基于领域本体科学效应知识语义检索的研究①

杨政国, 马建红

(河北工业大学 计算机科学与软件学院, 天津 300401)

摘 要: 当前传统的信息检索技术并不能准确的捕获用户的信息需求, 基于本体的方法虽然考虑到语义搜索的复杂性但是却迫使用户使用一个十分难以掌握的查询语法. 通过对用户查询习惯和查询短语的分析, 我们发现查询短语通常为简单的动宾结构短语. 针对化学领域科学效应知识和用户的查询习惯的特点, 给出了一种从自然语言查询到本体知识映射的语义检索的方法.

关键词: 本体; 语义检索; 语义理解

Reseach of Semantic Retrieval of Science Effect Based on Domain Ontology

YANG Zheng-Guo, MA Jian-Hong

(School of Computer Ecience and Engineering, Hebei University of Technology, Tianjin 300401, China)

Abstract: Current information retrieval (IR) approaches do not formally capture the explicit meaning of a query. Ontology-based approaches allow for sophisticated semantic search but impose a query syntax more difficult to handle. Through the analysis of user's query habits and query phrase, we found that a simple query phrase usually verb-object phrase. For the field of chemistry knowledge and scientific effects of the characteristics of the user's query habits, we present an approach for translating natural language queries to ontology knowledge.

Key words: ontology; semantic retrieval; semantic comprehension

1 概述

语义检索是当前研究的热点之一. 传统的单纯基于统计学的检索技术, 并不能理解查询的语义上的含义, 对于用户查询只能返回基于向量空间相似度计算的统计学最合适的结果. 基于统计学的本性和成熟, 相比语义技术, 他更加健壮并且适应网络规模[1]. 但是由于搜索引擎技术自身的问题, 返回的结果与用户的信息需求仍有一定的距离, 用户对搜索引擎的满意度不高. 主要表现为搜索结果数量大、结果不精确、有用的结果淹没在无用的结果之中等问题, 这种状况无法适应用户对高质量的专业化的信息服务的需求.

在可以使用本体来形式化描述问题的专业领域, 使用语义技术能够提供更好的搜索体验. 这个工作的一个决定性步骤是精确地捕获用户的信息需求. 然而, 目前大部分的基于本体的语义检索方法只是用本体来提供一些较浅的类似于关键词技术的检索体验, 或者仅仅是在本体知识中做是非判断, 存在着很大的局限性[2]. 在自然语言词汇到本体知识映射方面, 多是以字符串匹配方式简单而直接地完成映射, 使得映射成功率相对较低.

科学效应是 TRIZ(发明问题的解决理论)的理论核心之一, 是在特定条件下, 在技术系统中实施自然规律的技术结果, 是场(能量)与物质之间的互动结果. 效应也能看作是一种功能, 它使物质、场或两种的组合, 将输入作用转变为所需的输出作用. 通过选择不同的效应、物质参数, 可以控制效应的转换效果. 总之科学效应是科学原理、现象、定理和定律的集中表现形式和实施的必然结果.

本文提出一种通过使用依存语法分析对用户的查询语句进行语义理解, 将效应功能映射成流的参数变化, 从而准确检索科学效应知识的方法. 实验表明这种方法在计算机辅助创新的科学效应检索领域可以很

① 收稿时间:2013-07-10;收到修改稿时间:2013-08-16

大程度的提高检索的召回率和准确率.

2 科学效应知识本体库建立

科学效应是一种知识, 也是一种科学现象, 只不过是某种特定原因下产生的科学现象. 科学效应可以通过将输入流和输出流联系起来的定律来客观描述,并按照定律规定的原理将输入流转化为输出流以实现相应的功能. 依据国际上认可的科学效应的分类标准, 将科学效应本体分为: 化学效应、物理效应、几何效应和生物效应本体. 效应本体库包括功能、流和关系即:

$\Theta=(F,W,R)$

$F=[F_1,\ldots,F_n]$ F_x表示效应本体中的功能,

$W=[W_1,\ldots,W_n]$ W_x表示效应本体中的流,

$R=[R_{xy}]$ R_{xy}表示效应与效应、流与流、效应与流之间的关系.

在介绍功能定义之前, 先给出流的定义: 流是指功能中的输入和输出量, 流是一个具体的实体. 流又分为物料流、能量流和信号流三类. 流属性用来描述流的状态, 功能的输入流和输出流的转换有属性转换和关系转换两种, 其中的属性转换是指属性的类型或属性值发生变化, 关系的转换是指流之间的组合关系、包含关系等元组关系发生改变.

接下来给出功能的定义: 功能就是对一定设计环境下用来表述设计者意图的输入流、输出流之间的关系的抽象描述. 功能可用如下公式表示:

$$F=F(O,W_{in},W_{out})$$

其中, F 表示的是功能的名称, O 表示具体的操作, $W_{in}\in W$, $W_{out}\in W$, W_{in}表示输入流, W_{out}表示输出流, $F=F(O,W_{in},W_{out})$表示从输入流开始, 经过一定的操作处理, 到产生输出流的过程. 所以功能中至少要包含表示操作的动词和表示流的名词两个部分.

从功能的公式中总结出功能具有以下特征:

1) 功能是由行为发起的, 伴随着行为的发生而产生, 功能把行为提升到抽象层次上来表述, 产品设计者通过高层的功能抽象层次向下展开行为的确定, 继而得出问题的解决方案, 有效地扩展了设计思路.

2) 功能依赖于流而存在, 功能是流发生改变的一个黑盒, 如果没有输入流和输出流存在, 那么功能也就无法体现, 失去了实际的意义.

3) 功能是可变的. 由于系统的输入流的温度、大小、速度等流属性的不同, 系统的工作流程受控制信号影响也会有差异, 这些因素都直接导致功能发生变化.

本文以化学效应为例. 本体库中包括效应功能和物料流两大类, 效应功能的输入流输出流均可由物料流构成, 具体构建方法如下:

(1) 定义类以及类与类之间的层次关系, 从科学效应知识化学中抽象出基本的类及其类与类之间的层次关系. 这里的类对应本体体系中的 Class, 对象对应本体体系中的 Individual. OWL Thing 为基类, 在此之下建立效应功能和物料流两个大类. 在效应功能之下建立: 产生、保持、分离、变化、测量、消除、积聚、结合和运动九个子类; 物料流下根据化学物质分类建立子类. 接着每个子类下再建立子类, 从而形成类与类之间的层次关系.

(2) 定义概念与概念之间的关系, 这里的关系对应到本体库中的Object Property, 从而建立起概念与概念之间的映射关系, 以便推理出科学效应知识库存在但却隐含的知识.

(3) 定义概念(类和个体)的属性, 属性是两个个体之间的双重联系, 或者可以认为是两个 Individuals 之间的桥梁, 包括属性的名称、定义域、值域及其他约束, 对应到本体库中的数据属性(Datatype Property), 如物料有颜色、毒性、气味和状态等等属性, 效应有输入流、输出流和控制流等属性等等.

(4) 将前面几步定义的关系或者属性映射为本体体系.

(5) 对第 1 步中建立的类进行实例化, 即在类下添加具体的实例;

(6) 反复的采用 1~5 六个步骤对科学效应知识本体库中的本体进行修改和完善.

3 检索模型

基于领域本体和依存语法的语义检索大致流程是: 用户向系统提交问题, 使用句法分析解析用户查询语句, 从词语依存的层面上提取用户的信息需求, 语义分析模块对依存语法树进行解析产生本体查询语句, 分析理解用户的信息需求并检索组织呈现给用户.

科学效应描述的是输入和输出之间的关系, 用以

实现相应的功能. 对每个效应而言都有其输入和输出, 用一个黑盒来表示效应, 将效应的输入端和输出端称为两极, 那么该效应就成为两极效应. 但是对大多数效应来说, 除了输入端和输出端还会有一个控制端口来控制此效应的发生, 如果控制条件不同, 该效应产生的结果也就不同, 因此再给效应增加一个控制端口, 那么该效应便成为一个三极效应.

按照公认的标准对科学效应知识中对操作的划分, 可以分为九个基本类型: 产生、保持、分离、变化、测量、消除、积聚、结合和运动, 这九种操作基本上概括了工程求解过程中的所有情况, 针对这九个操作的基本类型的分析得出如表 1 的语义框架. 其中"-"表示属性在此流中不存在, "+" 表示属性在此流中存在, "↕"表示属性在此流中值发生了变化.

表 1 操作语义框架

动词类	动词子类	句势	含义
产生	合成 生产	1.[疑问词]+产生+属性限定词+名词 2.属性限定词+名词+是+[疑问词]+产生的 3.产生+属性限定词+名词的条件	输入流- 输出流+
变化	增加 控制	1.[疑问词]+变化+名词+属性 2.[疑问词]+使+名词+属性+变化	控制流↕
分离	分开 净化	1.[疑问词]+分离+属性限定词+名词 2.[疑问词]+从+属性限定词+名词+[中]+分离+[出]+属性限定词+名词 3.[疑问词]+把+属性限定词+名词+从+属性限定词+名词+中+分离+[出来]	输入流+ 输出流+
消除	破坏 去除	1.[疑问词]+消除+属性限定词+名词 2.[疑问词]+从+属性限定词+名词+[中]+消除+[出]+属性限定词+名词 3.[疑问词]+把+属性限定词+名词+从+属性限定词+名词+中+消除	输入流+ 输出流-
结合	混合 连接	1.[疑问词]+结合+属性限定词+名词+和属性限定词+名词 2.[疑问词]+把+属性限定词+名词+和+属性限定词+名词+结合 3.[疑问词]+把+属性限定词+名词+结合+到+属性限定词+名词+中	输入流+ 输出流+
积聚	吸收 存储 聚集	1.[疑问词]+积聚+属性限定词+名词 2.[疑问词]+把+属性限定词+名词+积聚	输入流+ 输出流+
测量	检测 度量 测定	1.[疑问词]+测量+属性限定词+名词	流参数
运动	移动 引导	1.[疑问词]+运动+属性限定词+名词 2.[疑问词]+使+属性限定词+名词+运动	控制流↕
保持	预防	1.[疑问词]+保持+属性限定词+名词 2.[疑问词]+使+属性限定词+名词+保持	输入流+ 输出流+

通过句势分类器[3]将自然语言查询归类到上述框架中, 从而确定问句的关注重点, 接下来进行语法分析得到依存语法树, 通过对语法树进行语义的分析得到问句所要表达的语义信息, 进而对科学效应本体进行查询推理. 检索模型如图 1 所示.

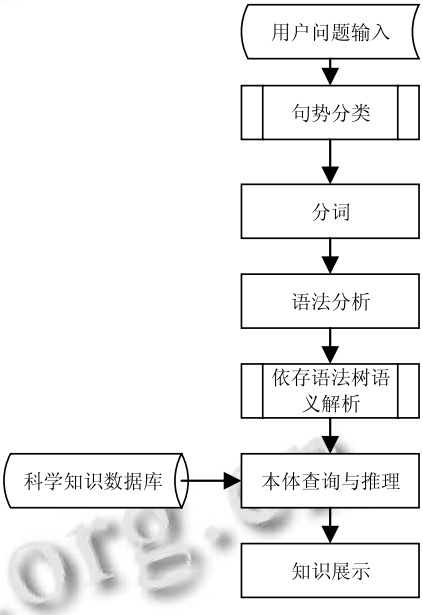

图 1 检索模型

4 基于依存语法的语义分析

依存语法是以语法成分之间支配与从属的依存关系为基础而建立起来的语法系统. 语法结构的组成成分之间存在着各种各样的支配与从属的关系, 这些成分依靠这类关系而彼此相联. 因此通过对句势分类器分类的句子进行语法分析, 根据支配词和从属词之间关系可以得到较为准确的语义信息. 分析流程如图 2 所示.

以产生的句势为例, 例如用户输入为"如何产生不可燃的黄绿色有毒气体", 通过句势分类器将其划分为产生类第一种句势, 使用语法分析器得到如下语法树:

[*advmod*(产生-2, 如何-1), *root*(*ROOT*-0, 产生-2), *neg*(可燃-4, 不-3), *rcmod*(气体-8, 可燃-4), *cpm*(可燃-4, 的-5), *nn*(气体-8, 黄绿色-6), *amod*(气体-8, 有毒-7), *dobj*(产生-2, 气体-8)]

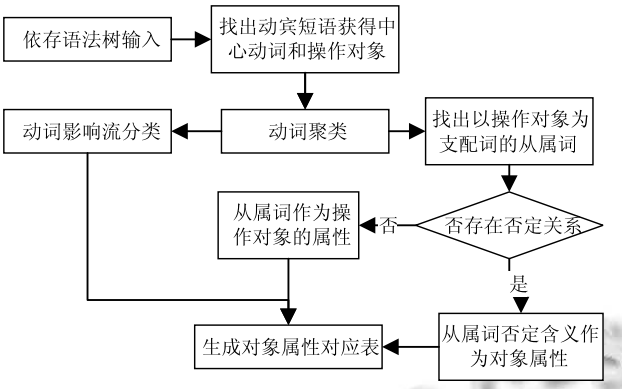

图 2　语义分析流程

根据句势模版得到此类句势核心为动宾短语结构, 而句中 dobj 为动宾修饰结构, 因此可以得出中心功能动词为产生, 名词为气体, 针对动词经过聚类后得出其蕴含的语义为产生新的物质即输出流里存在, 输入流里不存在的物质, 针对名词遍历语法树找到以名词为支配词的修饰关系并将从属词归类, 再将每个从属词作为支配词查找是否有否定关系, 如果有则把这个词的否定含义作为语义, 最终得到的语义分析结果为:

{中心功能动词=产生, 颜色=黄绿色, 毒性=有毒, 物质状态=气态, 可燃性=不可燃, 影响流=输出流}

最后通过本体推理查询模块生成并进行查询和展示.

5　仿真实验

根据上面理论实现了一个化学效应的小型语义检索系统, 使用 protégé进行本体构建, 由于在语义搜索领域还没有一个公认的测设数据集和评价方法, 而文中的工作也只是针对特定的知识创新的化学效应领域, 因而, 测试时使用的数据都是自己建立的化学相关领域本体知识及化学相关的资源, 而这些资源都经过手工方式标注过的.

召回率(Recall)和精度(Precise)是广泛用于信息检索和统计学分类领域的两个度量值, 用来评价结果的质量. 其中召回率是是检索出的相关文档数和文档库中所有的相关文档数的比率, 衡量的是检索系统的查全率. 精度是检索出的相关文档数与检索出的文档总数的比率, 衡量的是检索系统的查准率. 对于一个检索系统来讲, 召回率和精度往往不可能两全其美: 召回率高时, 精度低, 精度高时, 召回率低. 所以常常将这两个度量值融合成一个度量值 F 度量(F-measure). 其计算方法分别为:

召回率: $\mathrm{Re}call = \dfrac{A}{A+C}$

精度: $\mathrm{Pr}ecise = \dfrac{A}{A+B}$

F 度量: $F = 2 * \dfrac{\mathrm{Re}call * \mathrm{Pr}ecise}{\mathrm{Re}call + \mathrm{Pr}ecise}$

其中 A 为系统检索到的相关文档, B 为系统检索到的不相关文档, C 为相关但是系统没有检索到的文档.

图 3 为检索界面示意图, 从图中可以看出本文所述方法可以有效滤除无用信息, 提高准确率, 而基于关键词的传统方法虽然召回率和本文方法相同, 但是准确率较低.

表 2 为查询实验的数据, 四条语句分别为: 1.如何产生不可燃的黄绿色有毒气体; 2.如何产生有毒气体; 3.如何消除有毒气体; 4.如何降低温度.

表 2　实验数据表

语句编号	本文方法			传统方法		
	召回率	精度	F 度量	召回率	精度	F 度量
1	100%	100%	100%	100%	3.7%	7.1%
2	90%	100%	90%	100%	50%	66%
3	90%	100%	90%	80%	45%	62%
4	100%	100%	100%	0	0	0
均值	95%	100%	95%	70%	24%	33%

由表 2 的结果可以看出当检索语句简单且和句势框架接近时, 查询结果非常理想.

6　结语

以上所述的语义理解是对中文自然语言查询在科学效应知识检索领域的一种尝试和研究, 提出了一种利用依存语法和句势框架相结合的语义理解方法, 实验表明确实可以提高检索的准确率和召回率, 但是作为计算机专业人员, 本方法存在内存占用大的问题, 语义理解也只限于科学效应领域, 处理方法也不是很完善. 总之, 仍有进一步改善的空间.

图 3 检索界面

参考文献

1 Tran T, Cimiano P, Rudolph S, Studer R. Ontology-based interpretation of keywords for semantic search. The Semantic Web, Lecture Notes in Computer Science, 2007, 4825: 523–536.

2 陈叶旺,李海波,余金山.一种基于农业领域本体的语义检索模型.华侨大学学报(自然科学版),2012,(1):33–38.

3 刘宏哲.一种基于本体的句子相似度计算方法.计算机科学,2013,(1):257–2.

(上接第 204 页)

2 Paul D. Logic modeling as a tool for testability. IEEE Autotestcon'85. Long Island, New York. 1985. 1–12.

3 Gould E. Modeling it both ways: hybird diagnostic modeling and its application to hierarchical system designs. IEEE International Automatic Testing Conference. Orange, CA, USA. 2004. 576–582.

4 张烈刚.军用飞机通用 ATS 体系结构研究.计算机测量与控制,2005,13(4):346–347.

5 连光耀,黄考利,吕晓明,等.基于混合诊断的测试性建模与分析.计算机测量与控制,2008,5(1):601–603.

6 李行善,左毅,孙杰.自动测试系统集成技术.北京:电子工业出版社,2004.

7 高运征,万晓冬,杨春英.机载 ATE 总体技术指标确定方法的研究.计算机测量与控制,2008,16 (3):304–305.

8 Testability Timeline. http://www.testability.com. 2007.