用前景。例如,"体育"这一概念根据上下位类的关系可以细分为足球、排球、奥运会、亚运会等,单纯的字面匹配会漏检甚至误检很多与之相关的信息。通过概念控制就可以将一个上位类的概念扩展为多个子概念。体育新闻的检索就可以扩展为:体育新闻、球类(足球、篮球、排球)、田径运动、体育赛事(奥运会、亚运会、世界杯)等概念,同时对那些具有下位概念的词汇可以再次扩展,这样就大大地提高了检索效果。"计算机"和"电脑"是同一事物的不同表述,机械匹配的话就只能检索到有关"计算机"或"电脑"的信息,采用概念控制的相关方法可以将这些相同概念的词汇统一到检索匹配中,这样就扩大了检索面,提高了检全率。在自然语言检索系统中进行概念控制,就是把信息检索从目前的基于关键词层面提高到基于知识(概念)层面,能够从概念意义层次上来认识和处理检索用户的请求,从而提高检全率和检准率。

1.2　概念控制的主要方法

目前虽然没有一个检索系统可以完全实现理想状态下的高层次的语义检索,但有些自然语言检索系统已经采用概念控制查询。主要的方法是利用知识体系建立概念间的关系进行查询扩展,深度匹配,优化检索效果。

概念控制的内容包括:提问句概念语义块的抽取,从提问句中切分出概念词或词组等语义单位;基于知识体系对抽取出的语义单元进行概念扩展;概念的组配,将选择出的各检索单位基于知识体系的组织信息转换成体现概念关系的逻辑表达式。

概念抽取不等同于分词处理,其中包括普通概念的识别和人名、地名、事件名等专有名词的识别,并进行概念提取。对于普通的概念字串采用逆向最长匹配算法(或正向最长匹配算法),并综合切割标记等分词手段切分就可以进行概念抽取。对于词典中未

收录的概念词,可以采取基于句模、句子结构分析、词和词组构成规则、句内结构性标志字、标点符号等来进行切分。除切割标志外,已知的词也可作切割标记使用。

自然语言检索系统的本质是查询满足特定主题概念的文本,因此被检索的内容不是和提问句的字面匹配。对提问进行分析后抽取出的主要是概念或概念的组合,需要进行概念匹配,这就需要对检索句中的词进行概念扩展,即考虑提问句中词的同近义词、上下位词和关联词。可以通过知识体系保存同近义关系、上下位关系和其他关联关系,当处理用户检索需求时,通过查询知识体系可对提问句中的词按概念进行扩展。如"我国今天的体育新闻",可以通过知识体系对"体育"进行扩展,查询包括"篮球"、"赛车"、"奥运会"、"世界杯"等方面的内容,"今天"一词指检索者进行检索时的日期,因此需通过规则将"今天"映射为检索时的时间,将"我国"扩展为"我国"V"中国"V"中华人民共和国"(知识体系中可能只存在"中国"和"中华人民共和国"之间的同义关系,没有"我国"这一词条)。

概念组配,按其内在逻辑关系,可分为限定组配和相交组配。限定组配将一个概念的内涵增加到另一个概念的内涵中,从而加深概念的内涵,缩小或限定了概念的外延。相交组配将具有共同的属概念、概念之间具有相交关系、外延部分重合的概念组合成一个新概念。在构成查询表达式时,基于知识体系的扩展词和原词间为"逻辑加"的关系。如"美丽"扩展为"美丽"V"漂亮"V……V"标志"。提问句中语义块间的关系通常为逻辑乘的关系。语义块间的概念组配通常存在如下逻辑关系:(1)主谓结构,描述的是一种陈述与被陈述的关系。(2)偏正结构,描述的是一种修饰与被修饰的关系。(3)动宾结构,描述的是一种作用和被作用的关系。(4)并列结构,描述的是一种成分间的并列关系[1]。主谓结构、偏正结构和动宾结构间为"逻辑乘"关系,并列结构为"逻辑加"关系。但通过对检索提问句进行分析后发现,部分并列结构在用户的检索概念中为"逻辑乘"关系,于是采用通过句模分析和指示标志来确定语义块间的"逻辑乘"或"逻辑加"关系。提问句语义块之间的"逻辑加"关系通常存在显式指示标志,如"或"等。分析传统的主、谓、宾、定、状、补六大成分与句型的关系,可以辅助获取语义块的逻辑关系。检索提问句的语义概念和提问句的结构紧密相关。需要分析谓语的性质、句子的结构,如"把"字、"被"字句等。

概念控制的3项关键技术中,概念扩展和概念组配都离不开知识体系的支持,知识体系的好坏直接关系到检索效果的优劣。

1.3 概念控制系统

国内外已有一些检索系统在不同程度上实现了概念控制,代表系统有首信搜索引擎、孙悟空、UMLS等。下面以UMLS为例进行介绍。

UMLS(Unified Medical Language System,美国统一医学语言系统),是美国国立医学图书馆(NLM)于1986年开始研制的一项长期开发研究计划,旨在克服计算机生物医学信息检索中的两个显著障碍(相同的概念具有不同的表达方式;有用的信息分散在不同的数据库系统中),使用户很容易地跨越了在病案系统、文献摘要数据库、全文数据库以及专家系统之间的屏障。UMILS包括4个部分:专家词典、超级叙词表、语义网络、情报源图谱。专家词典是一个包含众多生物医学词汇的英语词典,可以确定英语词汇的范围以及识别生物医学术语和文本词的词形变异,也为超级叙词表提供了确定范围的医学术语和词汇。超级叙词表是生物医学概念、术语、词汇及其涵义等级范畴的广泛集成。语义网络是为建立概念、术语间错综复杂的关系而设计的,它为超级叙词表中所有概念提供了语义类型、语义关系和语义结构。情报源图谱是一个关于生物医学机读情报资源的数据库,其目的是利用超级叙词表和语义网络实现以下功能:确定情报源与特定提问的相关性,以便选取最合适的情报源;为用户提供特定情报源的范围、功能和检索条件等人工可读的信息;自动链接相关情报源;在一个或多个情报源中自动检索并自动组织检索的结果。

图1矩形框中所示的部分即为概念控制部分。

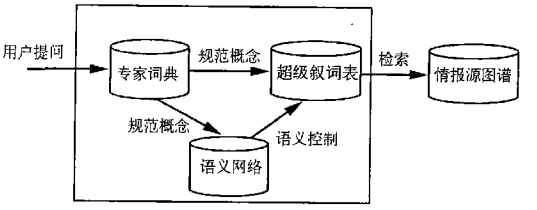

图1 UMLS的概念控制系统

2 支持概念控制的知识体系

实现概念控制离不开知识系统的支持,没有合理的知识系统,就无法实现概念扩展和概念组配,也就无法达到概念控制的目的。进行概念控制的核心是

构建合适的知识体系。

2.1 知识体系及其形式

目前，有关知识体系的研究主要集中在对Ontology的探讨中。国内关于Ontology的中文译法也不统一，有“知识体系”、“本体”、“概念集”、“概念体系”、“本体论”等多种提法。Ontology是一个源于哲学的概念[2]，原意指关于存在及其本质和规律的学说，后来被人工智能研究领域引入，特指对共享概念模型的明确的形式化规范说明。Ontology能够将词汇有关概念关系显式地表示出来，从而将术语的语义和概念关系显式化表示出来，因而在语义查询、概念控制方面发挥着重要作用。

Ontology中的概念表示一般采用框架结构，使用槽来表示概念的属性以及概念之间的关系[3]。借助概念之间的关系，Ontology在整体上形成了一个语义网。概念之间有4种基本关系：part-of表达概念部分与整体的关系；kind-of表达概念间的继承关系，类似面向对象中的父类和子类之间的关系；instance-of表达概念间实例和概念之间的关系，类似面向对象中的对象和类之间的关系；attribute-of表达某个概念是另一个概念的属性，例如概念“价格”可以作为概念“桌子”的一个属性。在实际应用中，概念之间的关系将不会局限于上述4种关系，可以根据特定领域的具体情况定义相应的关系。

2.2 语义网络

概念与概念之间有着横向或纵向的联系，形成了语义网。语义网络(Semantic Network)是美国语言学家奎廉(R. Quilian)于1968年提出的。1972年美国人工智能专家西蒙斯(R. F. Simmons)和斯乐康(J. Slocum)将语义网络用于自然语言理解系统中。语义网络是当今网络的扩展，这可扩展的网络给出了信息明确的定义，同时优化了人与计算机的合作[4]。

如何构建语义网络更好地将信息定义明确化目前仍没有统一的准则，这涉及到语言学、认知心理学等学科方面的知识。但是构建语义网络的统一原则都是将概念之间的横向或纵向联系显式化，组织成一个有机的结构形式。

Wordnet是由普林斯顿大学认知科学实验室开发的在线词汇参照系统。它将所有英语词汇分成5类：名词、动词、形容词、副词和功能词。名词按照3种关系被加以组织：部分关系、上下位关系以及物质与材料。同时有反义的名词被标注了反义关系，这样形成了一个互相高度连通的名词网络。动词的多义性比名词更高，在Wordnet中动词被组织成各种推演(蕴涵)关系，而组织动词的不同关系可以被总结成一个覆盖它们的基础词汇的推演，包括四种(见图2)[5]。

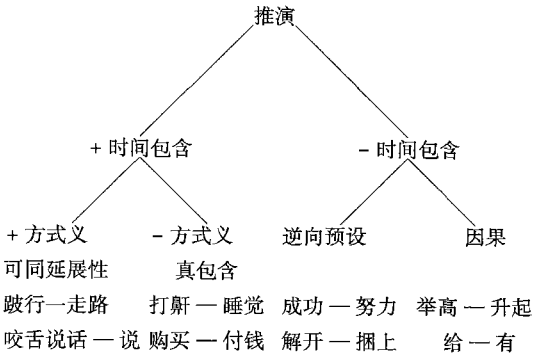

图2　动词的4种推演关系

2.3 概念词表

知识体系中除了概念间相互关系形成语义网络，对于每个单独的概念还必须有概念词表。没有概念词表的语义网络只是一个单纯的概念关系网而不是与词汇相结合的知识体系。Hownet中的词表就是一个概念词表。Hownet是一个以汉语和英语的词语所代表的概念为描述对象，知网是一个利用一种知识词典描述语言来描述概念与概念之间的关系以及概念的属性与属性之间的关系的知识系统[6]。Hownet包括词表和义原体系表。词表中记录了每一个词语的概念及其描述，每一个概念用一个记录来表示，如下所示：

NO. =017144

W_C = 打

G_C = V

E_C = ~网球, ~牌, ~秋千, ~太极, 球~得很棒

W_E = play

G_E = V

E_E =

DEF = exercise|锻练, sport|体育

其中NO.为概念编号，W_C, G_C, E_C分别是汉语的词语、词性和例子，W_E, G_E, E_E分别是英语的词语、词性和例子，DEF是知网对于该概念的定义，是知网的核心。每个DEF被称为一个语义表达式，采用知识描述语言，将与词汇有关概念关系采用显示的表示出来。如从概念编号为017144的DEF可知

"打"的概念之一是"锻炼",属于体育范畴。

2.4 构建知识体系的问题

构建一个适用合理的知识体系对于改善检索效果至关重要。但目前已成形的知识体系都存在一些不足。

构建一个知识体系最重要的是词的构造性信息,不完善的知识体系中被遗漏的信息大部分是关于词的构造性信息而非事实性信息。传统词典的定义尽力涵盖了有关每个词义的所有事实性信息,但却忽略了词汇概念之间内在的各种关系[7]。

如何选择知识体系的用词目前都在探讨和实验中,世界知识体系(Ontology)给出了一些选词的规定[8]:语义网络的非叶子节点不可以是多个类的词,如"蔬菜和水果";也不可以是没有下位类的选词,如"职业者"。要避免使用混合类的词,不要将叙述属性和抽象概念的词进行搭配作为语义网络的非叶子节点,如"空杯子"、"破车",要避免选词时加入个人的判断因素,将一些主观的属性值与类名组合成非叶子节点,如"热咖啡"、"明亮的车"。但是要使知识体系能理想地添加新的类、新的属性和关系,仅有这些规定是远远不够的。

是将概念领域化还是通用化各有优劣,目前都没有定论。通用的知识体系有 Wordnet、Hownet 等,专门领域的知识体系有 UMLS 和首信等。

构建语义网络结构本身存在一定的缺陷。用有限的结点和弧不可能代表万事万物及其相互之间的所有联系,语义网络对知识的表达有一定的局限性。单纯增加概念和联系会大大增加网络的复杂度。语义网络结构本身没有语义上的约定,不具备逻辑系统那样的有效性。单层的语义网络结构容易产生语义解释循环或语义悖论。

国外语义研究的理论与方法,并不完全通用。汉语是语义型语言,具有语义先决性、句法强制性和语用选定性等特点。汉语语义结构上的复杂性与多变性以及词与词之间无自然界限、无词尾形式标志、无形态变化的"三无"现象的存在,给语义分析带来了困难。

知识的获取与表示,其中较难解决的问题就是如何把复杂多样的专业知识系统化。如果把人工智能技术应用到一个多学科综合的检索系统中,如何辨别某个多义词当前的具体含义,如何辨析用户特定的需求,这些都有待于继续研究。

要想使计算机准确地分析、表达和传输知识,必须使它具备理解自然语言的能力。目前对自然语言的处理,虽然已从语法阶段上升到语义阶段,但对自然语言的理解能力还限制在一些规范的语句和语法范围内,这就决定了智能信息检索系统所能具有的智能化表达程度。

3 结束语

持续增长的海量网络信息让传统的检索方式面临着严重挑战,也加重了人们的检索负担。在自然语言检索系统中采用概念控制就是为了优化检索效果。而实现概念控制需要有合适的概念体系。目前没有一个标准的 Ontology 构造方法,对当前已有的 Ontology 的性能评估也没有一个统一的标准,这些都需要进一步研究。不过在构造特定领域 Ontology 的过程中,有一点是得到大家公认的,那就是需要该领域专家的参与。

参考文献

1 吴起立,李朝晖. 题名自动分类标引探讨. 情报学报,1999(1)

2 王洪伟,吴家春,蒋馥. 基于本体模型的信息检索机制研究. 情报学报,2004(1)

3 潘宇斌,陈跃新. 基于 Ontology 的自然语言理解. 计算技术与自动化,2003(4)

4 Tim Berners-Lee, James Hendler, Ora Lassila. *The Semantic Web*. Scientific American, May 2001

5,7 姚天顺等. 自然语言理解——一种让机器懂得人类与源的研究. 北京:清华大学出版社,2002

6 知网辟蹊径 共享新天地—董振东先生谈知网与知识共享. http://www. keenage. com/html/c _ index. html (Hownet 04. 08. 02).

8 Why we need guideline? http://www. cs. uoregon. edu/ ~ tdbreaux/poster/guidelines. html (04. 08. 02)

龚 芳 北京师范大学管理学院情报学2002级硕士。通信地址:北京。邮编100088。

耿 骞 北京师范大学管理学院副教授。通信地址同上。

王 洋 北京师范大学管理学院情报学2001级硕士。通信地址同上。

(来稿时间:2004-09-29)