语义知识库存储方案研究

殷 浪
（武汉理工大学 计算机科学与技术学院，湖北 武汉 430063）

摘 要： 讲述了目前语义知识库的一些存储方案，针对由 Lehigh 大学提出的语义 Web 数据测试集 LUBM，选取了其提供的 14 种查询中的 3 个做了相应实验，分析比较了知识库的各种不同存储方案之间对于查询相应时间和存储空间的差别。实验结果表明，与其他存储方案相比，采用 TDB 存储方式能大幅度提高用户检索的效率，并且降低了存储空间。

关键词： 知识库；TDB；语义检索；存储方案

中图分类号： TP391 **文献标识码：** A **文章编号：** 1674-7720(2012)09-0053-03

Research on storage scheme of the semantic knowledge base

Yin Lang
(College of Computer Science and Technology, Wuhan University of Technology, Wuhan 430063, China)

Abstract: This paper describes the storage scheme of the current semantic knowledge base system. It choose three of fourteen query as the experimental query which are supplied by a semantic web data test set LUBM proposed by the Lehigh University, compares the query time and the storage space of different kinds of storage schema of the knowledge base. Finally, the experimental results show that the TDB storage schema can improve the retrieve efficiency and reduce the storage space.

Key words: knowledge base; TDB; semantic retrieve; storage scheme

随着网络的发展，当今社会已经步入了信息时代。网络资源呈指数增长，互联网已成为一个巨大的信息源，如何提高检索质量，寻求令人满意的检索模式和技术已经是全球的研究重点。当前基于关键词的信息检索由于数据缺乏语义信息及其查询处理缺乏语义支持，只能查找出与用户在语法层上匹配的信息，而无法给出与其在语义层上具有相关性的其他信息，从而导致信息查询结果质量低下。Tim Berners-Lee 提出了语义网，使网上信息提供具有计算机可以理解的语义，它的发展和成熟使得高效、高质的语义检索成为可能，以克服基于关键词的信息检索技术的缺陷。但是基于知识库的语义查询通常比较缓慢。本文研究了语义知识库的相关存储方案，并采用 LUBM 测试集从查询时间和存储空间这两个方面比较几种不同的存储方案[1-3]。

1 语义检索技术

1.1 Protégé

Protégé 是一款基于 Java 的图形界面本体工具，是由美国斯坦福大学开发的免费开源平台。它为用户提供了一系列的工具支持构建领域本体模型和基于本体的知识库应用，常用于定义本体模式。

1.2 Jena

Jena 是惠普实验室提供的针对语义 Web 应用的开源 Java 开发包[4]。它支持应用程序调用 Jena 提供的接口操作本体数据；支持主流的本体描述语言，如 RDF、RDFS 和 OWL；支持多种本体的存储模型，如内存模型和数据库模型等。Jena 常用于定义实例并对本体数据进行查询维护等。本文将使用 Jena API 进行相关查询分析。

1.3 Sparql

Sparql(Simple Protocol and RDF Query Language)是为 RDF 开发的一种查询语言和数据获取协议，它为 W3C 所开发的 RDF 数据模型所定义，但是能用于任何可以用 RDF 来表示的信息资源[5]。RDF 的三元模式构成了图形模式，而 Sparql 的查询解决方案试图将每个图形模式变量的绑定与查询模型节点进行匹配。

Sparql 协议和 RDF 查询语言(Sparql)目前是 W3C 的

工作草案或推荐标准，还在讨论中。Sparql 构建在以前的 RDF 查询语言（例如 rdfDB、RDQL 和 SeRQL）之上，拥有一些有价值的新特性。

2 本体知识库存储方案

2.1 文件系统

Jena 可以在文件系统中持久化本体知识库，即基于文件系统的存储。该方式实现起来比较简单，很多本体相关工具都支持对文件格式的本体进行存取。但是，这种方法不仅效率低，而且很难适应数据量较大的情况。基于文件系统的存储方式一般只适用于规模较小的本体。

早期的本体数据管理工作是基于文件系统实现的，它们用简单的文件格式存储本体数据并支持一些基本的操作。这类工作主要用来编辑和建立本体，并不是为大规模本体数据的存储和查询管理服务的，如 Protégé。

2.2 关系数据库

由于关系数据库技术发展成熟，大多数现有的本体数据管理工作使用关系或对象-关系数据库管理系统作为后台存储。Jena 就可以在关系数据库（Relational Database）中持久化本体知识库。当前支持的数据库引擎有 Oracle、PostgreSQL 和 MySQL。以 MySQL 为例，下面的代码说明了如何将 OWL 文件导入到 MySQL 持久化模型。

```
Class.forName("com.mysql.jdbc.Driver");
DBConnection connection=new DBConnection(DB_URL, DB_USER, DB_PASSWORD, DB_TYPE);
ModelMaker maker = ModelFactory.createModelRDBMaker(connection);
Model model=maker.createModel(modelName);
model.begin();
InputStream in=new FileInputStream(new File(owlfile));
model.read(in, null);
commit();
```

在持久化到数据库后，可以通过 ModelMaker.openModel(modelName) 来访问该模型。

2.3 TDB

TDB 是 Jena 的一个组件，可大规模地存储和查询 RDF 数据集，且支持 Sparql 查询[6]。TDB 是一个具有高性能、非事务性的 RDF 数据存储器，可以通过命令脚本和 Jena API 来访问和管理 TDB 存储。以下代码是说明如何将 OWL 文件存储为 TDB 的。

```
DatasetGraphTDB graph =TDBFactory.createDatasetGraph(TDBlocation);
TDBLoader.load(graph, "file:///"+owlfile);
```

3 实验设计和性能评估

3.1 实验设计

3.1.1 硬件环境

本实验测试是在个人电脑上进行的。具体环境是：2.20 GHz Intel(R)Core(TM)2 Duo CPU T6600，2 GB 内存，250 GB 的硬盘，Windows XP 操作系统，Java SDK 1.6.1。

3.1.2 测试数据集

LUBM 是 Lehigh 大学提出的语义 Web 数据测试集。它基于大学这个领域，采用机器自动生成的数据作为测试数据，提供 14 个测试查询和一套性能指标[7]。它可以根据用户指定的参数产生不同规模的数据，由此测试在不同规模的环境下，系统的实例查询性能。LUBM 测试集是目前最流行的语义 Web 测试集。它生成的数据满足本体层的规范，因此，也可以作为推理系统的测试数据集。但是 LUBM 测试结果也存在一个问题，即生成的数据中属性的个数是固定的，仅有 64 个。随着数据量的增加，数据会失去语义 Web 的一大特点——稀疏性，导致测试的结果不能反映实际应用的效果。

这 14 个测试查询，有的涉及推理机，由于篇幅有限，只做了部分测试。以下是 3 个测试查询语句。

（1）Query1，查询所有参加课程 <http://www.Department0.University0.edu/GraduateCourse0> 的学生。

```
SELECT? X WHERE
{? X rdf:type ub:GraduateStudent.
? X ub:takesCourse <http://www.Department0.University0.edu/GraduateCourse0>}
```

（2）Query2，查询学生及其所在大学和所在系。

```
SELECT? X? Y? Z WHERE
{? X rdf:type ub:GraduateStudent.? Y rdf:type ub:University.? Z rdf:type ub:Department.
? X ub:memberOf ? Z. ? Z ub:subOrganizationOf? Y? .? X ub:undergraduateDegreeFrom? Y}
```

（3）Query3，查询在 <http://www.Department0.University0.edu> 教学的所有教授的信息，包括姓名、email 地址和电话。

```
SELECT? X? Y1? Y2? Y3 WHERE
{{? X rdf:type ub:FullProfessor.}UNION {? X rdf:type ub:AssociateProfessor.}
UNION { X rdf:type ub:AssistantProfessor.}
? X ub:worksFor <http://www.Department0.University0.edu>.? X ub:name? Y1.
? X ub:emailAddress? Y2.? X ub:telephone? Y3.}
```

3.2 实验结果对比

库容量和转载时间的比较如表 1 所示。其中，库容量是指各种不同的存储方式所占用的磁盘空间的大小；转载时间是指从文件形式的知识库转换到其他存储方式所需要的时间。

由于关系型数据库会保存知识库中所有的三元关系，因此耗时会比较多。对于 1 个 50 MB 左右的 OWL 文件，就已经耗时 4 个多小时。因此，如果是较大的本体知识库，想借关系数据库来改善检索效率的话，其可行性需要斟酌。相对于关系数据库，TDB 所用时间要少很多，值得借鉴。

表 1 库容量和转载时间比较

	数据集	实例数量	库容量/MB	转载时间/s
文件系统			8.02	–
关系数据库	LUBM(1,0)	103 074	22.7	2549
TDB			14.8	23
文件系统			50.3	–
关系数据库	LUBM(5,0)	645 649	141	15533
TDB			91.7	127
文件系统			102	–
关系数据库	LUBM(10,0)	1 316 322	295	90154
TDB			187	260
文件系统			540	–
关系数据库	LUBM(50,0)	6 888 642	–	–
TDB			996	1500

对于上面提到的 3 个 Sparql 查询语句，在用文件系统、关系数据库和 TDB 这 3 种不同的存储方式存储时，查询所消耗的时间和查询结果如表 2 所示。

表 2 查询测试结果

查询		LUBM(1,0) 文件系统	LUBM(1,0) 关系数据库	LUBM(1,0) TDB	LUBM(5,0) 文件系统	LUBM(5,0) 关系数据库	LUBM(5,0) TDB	LUBM(10,0) 文件系统	LUBM(10,0) 关系数据库	LUBM(10,0) TDB	LUBM(50,0) 文件系统	LUBM(50,0) TDB
1	耗时/ms	150	120	94	609	350	193	1 342	1 098	702	–	2 310
	结果	4	4	4	4	4	4	4	4	4	–	4
2	耗时/ms	170	140	109	702	521	335	1 589	1 235	989	–	–
	结果	0	0	0	9	9	9	28	28	28	–	–
3	耗时/ms	195	163	114	809	611	398	1 850	1 545	1019	–	5 780
	结果	34	34	34	34	34	34	34	34	34	–	34

由表 2 可知，TDB 在查询方面要比文件系统和关系数据库的效率高。

针对目前语义检索领域中基于文件或者关系数据库存储方案下检索效率慢的问题，本文分析了这几种存储方案在查询响应时间和存储空间上的区别，并提出了基于 TDB 的知识库存储方案。实验证明，该方法能较大程度上提高用户检索效率，并且能降低存储所需空间。基于本体的语义检索的知识中，推理机还没有涉及。如果添加了推理机，语义检索的速度将会更慢，因不属于本文研究内容，故没有作比较。

本体知识库的存储方案其实还有很多方式，如 4store、BigData 和 BigOwlim 等。由于能力有限，无法对每种方式进行比较，只对研究项目所用到的存储方式比较。这些将是以后研究工作中的重点。

参考文献

[1] JARRAR M, MEERSMAN R. Ontology engineering –the DOGMA approach [C]. Advances in Web Semantics I. Lecture Notes in Computer Science, 2009,4891:7–34.

[2] MILLER E. Semantic web applications [J]. INTAP Interoperability Technology Association for Information Processing, 2003(34):210–212.

[3] GRUBER T R. A translation approach to portable ontologies[J]. Knowledge Acquisition, 1993,5(2):199–220.

[4] 栾艳，丁二玉，骆斌. 基于 Ontology 的语义检索技术[J].计算机工程与应用，2005,28(41):156–159.

[5] 于水明. 基于本体的语义检索的应用研究[D].大连:大连海事大学,2007.

[6] 谢圣献，谢光.语义检索在电子商务中的应用研究 [J]. 微计算机信息，2008,24(12):50–56.

[7] Gao Yuanbo, Pan Zhengxiang, HEFLN J. An evaluation of knowledge base systems for large owl datasets[C]. Third International Semantic Web Conference, 2004:6–7.

（收稿日期：2012–01–14）

作者简介：

殷浪，男，1987 年生，硕士研究生，主要研究方向：软件工程。