知识网格中基于领域本体的智能检索*)

胡艳丽 白 亮 张维明 肖卫东 汤大权
(国防科学技术大学信息系统与管理学院 长沙 410073)

摘 要 本文提出知识网格环境下基于领域本体的智能检索模型,采用 OWL DL 语言对领域知识进行形式化描述,支持推理和深层语义检索。"标注"和"查询优化"是检索的两个关键技术。通过规范的概念和概念间语义关系对文档片段进行标注,并针对"一词多义"问题提出"主题-概念"两阶段消歧算法。"查询优化"过程中,基于 OWL DL 推理的优化算法实现查询概念的自动扩展,提高了查全率和查准率。基于以上方法,建立航天领域本体,利用网上数据库开放资源作为测试集进行评测。实验显示,与传统基于关键词的检索方法相比,基于领域本体的检索方法具有较高的查准率和查全率。

关键词 知识网格,智能检索,领域本体,标注,查询优化,消歧,OWL DL

Domain Ontology-based Intelligent Information Retrieval in Knowledge Grid

HU Yan-Li BAI Liang ZHANG Wei-Ming XIAO Wei-Dong TANG Da-Quan
(Sch. of Information System & Management, National Univ. of Defense Technology, Changsha 410073)

Abstract Intelligent information retrieval model based on domain ontology is proposed in this paper. Domain knowledge is described with OWL DL language, which supports reasoning and uses deep semantics to guide the process of retrieval. Annotation and query optimization are two critical technologies in information retrieval. The annotation of document segments is implemented with formal definition of concepts and inter-relationships of concepts in domain ontology. "Topic-Concept" Disambiguation algorithm is suggested to solve the problem of polysemy in natural language processing. Query optimization algorithm is put forward to perform automatic expansion for users' query with reasoning mechanism of OWLDL. Based on the above technologies, ontology-based intelligent retrieval is realized in aerospace domain. The evaluations show that the effectiveness of retrieval can be improved with the ontology-based retrieval method, compared with traditional retrieval method based on keyword matching technologies.

Keywords Knowledge grid, Intelligent retrieval, Domain ontology, Annotation, Query optimization, Disambiguation, OWL DL

1 引言

上世纪 90 年代以来,研究者们逐渐意识到在分布异构环境下构建大规模、健壮知识系统的关键是知识共享和重用,并进行了大量研究,探索支持共享和重用的知识基础设施[1]。

知识网格就是一个支持有效获取、发布、共享和管理知识资源的智能互联环境,并提供所需要的知识服务,辅助实现知识创新、协同工作、问题解决和决策支持[2]。

F. Berman 于 2001 年底在"Comm. ACM"撰文提出知识网格以来[3],知识网格及相关研究得到密切关注。GGF 专门成立了 Semantic Grid Research Group,英国 Southampton 大学在这方面进行了深入研究[4];C. Mario 和 D. Talia 认为语义与知识网格是网格未来的发展方向,是下一代 e-Science 的基础设施[5];R. W. Moore 于 2001 年提出 Knowledge-based Grids 并探讨了它的应用实例[6,7];我国中科院、浙江大学等单位的学者也纷纷提出 Knowledge Grid[2]、Knowledge base Grid[1],并于 2001 年 7 月成立了中国知识网格研究组,开发了中国 e-Science 知识网格环境 IMAGINE,并在科技部 973 计划的支持下开展中国语义网格计划的研究,建立了知识网格论坛。

知识网格的主要研究内容是:利用网格、数据挖掘、推理等技术从大量在线数据集中抽取合成知识,使搜索引擎能够智能地进行推理和回答问题,并从大量数据中得出结论[3]。作为其关键技术之一,智能搜索的研究日益得到重视。

有效利用语义信息和本体论替代传统的关键词匹配是检索智能化的重要途径。知识网格包含了反映人类认知特性的认识论和本体论。本体定义了组成主题领域的词汇的基本属性和关系,以及用于组合术语的关系以定义词汇外延的规则,将领域概念化抽象出来的对象、关系和类等用一个词汇集来表达。利用该词汇集,可以表示领域知识。

信息检索和本体技术的结合是目前信息检索领域的研究热点之一,国内外众多科研人员在这一领域进行了探索。

L. Khan 和 D. McLeod 主要研究本体技术在音频检索领域的应用[8,9]。Marco Bertini 等探讨了如何将本体用于视频检索[10]。L. Hollink 等进一步提出为视频检索构建可视化本体的技术[11]。Apple W P Fok 将本体技术应用于个性化教育中的内容搜索,取得了较好的效果[12]。

AT&T 建立的 FindUR 系统[13]是一个应用了本体技术的信息检索系统。通过使用描述逻辑系统 CLASSIC[14]规定的描述逻辑语法来表达 Wordnet[15]中定义的词汇间的同义/

*)本研究得到国家自然科学基金(60572137)资助。胡艳丽 博士研究生,主要研究方向为信息管理、网格。

上义/下义关系，得到简单的背景知识；调用CLASSIC推理系统来完成推理任务，得到某个词汇的同义词集合、上义词集合、下义词集合，对查询输入的词汇使用该词汇的同义词集合和下义词集合来扩展，从而可以实现查询扩展。但该系统从本质上讲仍然是基于语法的，因为它并没有使用本体中的词汇去标记文档，只是强调利用本体来实现查询扩展，而查询输入的词汇本身也并非是依据本体中的词汇来书写的。

中科院计算所智能信息处理开放研究实验室建立的基于本体论和多主体的信息检索服务器[16]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，集成了界面主体、预处理主体、管理主体、信息处理主体和具有移动性的信息搜集主体，并利用本体对文档进行领域分类，同时对用户的查询信息进行规范。

武汉大学信息资源研究中心设计开发了一个基于本体的智能化知识检索原型——Kretrieval，依据本体原理和存储检索领域知识，实现动态启发式概念扩展算法和相关反馈，对情报学和人工智能领域的文献知识进行检索[17]。

文[18]介绍了一个基于本体的信息检索主体MELISA，用于在医疗专业领域检索参考文献。但该研究没有使用形式化的本体语言来建立本体，没有考虑本体的推理问题，导致难以具体实现本体并以之为基础进行推理，对本体的应用还是很粗略的。

分析发现，基于领域知识采用本体技术对检索对象进行高层语义建模，形式化描述概念和概念间的语义关系并提供推理机制；以此为基础标注检索对象，对用户查询进行扩展和优化，消除“一词多义”及“一义多词”等问题对检索的干扰，是深入应用本体技术、提高智能检索性能的关键。

本文提出一种基于领域本体的智能检索模型，采用OWL DL语言构建领域本体；基于文档片段实现较小粒度的标注，提出“主题-概念”两阶段消歧算法提供最接近文档片段语义内容的本体概念映射；基于OWL语言推理实现用户查询概念自动扩展及优化，从而实现深层语义检索。实验证明，与传统关键词检索相比，基于本体的智能检索可以获得更高的查全率(recall)和查准率(precision)。

我们用文档来表示一个信息的单元，文本是它的一种典型的形式，但是文档也可以包含其它的媒体，例如图像、视频和音频。在这里，把文档广义地看成是包括普通文本文档、扩展的多媒体文档、多媒体数据在内的所有形式的数据单元。

2 基于领域本体的智能检索模型

在传统的信息检索模型中，没有具有语义特征的规范词汇集对文档做有意义的标注，通常只是从语法角度出发孤立地抽取索引项，再运用一定的项加权策略(如tf-idf加权策略[19])对索引项赋以权值。这种依赖于词汇语法表现形式的索引不能有效地反映文档的语义，从而不能有效地代表文档；同时，由于缺乏具有语义特征的规范词汇集指导，用户信息需求五花八门且不准确，也不能有效地反映用户信息需求的语义。简单的词形匹配忽略了普遍存在的“一词多义”、“一义多词”和一个语义内容可以有多种表达的问题，语法表示与语义内容的脱离使得检索性能总是不能令人满意。

我们提出基于领域本体的智能检索模型，如图1所示。

首先对领域知识进行语义建模，建立领域本体；对文档进行预处理，按照表达意思的完整性分割成若干文档片段，基于文档片段进行标注；根据标注结果，一个或若干文档片段组成文档对象，其全局唯一标识和标注概念集作为元数据存储到数据库中；用户查询请求在领域本体指导下进行优化和扩展。检索任务并非直接对原始文档和数据进行搜索，而是对元数据库进行搜索。匹配结果通过标识映射到相关文档并返回给用户。

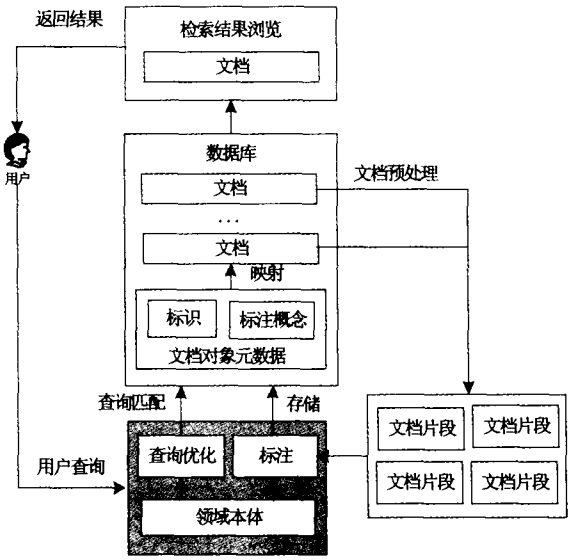

图1 基于领域本体的智能检索模型

本文重点研究领域本体构建、标注和查询优化技术，目标是将文档与用户查询请求抽象为一组具有语义代表性且机器可处理的概念集，基于语义进行检索。

3 领域本体构建

3.1 基本概念

定义1 领域本体可以用一个五元组表示：

$O=\{C, R, H^c, rel, A^o\}$

其中：C表示概念的集合；R表示关系的集合；H^c表示概念层次；rel表示概念间的关系；A^o表示本体公理。

定义2 本体中的概念C可表示为一个三元组：

$C_i=\{Name_i, ID_i, SynList_i\}$

其中，$Name_i$：概念i的名称；ID_i：领域本体中概念i的全局唯一标识；$SynList_i$：概念i的同义词列表。

本文主要考虑三类概念间的基本语义关系：*Kind-of*，*Instance-of*和*Part-of*。

定义3 关系的集合R包含三种关系：

$R=\{Kind\text{-}of, Instance\text{-}of, Part\text{-}of\}$

其中，*Kind-of*关系表示概念间的包含关系。例如：“载人飞船”和“载人航天器”，“载人飞船”是“载人航天器”的一种；除此之外，“载人航天器”还包含“航天飞机”、“航天站”等。

*Instance-of*关系表示概念与实例的关系。例如：“载人飞船”和“神舟六号”，“神舟六号”是“载人飞船”的一个实例。

*Part-of*关系表示概念间部分和总体的关系。例如：“杨利伟”和“神舟六号”，“神舟六号”是一个整体，由宇航员和飞船本身共同组成，“杨利伟”就是“神舟六号”的组成部分。

3.2 本体语言

本文采用描述逻辑来构建本体。相对于语义网(semantic networks)、框架(frame)等非基于逻辑的形式化方法而言，描述逻辑能够精确刻画语义；相对于一阶谓词逻辑推理问题的半可判定性而言，描述逻辑既具有较强的知识表达能力，又

保证推理是可判定的。

目前常见的本体语言有 XOL、SHOE、OML、RDF(S)、OIL、DAML+OIL 和 OWL，这些语言都是以描述逻辑为基础的。其中 OWL 采用 DAML+OIL 作为起点，是对本体语言进行研究的最新成果。

OWL 有三个子集：OWL Lite、OWL DL、OWL Full。OWL Lite 的表达能力最有限，推理效率高；OWL DL 在保证推理的完备性和可判定性的前提下，有尽可能强的表达能力；OWL Full 有最强的表达能力，但不对推理做任何保证。在构建本体的过程中，本文采用高效推理与表达能力兼顾的 OWL DL 作为本体描述语言，并基于 OWL DL 语言给出下列定义。

OWL DL 语言中，概念对应于类(class)，实例对应于个体(individual)。类是一组拥有公共属性的个体集合，个体是类的实例。在标注和查询优化过程中，对于概念和实例的操作是相同的，所以暂不区分，统称为概念。但基于 OWL DL 描述时，两者是严格区分的。

定义 4(Instance -Of)　对于概念 C_i 和实例 d_i，在本体中，如果 d_i 被定义为 C_i 的“Individual”，则称 d_i 是概念 C_j 的实例。

定义 5(Kind -Of)　对于两个概念 C_i 和 C_j，在本体中，如果 C_j 被定义为 C_i 的“subClassOf”，则称概念 C_i 语义包含 C_j。

定义 6(Part-Of)　对于概念 C_i 和实例 d_i，在本体中，如果 d_i 被定义为 C_i 的“subPropertyOf”，则称 d_i 是概念 C_j 的一部分。

3.3　航天领域本体

领域本体是对特定领域内概念和概念之间关系的精确描述。根据上述基本定义，将领域本体中的概念作为节点，概念间的关系作为连接节点的边，则领域本体可用有向无环图(directed acyclic graph，DAG)描述。

图 2 显示了一个航天领域本体的例子，该本体基于通用的航天术语和领域专家的知识建成。

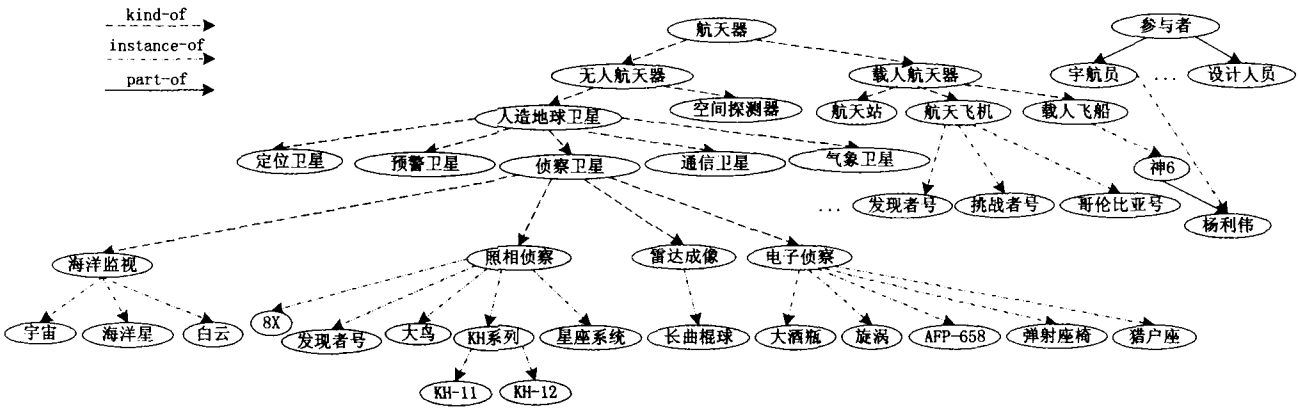

图 2　航天领域本体示例

下面以“杨利伟”和“载人飞船”为例具体描述本体中的概念和实例。

载人飞船：
-Label：airship
-RelationSet：
　-Kind-of：载人航天器
　-Instance-of：NULL
　-Part-of：NULL
-SynList：载人飞船，宇宙飞船，飞船…
杨利伟：
-Label：cosmonaut1
-RelationSet：
　-Kind-of：NULL
　-Instance-of：宇航员
　-Part-of：神六
-SynList：杨利伟、中华飞天第一人，首位中国太空人，航天英雄…

4　标注

标注前对文档进行预处理，按照表达意思的完整性分割成若干文档片段，基于文档片段进行标注，而非全文标注。因为全文标注往往会涉及若干领域多个概念，包含多个主题，不相关的概念可能互相干扰。同时全文标注会使标注概念过多，增加机器的处理时间，难于直接提高检索有效性。这个问题对于长文档标注尤其明显。

4.1　概念标注

标注是建立文档片段和领域本体中概念之间关联的过程，标注后的文档片段根据标注概念集生成新的文档对象或与已有文档对象合并。标注算法如图 3 所示。

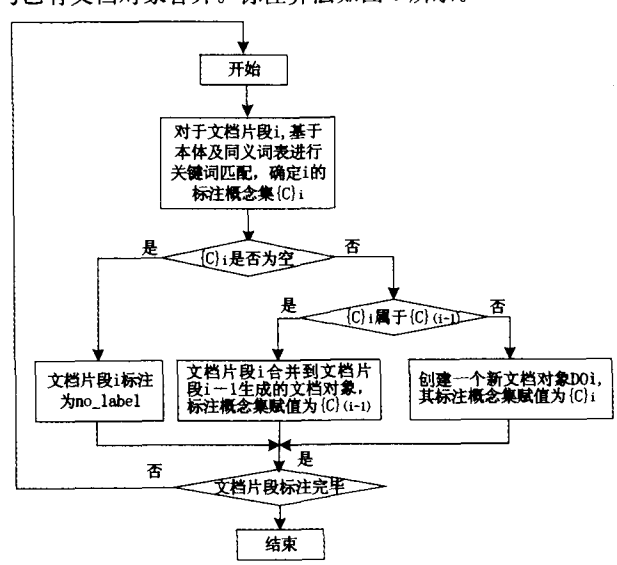

图 3　文档片段标注算法

需要注意的是，本体的构建与领域和构建者本身密切相关，常常由于领域非常复杂或本体构建者缺乏经验等因素，造

成领域本体不完整，或随着领域知识的变化本体需要调整，从而造成领域本体中定义的概念不可能涵盖现实中所有的概念。所以，对于没有标注任何概念的文档片段需要进行手工标注，或者将未标注文档片段合并入与之相邻的文档片段。

4.2 概念消歧

由于自然语言中普遍存在一词多义的现象，即一个关键词可能对应多个概念，或者说不同的概念同义词表中，可能会包含相同的关键词。这将导致标注给文档片段的概念出现歧义概念，即不是文档片段内容真正反映的概念。例如一个文档片段内容为"1985 年 1 月 25 日 14 时 50 分，'发现者'号航天飞机由肯尼迪航天中心发射升空。这是世界上航天飞机第一次执行全军事任务的飞行。除了进行军事试验外，'发现者'号还将释放一颗高级间谍卫星，以便获得军事上有用的情报。航天飞机进行一系列的军事行动，使得其对未来太空战争的重要意义更加明显…"，通过关键词"'发现者'号航天飞机"、"肯尼迪航天中心"、"航天飞机"、"'发现者'号"、"间谍卫星"与本体中概念同义词表的匹配结果，确定的标注概念为"'发现者'号航天飞机"、"肯尼迪航天中心"、"航天飞机"、"'发现者'号侦察卫星"、"间谍卫星"。显然，由于"'发现者'号"存在于多个概念的同义词表中，因此导致了歧义概念"'发现者'号航天飞机"、"'发现者'号侦察卫星"的出现，这会极大地影响检索的准确率。因此概念消歧是标注中的一个关键技术，将文档片段标注的歧义概念消除掉，保留反映文档片段内容的最合适的概念。我们注意到，文档片段内容中出现的关键词单独来看，具有很大的词义不确定性，但如果放在一起，可能就会具有一个特定含义，这种关键词之间相关性所蕴含的含义，为我们进行概念消歧提供了基本的思路。

本体中的一些相关概念组成一个特定的类别，本文称之为"主题(Topic)"。例如"航天飞机"和"侦察卫星"与"航天飞机"相关的概念和实例组成了一个描述"航天飞机"的主题；同样，与"侦察卫星"相关的概念和实例组成了一个描述"航天飞机"的主题。

一般说来，一个文档片段应该对应于一个特定主题，所以概念消歧首先应该是"主题"消歧，即确定适合于对该文档片段进行标注的概念属于本体中的哪个主题。然后在主题内，排除一词多义造成的歧义概念。基于上述想法，本文提出了一个"主题-概念"两阶段消歧算法(简称"T-C"算法)，并通过计算权重的方法实现算法的形式化描述。

设概念 C_i 存在 n 个同义词，其同义词集 $SynList_i=\{l_{i1}, l_{i2}, l_{i3}, \cdots, l_{ij}, \cdots, l_{in}\}$。

定义 7 同义词权重($SWeight_{ij}$) 概念 C_i 的同义词 l_{ij} 的权重等于 l_{ij} 在该概念标注的文档片段中出现的次数($Num\ of\ l_{ij}\ of\ C_i\ Matched$)与它在本体的概念同义词集中出现的总次数($Total\ Num\ of\ l_{ij}$)之比，如下式：

$$SWeight_{ij}=\frac{Num\ of\ l_{ij}\ of\ C_i\ Matched}{Total\ Num\ of\ l_{ij}}$$

定义 8 概念权重($CWeight_i$) 概念 C_i 的权重等于它的所有同义词权重中的最大值。

$$CWeight_i=\max SWeight_{ij}\ \text{where}1\leqslant j\leqslant n$$

定义 9 语义距离 SD(C_i,C_j) 概念 C_i,C_j 之间的语义距离为本体中两个概念间的最短路径。

定义 10 相关概念(Cor-Concept) 本体中与概念 C_i 的语义距离不等于无穷的概念。

文档对象标注过程中出现的相关概念为正确理解文档含义，进行有效消歧提供了有意义的上下文环境。如"'发现者'号航天飞机"与"肯尼迪航天中心"、"航天飞机"相关，那么消歧时定义增值权重量化相关概念的影响。

定义 11 增值权重(EWeight) 假设概念 C_i 的相关概念集为$\{C_k \mid k=1,2,\cdots,r\}$，则 C_i 的增值权重等于 C_i 的概念权重加上所有的相关概念权重除以与 C_i 之间的语义距离 $SD(C_i, C_k)$之和，即

$$EWeight_i=Weight_i+\sum_{k=1}^{r}\frac{Weight_k}{SD(C_i,C_k)}$$

由增值权重的定义可知，若标注概念集中 C_i 的相关概念越多，语义距离越小，那么其增值权重越大，与文档对象含义的相关性越大。

在上述定义的基础上，T-C 消歧算法可描述如下：

```
Algorithm T-CDisam (DocObject DO)
1  Begin
2    //get the annotation concept set of DO
3    ConceptSet CS=getAnnotationConceptSet(DO);
4    //Determine the possible topics according to CS
5    TopicSet TS=getTopicSet(CS);
6    for each (Tm∈TS)
7      float TWeightm=0;
8      //get all possible concepts of topic Tm in CS
9      ConceptSet TCSm=getTopicConceptSet(CS);
10    for each (Cn∈TCSm)
11    if (Cn is ambiguous concept)=true
12        //calculate the average weight of C with its
13        //ambiguous concepts
14        ConceptSet CS
15                  =getAmbiguousConceptSet(Cn);
16        j=getCount(ACS);
                                  sum_{i=1}^{j} CWeight_i
17        AverageCWeight_n=-----------------------
                                        j
18        TWeightm=Tweightm+AverageCWeightn;
19        TCS=TCS-ACS;
20      else
21        TWeightm=Tweightm+CWeightn;
22  //end of for each (Tm∈TS)
23  DocTopic DocT=getTopic(max(TWeightm));
24  CS=TCSm;
25  for each(Ci∈CS)
26    //calculate Eweight of each concept
                                    sum_{k=1}^{r} Weight_k
27    EWeight_i=Weight_i+-----------------
                                  SD(Ci,Ck)
28  maxWeight=max(EWeighti);
29  for each(Ci∈CS)
30  if (EWeighti<maxWeight * V)
31      CS=CS-{Ci};
32    return CS;
33  End.
```

对文档片段标注完成后，每一个片段对应一个或几个本体中的概念，并使用概念 Label 来具体描述该文档片段。

5 基于本体的查询优化机制

信息检索时，往往由于用户经验不足、缺乏领域知识、输入的查询式过于简单或查询处理方法存在缺陷等原因，导致形成的查询请求存在各种问题，如不能真实反映用户的实际检索需求、存在不一致性和重复性、内容不够全面等，难于形成有效的检索，从而大大影响知识检索的质量。

用户查询也可视为一个简单的文档对象。对用户原始查询请求进行优化，构造合理有效的扩展是开展有效查询的重要环节。面向自然语言查询是目前检索系统的一个重要发展趋势，因此我们假定用户的查询输入以汉语语言方式提交。

查询优化主要从下述两个方面进行：

(1)查询概念消歧

与文档片段相同，用户提交的查询中也会存在歧义概念，因此，采用 T-C 概念消歧算法对查询概念进行消歧。采用本体概念同义词表中关键词匹配，获得用户查询语句中包含的概念(称为查询概念——Query Concept，QC)，查询概念组成

的概念集称为初始查询概念集。

(2)查询概念扩展

用户提交的查询概念通常暗含了用户对该概念集合包含元素的关心，查询结果应该能自动满足用户的这一潜在需求。为了充分理解和正确表示用户的查询请求，基于本体对用户检索需求进行优化和扩展。

通过分析，我们提出一种基于 OWL 推理的查询优化算法，具体描述如下：

```
Algorithm QueryOptimization (DocObject DO)
1  Begin
2    //采用 T-C 消歧算法对查询概念消歧
3    ConceptSet CS= T-CDisam (DO);
4    for each(Ci∈CS)
5      OntoClass class=getOntologyClass(Ci);
6      if (Individual(class) is not null)
7        CS=CS∪Individual(class);
8      if (subClassof(class) is not null)
9        CS=CS∪subClassof(class);
10     if (Individual(subClassof(class)) is not null)
11       CS=CS∪Individual(subClassof(class));
12   return CS;
13 end;
```

如查询“载人航天器”的相关资料。“航天装备”本体中对“载人航天器”的 OWL 描述如下：

```
〈owl:Class rdf:ID="载人航天器"〉
  〈owl:0bjectProperty rdf: ID ="hasDriver "〉
    〈rdfs: domain rdf: resource="#载人航天器"/〉
    〈rdfs: range rdf: resource="& 宇航员"/〉
  〈/owl: ObjectProperty〉
  〈owl:Class rdf:ID="航天站"〉
    〈rdfs: subClass0f rdf: resource="#载人航天器"/〉
    〈航天站 ref: ID= "'礼炮'号"/〉
    〈航天站 ref: ID= "'钻石'号"/〉
    〈航天站 ref: ID= "'和平'号"/〉
    …
  〈/owl:Class〉
  〈owl:Class rdf:ID="航天飞机"〉
    〈rdfs: subClass0f rdf: resource="#载人航天器"/〉
    〈航天飞机 ref: ID= "'发现者'号"/〉
    〈航天飞机 ref: ID= "'挑战者'号"/〉
    〈航天飞机 ref: ID= "'哥伦比亚'号"/〉
    …
  〈/owl:Class〉
  〈owl:Class rdf:ID="载人飞船"〉
    〈载人飞船 ref: ID= "神舟六号"〉
      〈hasDriver rdf: resource="& 杨利伟"/〉
      …
    〈/载人飞船〉
    …
  〈/owl:Class〉
〈/owl:Class〉
```

其中关于杨利伟的描述来自“航天活动参与者”本体中对“杨利伟”的 OWL 描述：

```
〈owl:Class rdf:ID="宇航员"〉
  〈宇航员 ref: ID= "杨利伟"/〉
  …
〈/owl:Class〉
```

推理机通过推理得到“载人航天器”，包括“航天站”、“航天飞机”、“载人飞船”等子类，查询时这些概念自动扩展为查询概念。本文使用 RACER[20] 作为描述逻辑推理器，采用 nRQL[21] 作为查询语言。

6 实验评测

随着网上学术信息资源的丰富，大量的网络图书、网络期刊、会议论文、学术论文、技术报告、BBS、网络论坛为学术研究提供了重要参考。

信息检索模型实用性的验证一般是使用公认的测试参考文档集(如 TREC、CACM、CF 等)来进行试验，比较不同模型的查准率和查全率，分析模型的优劣。但目前还没有使用本体中的词汇标记的公认测试参考文档集，所以本文收集了来自 Internet、万方数据库、中国期刊网、维普期刊数据库的 3500 篇文档作为检索实验的数据源，主要集中于航天领域。具体数据如表 1 所示。

表 1 测试数据源

序号	文档名称	来源	数量
1	网络期刊/报道	Internet	1000 篇
2	学位论文	万方数据库	500 篇
3	期刊	CNKI 中国期刊论文库	1000 篇
4	期刊	维普期刊数据库	1000 篇

标注时对文档进行编号，依据作者组织文章的章节划分为若干片段。存储时，为每个文档对象提供全局唯一的标识号，由文档编号和该对象包含的片段编号共同组成。检索时根据标识号可直接定位到所属文档并返回结果。

结合航天领域相关知识建立了包含 500 多个概念的航天领域本体，并进行标注、存储。采用查准率和查全率作为评价检索性能的标准，其中查准率定义为查询结果中的相关文档数与查询结果包含的文档总数之比，查全率定义为查询结果中的相关文档数与测试集中的相关文档总数之比。通过分析查准率和查全率，验证本文提出的基于本体的信息检索模型的实用性。

为了保证试验的一般性，邀请非专业人员和专业人员提出了 8 个查询请求来进行实验验证。用户查询如表 2 所示。

表 2 用户查询语句列表

序号	用户查询
1	“查询关于神舟系列飞船的新闻”
2	“查询宇航员约翰. 格伦的经历”
3	“查询关于人类登月计划的实施情况”
4	“查询关于“发现者号”发射的情况”
5	“查询关于历届航空航天展览的情况”
6	“查询航空航天设施使用的材料”
7	“查询太空生物技术的发展现状”
8	“查询航天服设计的技术工艺”

在选定测试集的基础上，基于本体的检索实验结果与网上资源本身提供的基于关键词的检索工具获得的检索结果进行比较，如图 4、图 5 所示。

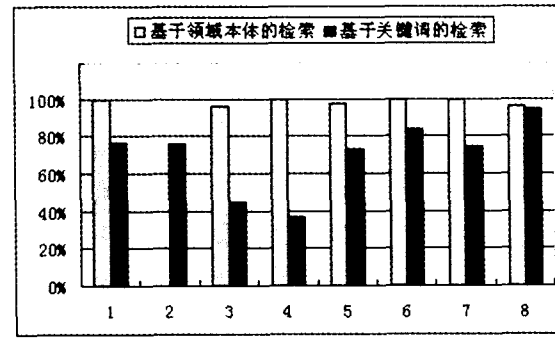

图 4 基于领域本体与基于关键词检索的查全率对比图

从实验结果可以发现：

1)与基于关键词的方法相比，基于本体的检索方法的查准率和查全率基本上是令人满意的，说明了该方法的有效性。可以看到基于本体的检索方法相对于基于关键词的检索方法其查准率有了大幅提升，查全率也有一定提高或是相当。查准率的提高是因为领域本体使文档含义与用户理解建立在统

一的语义基础之上，通过领域知识指导文档标注和概念消歧，减少了“一词多义”和“一义多词”等问题对查询的干扰，因此能够有效提高查询的准确率。查全率的提高是由于基于本体的检索方法利用本体对查询概念进行了优化扩展，相对于简单、单一的关键词可以更加全面地描述用户的查询需求，较好地反映了语义层次的信息检索与语法层次的信息检索的区别，即语义层次的信息检索依赖于词汇的语义而非表现形式。

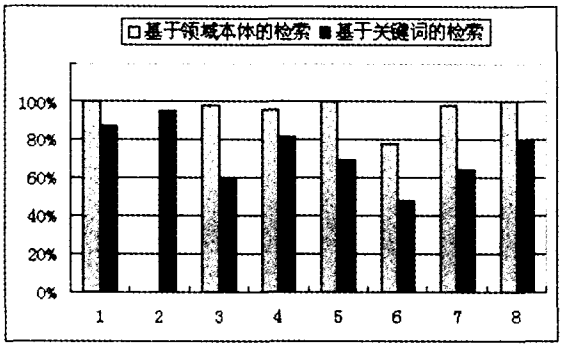

图5 基于领域本体与基于关键词检索的查准率对比图

例如对于查询1，CNKI中国期刊论文库查不到文档“神5搭载物亮相：奥运旗/人民币票样/台湾种子”，而基于本体的检索通过在“神舟五号”的同义词列表中包含“神5”成功查到这篇文档。而且由于现有检索工具主要是基于关键词的匹配，在查到“神舟系列飞船”的同时还查到“神舟笔记本”、“神舟”牌MP3的资料，从而降低了查准率；而基于本体的检索在消歧过程中通过计算“神舟”和“笔记本”的增值权重可以排除该文档属于航天领域这个主题。

同时我们注意到，在查询概念意义明确、不易产生歧义的情况下，消歧和查询优化扩展的效果并不明显，两种技术的查全率基本相当，如查询8。

2) 对于某些查询，基于本体的检索技术的结果并不尽如人意。如查询2，出现了查全率和查准率都为0的极端情况，这主要是本体的不完整性造成的。

如前所述，由于本体构建者对于领域知识的掌握和理解存在不完全性，或领域本身的复杂性导致建立本体的工作量很大，因此领域本体构建不可能或难于涵盖现实中所有的概念和关系。如美国的约翰·格伦是历史上年龄最大的宇航员，但开始由于忽略了该知识，本文的航天领域本体没有包含相关概念和知识，因此查询失败。这就需要扩展本体的概念，随着领域知识的扩展加入新概念。

结论 本文提出了一种基于领域本体的智能检索模型，将本体技术应用于信息检索的关键技术中，提供了一种深层语义的、基于概念的检索方法。

本文以航天领域为背景，对领域知识进行高层语义建模，设计了航天领域本体的构建方法，基于OWL DL语言实现对概念和概念间关系的形式化描述，建立了对文档内容和用户查询进行统一描述的语义基础。提出基于文档片段的标注方法，保持表达意思完整的同时实现细粒度标注，提高处理效率，减少噪声干扰，尤其适用于长文档标注。针对“一词多义”问题，提出“主题-概念”两阶段消歧算法，在标注基础上获得最接近文档片段语义内容的本体概念映射，有效地解决了概念歧义问题。本体中的概念及其同义词表为用户查询优化提供了具有语义特征的规范词汇集，基于OWL描述的推理机制利用概念和概念间的关系可以实现查询概念的自动扩展。利用网上数据库开放资源作为测试集进行评测，结果表明，与传统关键词检索相比，基于领域本体的检索方法可以获得更高的查全率和查准率。

目前，基于本体的检索技术研究方兴未艾，本文进行了初步探讨，还有很多问题值得研究。例如，本体的构建与领域和构建者本身密切相关。为了提高基于本体技术的动态性和适应性，如何采用本体进化技术改进领域本体不完整或随着领域知识的变化本体需要调整的问题，将是我们下一步研究的重点。

致谢 感谢本文合作者白亮在概念消歧方面的研究工作。

参考文献

1 WU Zhao-Hui, XU Jie-Feng. Knowledge base Grid: A Generic Grid Architecture for Semantic Web. J Comput Sci. & Technol, 2003,18(4):462～473
2 Zhuge H. THE KNOWLEDGE GRID. World Scientific Publishing Co, Dec. 2004
3 Berman F. From TeraGrid to Knowledge Grid. Comm ACM, 2001,44(11):27～28
4 De Roure D, Shadbolt N. The Semantic Grid: A Future e-Science Infrastructure. In: A. J. G. H. F. Berman, Fox G, eds. Grid Computing: Making the Global Infrastructure a Reality, John Wiley & Sons, 2003. 437～470
5 Cannataro M. Semantics and Knowledge Grids: Building the Next-Generation Grid. IEEE Intelligent Systems, 2004,19(1):56～63
6 Moore R. Knowledge-Based Grids: [tech report]. SDCS TR-2001-2. San Diego Supercomputer Center: San Diego, Calif, 2001
7 Moore R W. Knowledge-Based Grids: Two Use Cases. Sept. 2001. http://www-itg.lbl.gov/GPA/Moore.GGF-3.pdf
8 Khan L. Audio Structuring and Personalized Retrieval Using Ontologies. In: Proceedings of IEEE Advances in Digital Libraries, Library of Congress, Washington, DC, May 2000
9 Khan L. Disambiguation of Annotated Text of Audio Using Onologies. In: Proc. of ACM SIGKDD Workshop on Text Mining, Boston, MA, 2000
10 Bertini M, Torniai C. Enhanced Ontologies for Video Annotation and Retrieval. In: 2005 ACM Multimedia Conference, Singapore November 10-11, 2005
11 Hollink L, Schreiber A Th. Building a Visual Ontology for Video Retrieval. In: 2005 ACM Multimedia Conference, Singapore November 2005
12 Fok A W P. Ontology-driven Content Search for Personalized Education. In: 2005 ACM Multimedia Conference, Singapore November 2005
13 Baader F, Nardi D, et al. The Description Logic Handbook: Theory, Implementation and Applications. UK: Cambridge Univ. Press, 2003. 436～459
14 Brachman R J, Patel-Schneider P F, et al. Living with CLASSIC: When and how to use a KL-ONE-like language. Principles of Semantic Networks. Los Altos: Morgan Kaufmann, 1991. 401～456
15 Miller G A. WordNet: A lexical database for English. Communications of the ACM, 1995,38(11):39～41
16 武成岗，焦文品，田启家，等. 基于本体论和多主体的信息检索服务器. 计算机研究与发展，2001，38(6)：641～647
17 周宁，张玉峰，张李义. 信息可视化与知识检索. 科学出版社，2005
18 Abasolo JM. MELISA: An ontology-based agent for information retrieval in medicine. In: Proceedings of the First International Workshop on the Semantic Web (SemWeb2000), Lisbon, Portugal, 2000
19 李国辉，汤，武德峰. 信息组织与检索. 科学出版社，2003
20 Haarslev V. Description of the racer system and its applications. In: Proceedings International Workshop on Description Logics (DL-2001), Stanford, USA, Aug. 2001
21 Haarslev V, Wessel M. Querying the semantic Web with racer + nrql. In: Proceedings of the KI-2004 International Workshop on Applications of Description Logics (ADL'04), Ulm, Germany, September 2004