

基于本体的信息检索系统的设计与实现*

付燕宁¹, 金龙飞¹, 王开锋², 刘磊¹

(1. 吉林大学 计算机科学与技术学院, 吉林 长春 130012; 2. 北京航空航天大学 计算机学院, 北京 100083)

摘要: 详细地描述了使用语义 Web 技术实现智能信息检索的思想和工作流程, 提出了一个基于本体的信息检索系统的框架, 并以一个地理信息领域的应用为例分析了该检索系统的实现技术。

关键词: 信息检索; 语义 Web; 本体

中图法分类号: TP311 文献标识码: A 文章编号: 1001-3695(2006)11-0155-03

Design and Implementation of Ontology-based Information Retrieval System

FU Yan-ning¹, JIN Long-fei¹, WANG Kai-feng², LIU Lei¹

(1. College of Computer Science & Technology, Jilin University, Changchun Jilin 130012, China; 2. College of Computer, Beijing University of Aeronautics & Astronautics, Beijing 100083, China)

Abstract: This paper describes design idea and process of the intelligent information retrieval based on semantic Web technologies in detail, and proposes a framework of an ontology-based information retrieval system. What's more, this paper also provides analysis of the implementation technologies based on an application of geographical information domain.

Key word: Information Retrieval; Semantic Web; Ontology

因特网的广泛普及使得网上信息的数量和类型随之急剧增加, 从而造成用户查找信息越来越困难。如何高质高效地检索到所需的信息, 是当前信息检索领域迫切需要解决的问题。

目前因特网的主要信息检索方式是: 用户向搜索引擎公司提供的服务器发出指令, 由服务器通过基于关键字匹配技术检索出相关的内容, 并通过一定的排序法则呈现在用户面前。这种检索技术具有以下几点不足: ①检索结果的数量过于庞大; ②产生许多无价值的垃圾信息, 准确率没有保障; ③检索算法采用词形匹配而非词义匹配, 查全率难以保证。产生这些问题的根源在于关键字对于计算机而言没有任何含义, 或者说没有语义, 因而检索的结果不能完全满足用户检索的意图。语义 Web^[1]技术的出现可以改变这一现状。在语义 Web 中, Web 上的信息均被赋予了定义良好的含义, 能够被机器理解和处理, 这使得 Web 上的自动信息处理成为可能, 检索的全面性和准确性也随之大大提高。

本文提出一个基于本体的信息检索系统框架, 它能够提取和利用 Web 上的语义信息并根据用户的检索条件进行推理, 进而得出较为准确的结果。该框架具有以下特点: ①具有语义推理功能。使用推理机制扩展查询, 使得用户能够准确、全面地找到自己感兴趣的信息。②提供灵活多样的信息检索方式。用户可以根据需要和自己的喜好选择合适的检索方式。③检索结果以用户便于理解的方式排序。例如, 对于用户最有价值的结果以较为显著的形式展现出来。④友好的界面。便于用户快速地掌握本系统的使用方法。

1 信息检索系统的基本框架

1.1 功能描述

依据语义 Web 的观点, 将来 Web 上所有可用的信息均有合理的结构、标注和约束, 因此在描述本检索系统的体系结构之前, 首先作出这样的假设: Web 上的信息都已经进行了合理的标注, 能够被人和机器理解。基于上面的假设, 本信息检索系统的结构如图 1 所示。

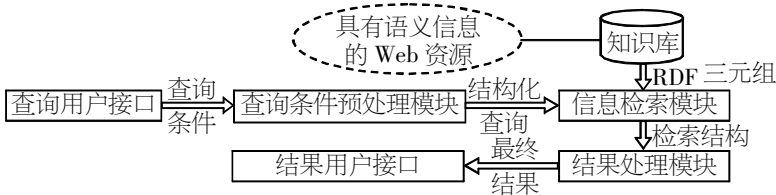


图 1 信息检索系统的体系结构

本系统的工作流程如下: ①建立起一个与信息源相关联的知识库, 并且将知识库中的信息编码成 RDF/XML 格式提供给信息检索模块, 以便计算机能够高效地处理这些信息。②用户或应用软件由查询用户接口输入查询条件, 查询条件预处理模块将查询条件转换为结构化查询条件, 提交给信息检索模块。③信息检索模块依据知识库提供的信息执行查询, 返回一系列满足查询条件的实例。④这些实例经过结果处理模块处理后, 返回给用户。

1.2 知识库

知识库是关于实体的形式化知识的实现, 它在整个系统中占有核心地位。知识库中包含着一个或几个域本体, 它们描述信息源中出现的概念, 知识库通过本体与信息源相关联。知识库中保存着类层次, 它是对 Web 上资源进行标注、分类的依据。例如在地理信息领域, 土地、水面、冰川、湿地均被定义为

土地类型的子类,这些类可以用于创建与 Web 资源相关的具体实例。此外知识库中还保存着真正文档所在的物理地址,在本文的系统实现中,它们由指向 Web 上文档的 URI 构成。

在知识库中,本体的建立是关键。本体是信息的组织框架,在原始信息的收集、资源对象元数据的提取和信息检索中都需要以领域本体为基本依据。本体由概念、概念的属性以及概念和约束之间的关系构成,它反映了对特定领域的语义的通用理解。一个通用的本体可以在不同的软件实体间交换并共享信息。不同领域有不同的领域本体,一个领域中包含的知识和词汇非常丰富,建立一个恰当的领域本体是十分重要的。这个工作通常是由熟悉某一领域的问题域专家完成。

本体中的知识和信息采用本体语言描述,本体语言使得用户能够清晰地、形式化地为领域模型构建本体。在该领域诞生了许多种本体描述语言,其中与 Web 相关的有 RDF/RDFS^[2], OIL^[3], OWL^[4~6]等。为了使计算机能够高效地处理语义信息,本系统采用 W3C 所发布的资源描述框架(RDF)作为元数据编码参考模型。知识库中的所有信息均被编码成 RDF/XML 格式。RDF 模型的核心是三元组描述,即任何复杂的事务描述均可以简化成一系列的三元组描述, RDF 描述的核心是事物与其他事物之间的联系。

1.3 信息检索模块

信息检索模块接收满足一定检索规范的结构化查询,并依据知识库中的 RDF 三元组执行检索,得到满足条件的结果。本模块主要具有推理和查询两种功能。为了在信息检索时能够向用户返回更多满足用户需求的检索结果,本模块首先要针对用户提供的检索条件,依据知识库进行相关的推理,这种推理是基于类层次和规则的,使用推理机制可以扩展查询,使得信息检索的全面性和准确性大大提高;然后依据推理结果提取出满足条件的实例,这部分功能相对较为简单,仅需要将经过处理后的查询条件和知识库中的信息进行匹配,将满足条件的元数据选出,交给后续部分处理即可。

有许多工具可以用于建立该模块。其中 Racer 是 Ralf M ller 和 Volker Haarslev 从 1999 年开始开发的系统; Jena 是惠普(HP) 实验室开发的一个构建语义 Web 应用程序的 Java 框架,两者均具有推理和对知识库的查询功能。Jena 是一个以 RDF API 为核心,用于实现语义 Web 的工具。它为 RDF, RDFS 和 OWL 提供了一个可编程的环境,允许用户来创建、解析和查询 RDF 模型,也提供对 OWL 和 RDF 的推理功能。Racer 是描述逻辑推理机。就推理而言 Jena 自身提供的是基于规则的推理机,性能上不如 Racer, Racer 在所有的描述逻辑推理机中性能比较出众,可以支持工业使用;但 Racer 使用的是 DIG 接口,需要通过 HTTP 或是 TCP 访问,所以如果知识库不是很复杂的话,可能会由于通信的时间消耗反而使性能下降。另一方面 Racer 在一定程度上来说只是单纯的推理机,它不对知识库的存储作很多考虑,而 Jena 对知识库的存储乃至持久化的支持都很完整。

考虑到设计本系统是为了研究需要而不是工业应用,权衡两者的优缺点,将使用 Jena 来实现信息检索模块。

1.4 查询用户接口与查询条件预处理模块

用户由查询用户接口提交查询请求,它可以是关键字查

询、简单的问答式查询、一个用户可以显式地选择本体的类及其属性值的格式化查询或者更为复杂的查询。这些查询条件可能不能直接被信息检索模块使用,查询条件预处理模块的作用就是将用户的查询条件转换为符合规范的结构化查询。由于知识库是采用 RDF 编码,查询条件需要经过查询条件预处理模块转换成 RDF/XML 格式的查询表达式。

1.5 结果用户接口与结果处理模块

对检索结果的处理对于信息检索系统也是尤为重要的。经过信息检索模块检索得到的满足条件的实例可能非常多,试想这些实例杂乱无章地展现在用户面前,很容易使用户迷失在信息的海洋中。结果处理模块的作用就是使得这些结果以用户易于接受的方式送给结果用户接口。这需要一个优秀的排序算法,能够较为准确地反映出各检索结果对用户的重要性。

1.6 语义标注

本文设计的基于本体的信息检索系统是建立在 Web 上的信息均具有语义的基础上的,然而,至今为止语义 Web 仍然是一个展望。现在 Web 上的信息几乎都是非结构化的,没有任何标注, Web 信息和语义信息检索之间存在着鸿沟。因此从 HTML 文档中自动地提取信息并且自动地转换成语义表示是十分必要的,一旦完成转换,数据可以被应用程序使用、存储到数据库或者通用本体中。然而,由于目前在世界范围内存在大量可用信息,这些信息都是以非结构化文本和媒体文档的形式存在的。在一个可以承受的花销范围内将这些巨量信息转换为规范的本体知识仍然是一个尚未解决的问题^[7]。然而,对特定领域的 Web 信息进行语义标注是可行的,在下面的实现中,将对地理信息领域的 Web 信息进行语义标注。

2 信息检索系统的实现

本节将以一个地理信息领域的应用为例讨论基于本体的信息检索系统的框架实现的具体问题。

2.1 知识库的建立

首先要建立起恰当的领域本体^[8],这是知识库建立的关键。本例中使用的是一个地理信息领域的本体,它基于德国的 ATKIS-OK-1000^[9] 目录系统。ATKIS 目录用于监控人类活动以及建筑物和其他设施对土地使用的影响,包括了对用作人类活动(如用作交通和运输的不同区域类型)的不同区域类型的详细区分,而对自然区域只作粗略区分。ATKIS 区域划分为居住区域、交通区域、植被区域和水面四个子区域,这四个区域又包含若干个子区域,图 2 展示了这个本体的类层次及其主要属性。本系统选用 Protégé 作为本体建模工具, Protégé 是由斯坦福大学开发的一个开放源码的本体编辑器,它是用 Java 编写的。用户无需考虑具体的本体描述语言,而只需在概念层次上设计领域本体模型。

现今的 Web 基本上都是没有语义的,对 Web 上浩如烟海的信息进行自动标注的问题现在尚未解决。本系统中的知识库采用一种十分简单的方法对 Web 上的资源进行标注,即在知识库中手动地添加指向 Web 上文档的 URI。当然,使用这种方法 Web 中能够包含的信息是十分有限的。

2.2 信息检索模块的实现

语义推理的目标是为了在信息检索时,能向用户返回全面

准确的检索结果,以满足用户的需求。何时进行语义推理在这里有两种选择:①信息检索模块针对用户提供的检索条件进行相关语义扩展^[10];②在知识库建立之后,对知识库进行语义推理,将所有隐含的信息均显式地描述出来,信息检索模块依据此结果执行检索。本文的实现采用后者,这样不仅降低了服务器的负载,也大大简化了信息检索模块的实现。但是,这是以牺牲检索的准确率为代价的。本模块的实现主要利用了惠普实验室开发的 Jena。

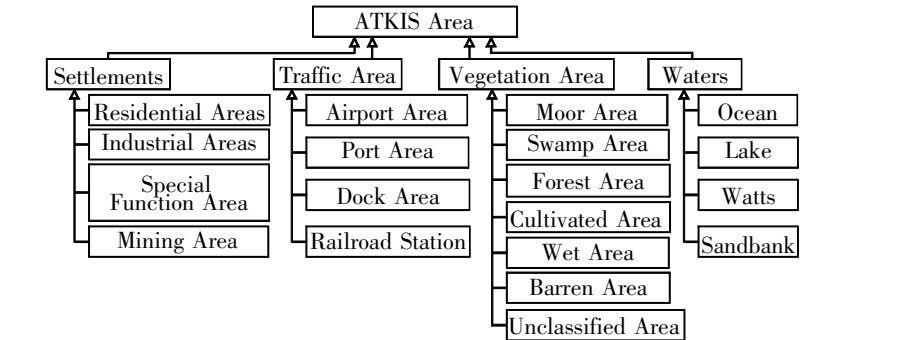


图 2 ATKIS 本体的类层次及其属性

信息检索模块首先接收结构化检索条件,然后读取知识库中的相关信息,信息检索模块依据知识库执行查询。知识库以 RDF 三元组的格式保存信息,即 (Subject, Predicate, Object) 格式。检索条件包括三个参数,信息检索模块将这三个参数与知识库中 RDF 三元组进行匹配,它返回与这三个参数匹配的所有三元组。如果在这三个参数的位置上任意一个为 Null,则认为该参数是匹配的。所以(Null, Null, Null) 返回知识库中的所有 RDF 三元组。这种信息检索方法从基于关键字的语法匹配提升至基于知识层面的语义匹配,用户可以找到包含句法不同但语义相似的词语的页面。在本文设计的信息检索系统中,对于某一个概念,用户可以根据需要指定其属性。在这里可以选择的属性有 SubClassOf(表示前者是后者的子类)、Type(表示前者是后者的实例)、Comment(对概念的描述)、InverseOf(相反关系)、DisjointWith(不相容关系)、URI (Web 资源所在地址) 等。例如,在本系统中输入概念 Residential-area(参见图 2 中的本体) 及其属性 SubClassOf, 我们将得到如图 3 所示的结果。



图 3 语义检索

在这个例子中反映了检索结果和检索条件之间的层次关系,这在传统的检索方式中是无法做到的,因为这种检索方式是基于语义层面而不是基于传统的语法层面。由此可见,与传统检索方式相比,语义检索方式具有一定的自主性和智能性,获得了对知识的表示、理解和处理等能力。机器可以对用户的查询条件和 Web 上的资源进行分析和处理,将用户最感兴趣

的信息返回,提高了信息检索的查全率和查准率。

此外,本文旨在利用本体进行语义层次上的信息检索,因此实现过程中不对查询条件预处理模块和结果处理模块两部分作进一步的介绍。用户直接将查询条件提交给信息查询模块,得到的结果直接返回给用户。

3 结束语

本文分析了传统的信息检索方式的不足之处,提出了一个基于本体的智能信息搜索系统的框架,解决了传统信息检索中查全率和查准率低下的问题;此外,本文还利用一个地理信息领域应用分析了智能信息搜索系统的实现技术。一旦建立起特定领域的本体,利用本文设计的框架就可以实现针对不同领域的智能信息检索。当然本文所给出的实现还不够完善,将其付诸于应用还有待于语义 Web 技术的进一步发展和对本文所提出的思想的进一步完善。

参考文献:

[1] im Berners-lee, J Hendler, O Lissala. The Semantic Web[J] . Si-
mantific American, 2001, 284(5) : 34-43.
[2] Eric Miller, Ralph Swick, Dan Brickley. Resource Description
Framework (RDF) [EB/OL] . http: //www. w3. org/RDF/, 2005-
10-04.
[3] I Horrocks, D Fensel, J Broekstra, et al. The Ontology Inference
Layer OIL[EB/OL] . http: //www. ontoknowledge. org/oil/TR/oil.
long. html, 2005-10-09.
[4] Michael K Smith, Chris Welty, et al. OWL Web Ontology Language
Guide[EB/OL] . http: //www. w3. org/TR/owl-guide/, 2004.
[5] Mike Dean, Guus Schreiber. OWL Web Ontology Language Refer-
ence[EB/OL] . http: //www. w3. org/TR/owl-ref/, 2004.
[6] Peter F Patel-Schneider, Pat Hayes, Ian Horrocks. OWL Web Onto-
logy Language Semantics and Abstract Syntax[EB/OL] . http: //
www. w3. org/TR/owl-semantics/, 2004.
[7] H Stuckenschmidt, Frank van Harmelen, Dieter Fensel, et al. Cata-
logue Integration: A Case Study in Ontology-based Semantic Transla-
tion[R] . Amsterdam: Computer Science Department, Vrije Univer-
siteit Amsterdam, 2000.
[8] Stefan Decker, Dieter Fensel, Frank van Harmelen, et al. Knowledge
Representation on the Web[C] . Aachen: Proceedings of the 2000 In-
ternational Workshop on Description Logics, 2000. 453-469.
[9] David Vallet, Miriam Fernández, Pablo Castells. An Ontology-based
Information Retrieval Model[C] . Heraklion, Crete: Proceedings of
ESWC, Springer, Lecture Notes in Computer Science (3532) , 2005.
455-470.
[10] Urvi Shah, Tim Finin, Anupam Joshi. Information Retrieval on the
Semantic Web[C] . Proceedings of the ACM Conference on Informa-
tion and Knowledge Management, 2002.

作者简介:

付燕宁(1965-), 男, 吉林四平人, 副教授, 博士研究生, 主要研究方向
为工作流、本体语言及语义 Web 服务; 金龙飞(1977-), 男, 吉林白山
人, 博士研究生, 主要研究方向为本体工程及语义 Web; 王开锋(1983-),
男, 山东兖州人, 硕士研究生, 主要研究方向为计算机应用; 刘磊
(1960-), 男, 吉林长春人, 教授, 博导, 主要研究方向为程序分析、本体
工程、语义 Web 等。