基于规则推理的语义检索若干关键技术研究

马 森[1,2],赵 文[2,3],袁崇义[1,3],张世琨[2,3],王立福[2,3]
(1.北京大学信息科学技术学院,北京 100871;2.北京大学软件工程国家工程研究中心,北京 100871;
3.北京大学信息科学技术学院软件研究所高可信软件技术教育部重点实验室,北京 100871)

摘 要: 针对专业领域复杂的检索需求,目前相关研究采用基于语义的方法来扩展检索范围并提高准确度.在语义推理方面,目前搜索引擎通常直接采用语义网中的推理算法,推理效率不高.在排序方面,基于关键字的搜索引擎的排序算法也不适合对语义检索结果进行排序.针对上述问题,本文给出了基于语义网的语义规则建立方法,并提出了一种基于闭合世界假设的反向链接推理算法,提高推理效率,同时给出了一种基于特征相似性排序算法,使检索结果排序方式更加符合语义检索的特点.基于本文提出的方法,构造了语义搜索引擎 MaterialHub,实验表明该搜索引擎提高了检索的准确率和查全率,有较好的查询响应时间,并已经得到实际应用.

关键词: 语义检索;语义规则;语义规则推理;语义相似性排序
中图分类号: **文献标识码:** A **文章编号:** 0372-2112 (2013)05-0977-05
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2013.05.022

Research on Critical Technologies of Semantic Retrieval Based on Rule Reasoning

MA Sen[1,2], ZHAO Wen[2,3], YUAN Chong-yi[1,3], ZHANG Shi-kun[2,3], WANG Li-fu[2,3]
(1. *School of Electronics Engineering and Computer Science, Peking University, Beijing* 100871, *China*;
2. *National Engineering Research Center for Software Engineering, Peking University, Beijing* 100871, *China*;
3. *Key Laboratory of High Confidence Software Technologies (Ministry of Education), School of Electronics Engineering and Computer Science, Peking University, Beijing* 100871, *China*)

Abstract: For some complex specified domain retrieval requirements, many researches use semantic related technologies to resolve such problems. In terms of rule inference, it always leverages the inference algorithm of Semantic Web directly. However, the efficiency is not good. On the respects of searching results ordering, the ordering algorithm for search engine, which is based on keywords, is not suitable for ranking on search results generated by semantic retrieving. Focusing on the above issues, this paper proposes a semantic rule modeling method, and gives a new rule reasoning algorithm based on closed world assumption backwards reasoning chain to get higher inference efficiency compared to most semantic inference engines. Moreover, this paper proposes a new ordering algorithm based on feature similarity. Taking advantage of the above methods this paper describes, the search engine-Material Hub has been built up. Experiments show this semantic search engine improves the searching precision rate, recall rate, and rational responding time. So far, this system has been applied in industry.

Key words: semantic retrieval; semantic rule; semantic rule inference; semantic rank

1 引言

语义搜索引擎作为基于关键字搜索引擎的扩展,近年来已经成为研究的热点.尤其针对某些专业领域,方便、快捷的检索就显得尤为重要.例如在生产制造领域,主要体现在以下几个方面:(1)同类产品具有不同参数约束的个体非常多,那么仅依靠关键字,检索者需要逐条去筛选可选参数约束是否符合标准.(2)在查询中需要查询者掌握相关领域的专业术语才能检索到,只了解大概的所属类别,无法通过关键字匹配.(3)如何为检索建立且有效的利用领域规则来提高检索的准确度.(4)如何对检索结果进行基于语义方式的排序等.这些是基于关键字的搜索引擎很难实现的,需要通过基于语义方法和技术实现.

由此,本文针对于上述检索需求,提出了领域规则建立、规则推理、以及基于语义相似性排序等解决方法.结合上述内容,本文构造了一个基于语义搜索引擎 MaterialHub,目前已经在航天某院实际应用.

收稿日期:2011-11-10;修回日期:2013-02-19
基金项目:国家973重点基础研究发展计划(No.2009CB320706);解放军总装备部预研项目(No.513150202)

2 相关研究

语义检索的研究主要集中在语义规则的建立、关键字转换语义查询的封装算法和语义搜索引擎的构造等方面.

(1)在语义规则建立方面,最常用的是 SWRL 很多常用推理机都支持基于 SWRL 的推理.如果规则并不复杂只是基本的约束表达,也可采用 W3C 的本体语言 OWL(Web Ontology Language)[1].

对于把关键字转换为语义查询语言是语义引擎常用的处理方式,例如文献[2]采用了基于自然语言处理的方式将关键字转换为形式逻辑的方式.文献[3]将查询语句先通过自然语言的方式进行处理与知识库进行同义词替换与匹配,然后封装成 SPARQL[3]形式进行基于本体的检索.

(2)在推理方面,大多数搜索引擎用到的是基于描述逻辑的包含关系推理,其推理算法如文献[4]中提到的 tableau 算法.此外,基于规则的推理也是语义推理的一种方式,文献[5]利用基于规则的推理对军事作战计划中的异构资源系统的互操作性进行研究.

(3)语义搜索引擎:第一类是基于表单的搜索引擎,"SHOE"语义搜索引擎[6]提供了基于表单的语义搜索;第二类是基于 RDF(Resource Description Framework)[7]搜索引擎,典型的代表是 Corese 搜索引擎[7];第三类是基于关键字,利用本体信息的方式来加强查询的语义搜索引擎,SemSearch[8]就是典型的这种搜索引擎,第四类是基于问答的方式,通过自然语言处理的方式强化语义处理的搜索引擎,AquaLog[9]和 ORAKEL[10]是这类搜索引擎的代表.

3 语义检索关键技术

本节所提及的关键技术是在本体已建立完成的前提下,领域本体的建立在其他文章中给出.图 1 给出了检索执行流程对应的关键技术.

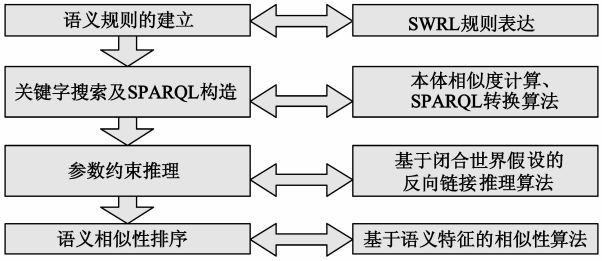


图1 语义检索执行流程

进行检索时,(1)系统预先设定语义规则,将领域规则通过 SWRL 的形式建立,将其作为知识库的一部分;(2)关键字用于对其所属领域本体的选取,选取过程采用关键字和本体的相结合的相似度计算方法,以确定搜索的内容是关于哪个领域本体;(3)如果检索中含有对领域规则的要求,系统需要进行基于规则的推理,然后采用本文提出的基于闭合世界的反向链接的推理算法来进行规则推理;(4)最后,检索排序采用了基于语义特征的算法对结果进行相似度量化计算.上述内容将在以下各节中给出.

3.1 基于 SWRL 的语义规则建立

本体的构造采用 OWL 语言,它是语义检索的前提.参数约束及规则采用 SWRL,其表达要依靠 OWL 语言来建立基本的术语及关系,所以规则的建立采用的是 OWL + SWRL 的形式.

对本体中大部分的内部参数约束可以使用 OWL 中的值约束、基数约束等来表达.但是外部参数、隐含规则,或选择表达式,例如属性的组合、限制假设,仅用 OWL 语言难以表达,需要借助 SWRL.

3.2 基于关键字的语义检索及 SPARQL 转换

3.2.1 基于关键字的领域本体匹配

语义检索是对一个或多个本体进行检索,首先需要确定被检索的本体.为此,需要存储所有领域本体的索引,并根据检索关键字找出所有相关的本体,为下一步的语义检索做准备.例如:检索机床,它可能存在于不同的本体中,需要找到所有涉及机床的本体.如果仅存在于一个本体中,那么后面的基于语义的检索就是针对这个本体进行.如果找出若干本体,那么将以问答的方式,选择涉及领域.在匹配时,除了把查询类型作为主要匹配项,还将进行参数匹配,并进行本体相似度计算并列出相关本体.相似度公式为:

$$sim(i,o) = ui_k + r\frac{\sum_{j=0}^{n} i_j}{n} \tag{1}$$

公式等号左边为关键字输入 i 与本体 O 的近似度.i_k 表示能够在本体 O 中,检索出表达类型的关键字.i_j 表示在检索出的关键字语义节点下,是否能够找到参数.如果能找到 $i_j = 1$,否则为 0.n 为参数的个数.u 为关键字系数,表示关键字在近似度匹配时的权重.r 为参数与目标本体在 i_k 下的相关度系数,$u + r = 1$.

3.2.2 基于自然语言的 SPARQL 转换

在语义检索中,检索者必须熟悉本体的语法、形式化查询语言、目标本体的结构和词汇等,不易于检索.为了方便检索,输入采用自然语言.因此,在这个阶段需要将自然语言转换为语义查询语言 SPARQL.

在关键字转换成 SPARQL 之前,参数需要进行预处理.首先要与本体中的概念谓词进行匹配,如果本体概念中包含参数的名称,则认为能够匹配.出于效率考虑,并没有采用 WordNet[11]或者中文词库等方式来找出

谓词的同义词.然后,根据 SPARQL 查询模板进行查询语句封装,将类型及参数约束等关键字转换成 SPARQL,进行检索.

3.3 基于参数约束的规则推理

在参数匹配时,除了参数约束的范围匹配外,还要结合领域规则进行推理验证.本节描述了一种基于闭合世界反向链接的规则推理算法.

3.3.1 推理机执行 SWRL 规则的不确定性

语义推理机执行 SWRL 推理是基于开放世界假设的,其推理算法具有不确定性的特点,即无法开发出一个保证在有限时间推理的算法,关于 SWRL 不确定性的证明参见文献[4].为了提高推理效率并且在不降低检索准确率的情况下,能够保证推理在有限时间完成,系统推理基于闭合世界假设(Closed World Assumption, CWA).在推理中对于未知的个体,推理机假设其真值为假.对于开放世界假设,认为数据源是互联网,一个无限大数据源,对于未知个体不能认为其真值为假,所以采用 OWA 更合理.而本文提及的应用是针对某个领域,其数据源是有限大小的,所以采用 CWA 推理效率更高.此外,本文描述的推理算法是基于反向链接推理,即根据结论向已存在事实推导.相比于正向链接推理,推理的目标更加明确,正向链接推理可能会推导出若干与搜索目标无关的结论,虽然保证了完备性但降低了效率.针对搜索引擎具有确定目标的推理且效率的优先级大于完备性的情况下,选择反向链接推理更为合适.

3.3.2 基于 CWA 的 SWRL 推理过程

推理机首先转换 OWL-DL 中的 TBox[12] 子句,建立统一的规则格式.TBox 在描述逻辑中是概念术语集合,通常包含了概念术语间的层级关系.例如:系统中可信设备为:由系统指定的供应商来提供,并且这些设备已经在项目中使用过且使用超过 3 年,TBox 为:

原子类:

$DesignatedSupplier$, $Device$, $Project$

定义的类:

$$TrustedDevice \equiv Device \cap \exists\ isMadeBy\ DesignatedSupplier \cap \exists\ isUsedIn\ Project$$

对象属性:

$isMadeBy(Domain: Device, Range: DesignatedSupplier)$

$isUsedIn(Domain: Device, Range: Project)$

然后,将 OWL-DL 转换而来的 SWRL 规则与知识库中已经存在的 SWRL 合并.这时规则将在推理机中加载,此外还需要加载 ABox[12] 中的信息来完成推理,谓词 isMadeBy 表示设备由哪个供应商制造,谓词 isUsedIn 表示设备用在哪个项目中等.ABox 定义如下:

$Device(H30\ Filter; CEFU\ Duster; TD\ Grinder)$

$DesignatedSupplier(S1; S2; S3; S4)$

$Project(P1; P2; P3; P4)$

$isMadeBy(H30\ Filter, S2; TD\ Grinder; S4)$

$isUsedIn(H30\ Filter, P4)$

3.3.3 推理算法

在执行推理时,规则会被触发执行.考虑一种情况,假设规则的结论中有 n 个变量,而每个变量都有 m 个值,那么推理机最多要执行 m^n 个组合,这样推理的效率很低.

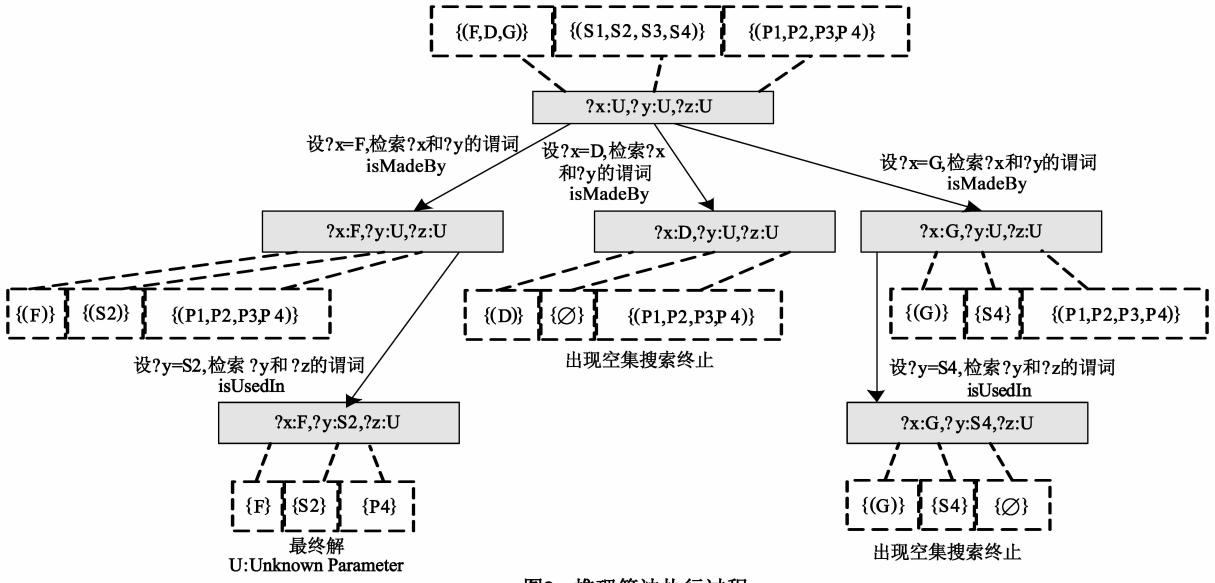


图2　推理算法执行过程

针对这个问题需要构造一棵树.树根列出推理结论中所有变量的值.选择一个变量,根据值的不同,进入不同的分支.在每个分支下,找到与这个变量关联的变量的取值范围,这样原有的范围就会被缩减.如果所有的变量都有值,那么这组值就可以作为一个解.如果一个变量取值为空,那么此分支代表一组变量和相关取值不能作为最终解.算法描述如下:

(1)预处理阶段:分析规则中的谓词,找到与知识库相关的规则,并给规则中的变量赋值.

(2)所有变量初始化后,按照算法进行递归查找.

(3)依据规则库中的规则,在给规则的前提中变量分配值时,规则结论不能找到相应的解与前提对应,那么就认为前面的取值不是解.直到找到规则中所有的变量都有取值,然后把取值集合作为解 s 加入容器 S 中.

(4)当搜索结束时,返回容器 S 为最终解的集合.

下面通过一个实例来说明推理的过程.推理的目标可信设备其定义为:由指定供应商提供并在项目中使用的设备.SWRL 规则形式如下:$Device(?\ x) \Lambda isMadeBy(?\ x,?\ y) \Lambda UsedsedIn(?\ x,?\ z) \Lambda DesignatedSupplier(?\ y) \Lambda y) \Lambda Project \rightarrow TrustedDevice(?\ x)$

图 2 描述了算法如何对规则进行推理,根据 3.3.2 节定义的 ABox 进行推理,找到规则的解.可以看到 H30Filter 为可信设备.

3.4 语义相似性排序

文献[13]提出了基于特征的语义匹配规则,本文采用的语义排序方式是对其进行的改进.即:

$$Sim(A,B,O) = u\frac{n(C_{super}(A,O)\cap C_{super}(B,O))}{n(C_{super}(A,O)\cap C_{super}(B,O))+\min\{P_{super}(A,B,O)\}} + v\frac{n(C_{sub}(A,O)\cap C_{sub}(B,O))}{n(C_{sub}(A,O)\cap C_{sub}(B,O))+\min\{P_{sub}(A,B,O)\}} + \varphi\frac{P_{sat}}{P_{tot}}$$

其中 $C_{super}(A,O)$和 $C_{super}(B,O)$表示在本体树 O 中,包含 A 和 B 的概念数量.其中,分子 $n(C_{super}(A,O)\cap C_{super}(B,O))$表示交集的个数,分母中 $\min\{P_{super}(A,B,O)\}$表示 A,B 的概念在本体图中的最短路径.公式中的 u、v、φ 为三个项的系数,其中 u 表示概念 A,B 相对于本体树 O 父概念的相似度的系数,v 是概念 A,B 相对于本体树 O 子概念的相似度的系数,φ 是参数约束满足的系数,如果参数很多,φ 的值就相对大一些,三个参数的和为 1,即 $u+v+\varphi=1$.

在上述公式中 $0<Sim(A,B,O)<1$.当 $A\equiv B$ 时,$Sim(A,B,O)=1$,当完全不相关时,或者没有共有的父概念和子概念以及共同的参数时,$Sim(A,B,O)=0$.

4 实例研究和实验分析

基于本文提出的闭合世界反向链接算法,本节给出了搜索引擎 MaterialHub 的基本框架,如图 3 所示.

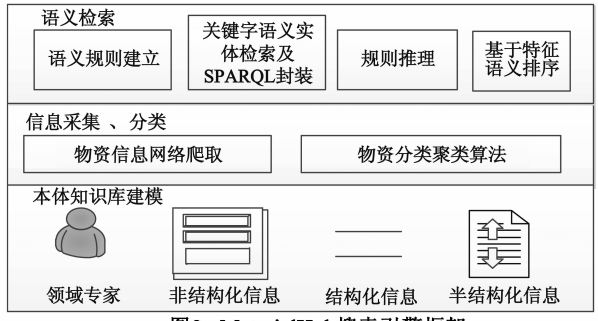


图3 MaterialHub搜索引擎框架

语义检索界面如图 4 所示,搜索的目标是过滤机、可信设备.过滤机的子类真空过滤机、板框压滤机都通过本体推理检索出来.然后进行规则匹配,根据可信设备的规则定义对检索结果进行规则推理.过滤可信设备.最后对检索结果根据第 3 节描述的基于特征相似性算法进行排序.

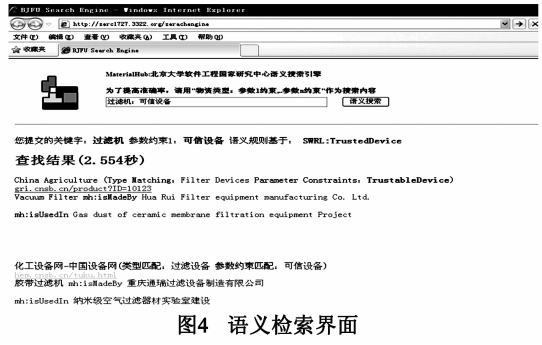


图4 语义检索界面

实验 Java 环境为 JVM1.6,Web 容器采用的是 Tomcat7.0,服务器配置了 Pentium(R) Dual-Core2.5GHz 处理器和 3GB RAM.

在实验中与其对比的是基于关键字的搜索引擎以及基于文献[13]Ameri&Dutta 的语义搜索引擎,实验对某制造集团相关厂商的产品信息进行了爬取,共采集到大约 2G 产品数据,建立索引大约 20M,本体库有 1875 个物资概念,16 个本体文件,133 个规则约束条件.针对于带有参数约束、隐含规则的、具有子类的概念的产品进行 1000 次检索并计算其准确率及查全率.结果如图 5,6 所示,MaterialHub 准确率、查全率更高.对查全率进行比对,如图 6 所示,得到同样的结论.

5 总结

本文描述了基于规则推理的语义检索若干关键技术.阐述了一种基于闭合世界反向链接推理算法,在保

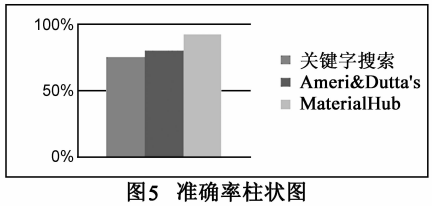


图5　准确率柱状图

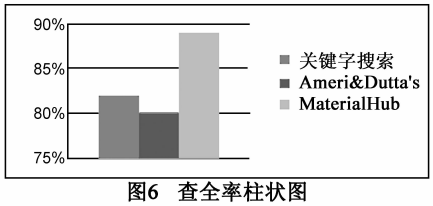


图6　查全率柱状图

证推理准确率的情况下使规则推理效率更高.提出了一种基于语义特点的排序算法,使得语义检索结果排序更加合理.基于本文开发的原型系 MaterialHub 已经得到实际应用,使用效果良好.本文提及的基于特征语义排序算法由于包含了最短路径算法,尽管使排序准确率有所提高,但效率相对不高,所以如何对语义检索的结果进行有效的排序也是项目将来研究的方向.

参考文献

[1] 高明霞,刘椿年.基于约束的自然语言问题到 OWL 的语义映射方法研究[J].电子学报,2007,35(8):1598 – 1602.
GAO Ming-xia, LIU Chun-nian. A constraints-based semantic mapping method from natural language questions to OWL[J]. Acta Electronica Sinica, 2007, 35(8): 1598 – 1602. (in Chinese)

[2] Qi Zhou, Chong Wang, Miao Xiong. SPARK: Adapting keyword query to semantic search[A]. Proceedings of the 6th International Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference[C]. Heidelberg: Springer, 2007. 4825:694 – 707.

[3] Angles R, Gutierrez C. "The expressive power of SPARQL"[A]. Proceedings of the 7th International Semantic Web Conference[C]. Heidelberg: Springer, 2008. 5318:114 – 129.

[4] I Horrocks, Patel-Schneider. Owl rules-a proposal and prototype implementation[J]. Journal of Web Semantics: Science, Service and Agents on the World Wide Web, 2005, 3(1):23 – 40.

[5] D Elenius, David Martin. Reasoning about resources and hierarchical tasks using OWL and SWRL[A]. Proceedings of the 8th International Semantic Web Conference[C]. Heidelberg: Springer, 2009. 5823:795 – 810.

[6] Heflin, Hendler J. Searching the Web with SHOE[A]. Proceedings of the AAAI Workshop on AI for Web Search[C]. California: AAAI Press, 2000. 35 – 40.

[7] Corby O, Dieng – Kuntz R, Faron-Zucker C. Querying the semantic web with Corese search engine[A]. Proceedings of 16th European Conference on Artificial Intelligence[C]. Valencia: IOS Press, 2006. 705 – 709.

[8] Yuangui Lei, Victoria Uren. "SemSearch: A search engine for the semantic web"[A]. Proceedings of the International Conference on Knowledge Engineering and Knowledge Management[C]. Heidelberg: Springer, 2006. 238 – 245.

[9] Lopez V, Pasin M, Motta E. AquaLog: An ontology-portable question answering system for the semantic web[A]. Proceedings of European Semantic Web Conference[C]. Heidelberg: Springer, 2005. 3532:546 – 562.

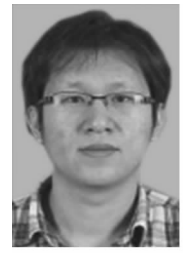
[10] Cimiano P. ORAKEL: A natural language interface to an F – Logic knowledge base[A]. Proceedings of the 9th International Conference on Applications of Natural Language to Information Systems[C]. Heidelberg: Springer, 2004. 401 – 406.

[11] Miller G. "WordNet: A lexical database for English"[J]. Communications of the ACM, 1995, 38(11):39 – 41.

[12] F Baader, Calvanese D. The Description Logic Handbook[M]. London: Cambridge University Press, 2002.

[13] Ameri. F, Dutta D. "A matchmaking methodology for supply chain deployment in distributed manufacturing environments"[J]. Journal of Computing and Information Science in Engineering, 2008, 8(1):1 – 9.

作者简介



马　森　男,1980 年 10 月出生于天津,北京大学信息与科学技术学院博士生.主要研究领域为语义网相关技术、应用集成.
E-mail: masen@pku.edu.cn

赵　文　男,1967 年 11 月出生于辽宁省大连市,博士,北京大学软件工程国家工程研究中心副研究员.主要研究领为软件工程、工作流技术.
E-mail: zhaowen@pku.edu.cn