基于航空领域本体知识库的语义检索研究*

李伟刚，张克亮，王慧兰
解放军外国语学院，洛阳 471003
E-mail: a611z@163.com; kliang99@sina.com

摘　要：基于本体的知识库对提高信息检索系统的查准率和查全率起到越来越重要的作用。本研究将本体的理论和方法应用于航空领域本体的构建，采用基于 Web 的本体共建模式建立航空领域本体知识库，并在此基础上研究基于本体知识库的语义检索。与传统基于关键词的信息检索相比，基于本体知识库的语义检索能够实现更高的查全率和查准率，在领域信息智能检索方面具有明显的优势和极大的潜力。

关键词：本体知识库；Web 共建；航空领域本体；语义检索模型

A Study of Semantic Retrieval Based on Ontology-based Aviation Knowledge Base

Li Weigang, Zhang Keliang, Wang Huilan
PLA University of Foreign Languages, Luoyang 471003
E-mail: a611z@163.com; kliang99@sina.com

Abstract: Ontology-based knowledge base plays an increasingly important role in improving the precision and recall rate of a retrieval system. By applying the theory and practice of Ontology to the construction of aviation ontology, we construct an ontology-based knowledge base of aviation field through the web. Then, we make a study of semantic retrieval based on the aviation knowledge base. Compared with the conventional keyword information retrieval approach, semantic retrieval can significantly improve the precision and recall rate, which has great advantage and potential in the field of intelligent information retrieval.

Keywords: ontology-based knowledge base; Web-based co-construction; aviation ontology; semantic retrieval model

1　前言

万维网的出现极大地方便了人们发布和获取信息，但是随着时间的推移，万维网上的信息呈爆炸式的增长，而且大多信息是异构和分布的，给用户快速、准确查找信息带来了极大的困难。传统的搜索引擎主要是基于关键词的搜索，它们越来越难以满足用户的需求。主要原因在于：第一，它们多是基于关键词的匹配而不是语义匹配，搜索时只判断关键词是否在网页中出现，而无法考虑这些词的语义，检索出大量无关信息；第二，不同领域的用户对同一概念可能使用不同的关键词表达，使得很多含有相同语义内容的信息无法获取；第三，多个关键词检索时，只能表达简单的“与、或、非”的关系，而无法揭示概念之间丰富、复杂的语义关系。

为了克服传统信息检索系统中存在的问题，本文尝试建立语义检索系统。通过构建航空领域本体知识库，进而实现基于领域本体知识库的语义检索系统，以提高领域信息检索的查准率和查全率。

2　本体和语义检索

传统的信息检索系统本质上采用的只是基于关键词的简单匹配，缺乏对知识的表示、处理和理解能力。解决问题的关键在于将信息检索从传统的基于语言表层的关键词匹配提升到基于语义层面的概念匹配，也就是实现信息在语义层面的表示，进而实现语义检索。

* 本研究获武器装备军内科研项目支持。

本体具有良好的概念层次结构和对逻辑推理的支持，能够通过概念之间的关系来表达语义。它提供了领域知识的共同理解，确定领域内共同认可的概念，从不同层次的形式化模式给出这些概念和概念间相互关系的明确定义，通过概念之间的关系来描述概念的语义（邓志鸿等，2002）。

而基于本体的语义检索实际上就是要将本体所反映的语义关系应用到对信息资源的标引和检索中，通过对相关文件的解析和推理在语义层面上实现信息检索（丁晟春、顾德访，2005）。构建基于本体的语义检索离不开领域本体的支持，近年来在信息检索领域，研究者越来越认识到本体的重要性，一些学者开始尝试构建直接服务于信息检索的本体。

武成岗等(2001）建立了基于本体论和多主体的信息检索服务器，它是一种利用多智能主体和本体理论设计的信息检索服务器，利用本体对文档进行领域分类，同时对用户查询信息进行规范。

Maki（2003）年提出了基于本体结构的方法，基本思想是利用本体中路径来进行用户查询的扩展，进而提高检索的查全率。

陈康和武港山（2005）将本体融合到信息检索技术中，利用本体中拥有的领域知识，提高了信息检索系统对自然语言文本的理解能力，同时也方便了用户以自然语言的方式提出检索请求。

许德山等（2008）利用建立的科技本体，采用 W3C 推荐的标准查询语言 SPARQL 实现了对中文信息的语义检索。

总的说来，基于本体的语义检索主要是基于概念匹配的检索方法，把传统方法中从用户查询和文档抽取出来的关键词替换为含有语义的概念，以此把关键词的检索提升到语义层面的检索，在一定程度上改善了信息检索的效果（王进，2006）。然而这些方法的侧重点大多停留在文档或用户查询中所涉及的本体概念，而没有充分利用到本体中的属性和其他语义关系，如果能够充分利用这些信息，就能更进一步把语义检索的作用发挥出来，这也是本研究的目的所在。

3 基于 Web 的航空领域本体知识库的共建

3.1 领域本体共建模式

本体构建的方法还没有成熟的理论作指导，目前的本体构建方法都是针对具体的项目提出的。比如，骨架法，TOVE 方法，METHONTOLOGY 方法，SENSUS 方法，IDEF5 方法以及七步法。在本项目的研究过程中，我们参照上述本体构建方法逐渐形成了一套比较成熟的领域本体构建模式。丁洁（2008）结合具体任务，对七步法进行了研究和改进，构建了军事情报领域本体知识库。费勤龙（2009）在分布式学习理论的指导下，对本体共建方法进行了初步探索，试验性构建了机载雷达领域本体知识库。Kcliang Zhang 和 Qinlong Fci（2010）以及费勤龙、张克亮、朱沛胜（2010）提出了基于 Web 构建领域本体知识库的模式和方法，以此为指导进行了领域本体知识库的工程实践，并对基于本体知识库的智能信息检索进行了研究。

3.2 航空领域本体知识库共建成果

在领域专家指导下，我们借助《中国航空百科辞典》、《世界导弹大全》、《航空发动机》和《世界空军武器装备》等工具书对领域知识的分类，建立了航空领域知识的概念层次结构。根据实际工作需要，我们选取了航空领域中几个重要的子领域作为知识库构建对象，包括：航空器、航空国家、航空机构和组织、航空发动机、机载武器。把它们作为一级概念，之后再对每一类概念进行细分，建立了层次结构最多为五级的航空领域概念分类体系。

定义好类和类的层级关系之后，就能得到领域知识的概念框架结构。为了更全面地描述领域知识，还需要挖掘类之间的语义关系，将隐性的语义知识显性地表达出来，建立起领域知识的网状联通，经过形式化之后被机器理解。

只有语义关系还无法完整描述类自身的特性，还需要为类添加大量的内在属性，这些属性是区别类的本质特征。比如对航空器子类“军用飞机”的属性定义就包括通用属性和特有属性，通用属性是所有军用飞机都具有的，可被子类的所有类型军用飞机继承，而特有属性则是军用飞机的某一子类所特有的，不能被其他子类继承的属性。

完成领域本体的框架设计之后，还需要添加大量领域实例，实例能够继承所属类所有的语义关系和内在属性。领域知识经过实例化之后，就初步实现领域知识语义层面的表示，经过语义表示的领域知识库是进行领域语义检索、文本分类、自动摘要等研究的基础。

基于 Web 的本体知识库共建模式，能够充分利用网络的优势，克服时间和空间上的限制，把领域专家和一线工作者丰富的知识添加到本体知识库中，最大程度地实现了知识的共享和重用。我们在本院多语种海量信息处理实验室中，搭建网络共建软硬件环境，并组织人员进行本体知识库的共建，在短时间内构建了相对完整的领域本体知识库，该知识库中包含 140 个类，21 条对象属性，81 条数据属性，1300 个实例，1268 条公理，6291 条对象属性的陈述和 4370 条数据属性的陈述。航空领域本体知识库达到了一定的规模，将领域知识由线性、无序的表示转变为网状的联通，为基于本体的语义检索研究奠定了良好的基础。

4 基于领域本体知识库的语义检索

本研究中我们暂时借助 Protégé 中的 Queries 检索工具进行实验，并与百度进行对比，检验基于领域本体知识库的语义检索能力。经过试验，基于本体的语义检索与传统的基于关键词的检索相比有如下方面的改进：

4.1 支持一定的语义扩展，提高检索查全率

当前基于关键词的信息检索方式存在很多问题。比如，自然语言中大量存在一义多词的现象，同一概念可以用不同词语表示，不同领域、不同背景的人对同一概念检索时会使用不同的关键词，因此，搜索的结果会完全不同。当用户想检索关于“战斗机的分类和性能信息”，我们分别用“战斗机”和“歼击机”这两个关键词在百度中进行检索，检索的结果却是完全不同。实际上这两个关键词代表同一个概念，指的就是进行空战、夺取制空权的军用飞机，无论使用哪个词进行检索，都应该返回大致相同的内容。造成这种的结果主要原因在于传统信息检索是基于关键词的匹配，而没有实现基于概念和语义的匹配。

借助本体强大的语义表达能力，我们可以定义“战斗机”和“歼击机”是 Owl:equivalentClass 关系，即它们是等同类，是表达同一概念的不同形式。所以基于航空领域本体支持的语义检索查询“战斗机”时也能得到大量“歼击机”的信息，同样检索“歼击机”时，也能检索到大量“战斗机”的信息，如图 1 所示。可见，基于领域本体的语义检索不再是简单的基于关键词的匹配，而是上升到概念和语义层面，采用该方法能够在一定程度上解决自然语言中“一义多词”的现象。借助大量预先定义的领域知识，有效提高领域知识查全率。

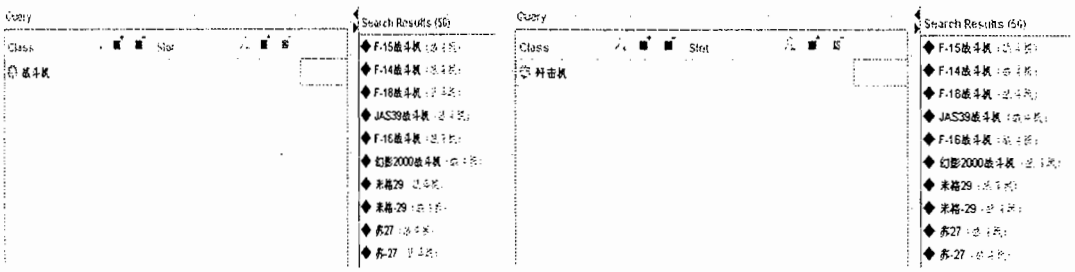

图 1 Queries 对战斗机和歼击机的检索结果

4.2 理解用户意图，揭示概念之间语义关系

为了准确表达意义，用户需要使用关键词组合进行查询，传统信息检索面对多个关键词时，不能准确理解它们之间隐含的语义关系，也就无法返回用户需要的信息。比如用户想查询“B52 轰炸机是由哪个公司生产的”，通常会使用关键词“B52 轰炸机 制造商”进行检索，图 2 显示的是百度的检索结果。

图2 百度对 B52 轰炸机制造商的检索结果

检索结果中大部分都是无关信息，用户真正需要检索的是有关生产商的信息，直接返回给用户生产商的信息才符合检索要求。而传统信息检索无法理解“B52 轰炸机”和“制造商”之间的语义关系，仅能返回给用户含有这两个关键词的网页。

基于本体的语义检索，首先要对这两个关键词之代表的概念间存在的语义关系进行理解。在航空领域本体知识库中，我们定义了“航空机构和组织”和“军用飞机”之间存在“研发”的语义关系，B52 是轰炸机的一个实例，而轰炸机是军用飞机的一个子类。经过推理，B52 轰炸机和航空机构和组织之间也存在类似的关系，因此，要在航空组织和机构的实例中查找才能符合用户的需求。图 3 显示的是查询结果，由此可以看出，基于本体的语义检索的确能够揭示关键词之间的关系，理解用户的意图，实现智能检索。

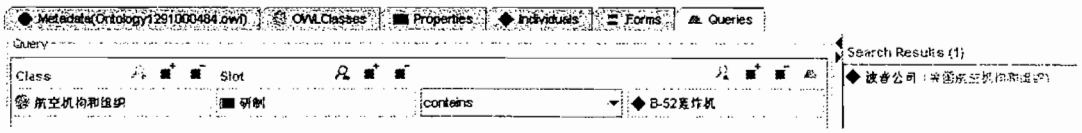

图3 Queries 对 B52 轰炸机制造商的检索结果

4.3 支持复杂语义查询，具有一定推理能力

传统的搜索引擎大多是基于关键词的匹配，无法支持关键词之间的语义推理，实现复杂语义的查询。比如要搜索“挂载有休斯公司研发的空对空导弹的战斗机”时，用户在搜索引擎中检索“战斗机 空对空导弹 休斯公司”，搜索结果都是关于空空导弹和休斯公司的网页，没有直接满足用户需求的信息，需要用户去总结。

回答这类问题，需要搜索引擎具有一定的推理能力，建立知识之间的关联，并将这些信息进行汇总展示。基于领域本体知识库的语义检索就能够完成这样的任务，我们在航空领域本体知识库中定义战斗机“挂载有”空空导弹，而航空机构“研发”空空导弹，建立起这样的语义关系后，可以从任何一个概念为入口找到用户需要的信息。语义检索能够将网页中的相关内容关联起来，提取出用户所需要的信息。比如某一战斗机挂载有空空导弹，而该导弹又是被休斯公司研发的，则该战斗机就是所要查询的信息，图4显示的是语义检索的结果。

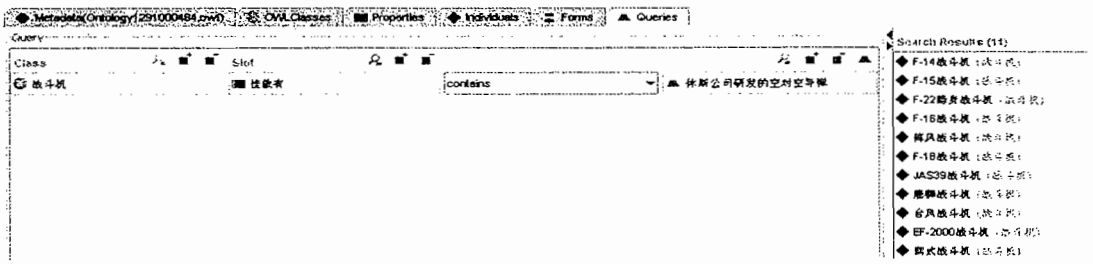

图4 Queries对挂载休斯公司空空导弹的战斗机的检索结果

从上述的实验可以看出，该检索程序已经具有了一定的语义检索能力，在一定程度上解决了传统信息检索系统中存在的问题，基于领域本体知识库的语义检索是今后信息检索的发展方向。

5 基于领域本体的Web语义检索模型

上述针对领域本体知识库的语义检索已经取得了比较理想的结果，在现实工作中发挥了重要作用。不过，真正能够检验语义检索威力、体现其全面优势的领域是面向互联网海量信息的智能检索。我们初步设计了基于领域本体的Web语义检索模型，其流程如下：

首先，我们在领域专家帮助下，建立相关的领域本体。领域知识是不断发展变化的，新的事物、概念不断的出现，因此在使用过程中要不断地更新和扩展领域本体知识库。其次，借助已有的本体在Web页面中插入语义元数据信息，从而使Web页面的内容机器可读，它是构建语义Web的基础性工作。再次，当用户输入检索请求时，查询转化器按照本体将查询请求转换成规定的格式，在本体的帮助下从元数据库中匹配出符合条件的数据集。在本模型中，我们采用SPARQL语言作为检索语言。最后，当系统检索时，需要生成符合语法要求的SPARQL检索问题表达式，并在本体——资源映射资源库中进行查找。如果资源库中有针对用户问题的直接答案，则直接呈现给用户；如果生成的结果有多个，还要按照与问题关键词或问句相关性进行排序，并将每一个答案，用超链接的形式与具体网页资源相连。

6 结语

借助基于Web的领域本体共建模式，我们建立了比较完善的航空领域本体知识库，在此基础上进行了基于本体知识库的语义检索实验，实验结果达到了预期的目标。为了充分发挥语义检索的优势，我们设计了基于领域本体的Web语义检索模型。在下一阶段的研究中，我们要依据该模型实现具体的Web语义检索系统，期待它能够提高互联网海量信息检索的智能水平，满足用户的现实需求。

参 考 文 献

[1] Keliang Zhang and Qinlong Fei. “Co-construction of Ontology-based Knowledge Base through the Web: Theory and Practice”, In Proceedings of the 6th IEEE International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China, 2010, pp. 345-350.

[2] R.Navigli, P.Velardi. An analysis of ontology-based query expansion strategies. In Workshop on Adaptive Text Extraction and Mining, in the 14th European Confernce on Machine Learning, 2003.
[3] 陈康, 武港山. 基于 Ontology 的信息检索技术研究[J]. 中文信息学报, 2005, 19(2): 51-57.
[4] 邓志鸿, 唐世渭等. Ontology 研究综述[J]. 北京大学学报 (自然科学版), 2002, 9: 730-738.
[5] 丁洁. 基于本体的英语军事术语知识库建设研究[D]. 解放军外国语学院硕士学位论文, 2008.
[6] 丁晟春, 顾德访. Jena 在实现基于 Ontology 的语义检索中的应用研究[J]. 现代图书情报技术, 2005, 10: 6.
[7] 费勤龙, 张克亮, 朱沛胜. 基于 Web 的机载雷达领域本体知识库的共建研究[J]. 微计算机应用, 2010(9): 21-28.
[8] 费勤龙. 基于 WEB 的机载雷达领域本体知识库的建构与查询研究[D]. 解放军外国语学院硕士学位论文, 2009.
[9] 世界导弹大全. 北京: 军事科学出版社, 1998.
[10] 王进. 基于本体的语义信息检索研究[D]. 中国科学技术大学博士论文, 2006.
[11] 武成岗, 焦文品, 田启家. 基于本体论和多主体的信息检索服务器[J]. 计算机研究与发展, 2001, 38(11): 641-647.
[12] 许德山, 乔晓东, 朱礼军, 宫丽环, 杨洁雄. 基于本体的中文语义检索系统[J]. 情报理论与实践, 2008, 31(3): 450.
[13] 中国航空百科词典[M]. 北京: 航空工业出版社, 2000.
[14] 钟华. 世界空军武器装备[M]. 长沙: 国防科技出版社, 2001.
[15] 张伟. 航空发动机[M]. 北京: 航空工业出版社, 2008.