《智能信息处理》课程考试

**研究论文题目研究论文题目研究论文题目**

蒋志新

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 考核 | 到课[10] | 作业[20] | 考试[70] | 课程成绩[100] |
| 得分 |  |  |  |  |

2016年11月11日

基于本体的搜索引擎研究

蒋志新（大连海事大学大连116026）

**摘要**由于目前搜索引擎的查全率和准确率太低,提出了一种基于本体的搜索引擎概念模型;这种搜索引擎不仅仅基于关键词检索,而且机器能够理解Web页面的内容,并进行逻辑推理来完成复杂的查询任务,最终返回精确的结果。

**关键词**本体;元数据;搜索引擎

**TheResearchofOntology-basedSearchEngine**

**JiangZhixin**

**(DalianMaritimeUniversity,Dalian116026,China)**

AbstractBecausetheRecallandPrecisionofthepresentsearchengineislow,aconceptualmodelforontology-basedsearchengineispresented;itwillnotbebasedmerelyonakeywordindex,itcanalsounderstandthemeaningofcontentonthewebpagesandcarryoutlogicalreasoningonthemtoperformcomplexsearchqueries,thenreturntheaccurateresults.

Keywords:Ontology;metadata;SearchEngine

1引言

如今,Web页面每天爆炸式的增长,用户想找到需要的信息已经变得很困难。目前Web信息检索方法主要是基于内容分类目录和基于关键词搜索的。目录分类常见于一些门户网站(如Yahoo!等),主要是通过相关链接获得一些浅层信息。基于关键词搜索是把用户的查询请求和Web页面、文档中的每一个词进行比较,只要发现某个网页中含有这个关键字符,就将该网页作为查询结果返回给用户。因此,目前信息检索的查全率(Recall,也称召回率,即被找到的信息/全部所需要的信息)和准确率(Precision,即有用的信息/全部查询结果)难以令人满意。究其主要原因,是因为对计算机而言,关键词几乎没有任何语义,计算机不理解Web页面中词语表达的语义及其相互关系,因此检索的性能还是难以得到本质的提高。本体技术的发展使上面的问题变得很容易解决;通过使用本体,能使机器理解包含语义的文档和数据,从而实现“精细、准确和自动化”的搜索。

2搜索引擎技术介绍

搜索引擎是一种能够通过Internet接受用户的查询指令,并向用户提供符合其查询要求的信息资源网址的系统。一个搜索引擎从逻辑上由三个功能模块:(1)信息搜索(2)信息处理(3)信息查询。信息搜索:也称为信息采集,功能是在Web中漫游,发现和搜索信息。一般是一个计算机后台程序,采用一定的搜索策略,不停地在WWW中搜集各种类型的新信息,同时具有定期更新功能,用于避免由Web上的信息更新带来的死链接和无效链接。信息处理:功能是将信息搜索模块搜集到的信息进行分类整理,采用不同的索引技术建立索引数据库,同时也要对索引数据库定时更新。信息查询:每个搜索引擎都必须向用户提供一个良好的信息查询界面,一般包括分类目录及关键词两种信息查询途径。分类目录查询是以资源结构为线索,将网上的信息资源按内容进行层次分类,使用户能依线性结构逐层逐类检索信息。关键词查询是利用建立的网络资源索引数据库向网上用户提供查询“引擎”。用户只要把想要查找的关键词或短语输入查询框中,并按“搜索”按钮,搜索引擎就会根据输入的提问,在索引数据库中查找相应的词语,并进行必要的逻辑运算,最后给出查询的命中结果(均为超文本链接形式)。用户只要通过搜索引擎提供的链接,就可以立刻访问到相关信息。

3本体技术

本体(Ontology)的概念起源于哲学领域,即"对世界上客观存在物的系统地描述"。在人工智能界,最早给出本体定义的是Neches等人,他们将本体定义为"给出构成相关领域词汇的基本术语和关系,以及利用这些术语和关系构成的规定这些词汇外延的规则的定义"。在计算机界最著名并被引用得最为广泛的定义是由Gruber提出的:本体是概念模型的明确的规范说明。本体的实质是把本体当作是领域(特定领域,或更广的范围)内部的不同主体(人、机器、软件系统等)之间进行交流(对话、互操作、共享等)的一种语义基础,即由本体提供一种明确定义的共识。

本体的目标是获取、描述和表示相关领域的知识,提供对该领域知识的共同理解,确定该领域内共同认可的词汇,并从不同层次的形式化模式上给出这些词汇和词汇间相互关系的明确定义。目前大家公认在构造特定领域的本体的过程中需要领域专家的参与。而一般最典型的本体应具有一个分类系统和一系列推理规则,分类系统定义对象的类别和类目之间的关系;借助推理规则,可以提供更强的推理能力。本体的类型有如下几种:(1)顶级本体:描述的是最普通的概念及概念之间的关系,如空间、时间、事件、行为等等,与具体的应用无关。(2)领域本体:描述的是特定领域中的概念及概念之间的关系。(3)任务本体:描述的是特定任务或行为中的概念及概念之间的关系。(4)应用本体:描述的是依赖于特定领域和任务的概念及概念之间的关系。

4基于本体的搜索引擎概念模型

结合本体技术,我们提出了一种基于本体的搜索引擎概念模型,如图1所示

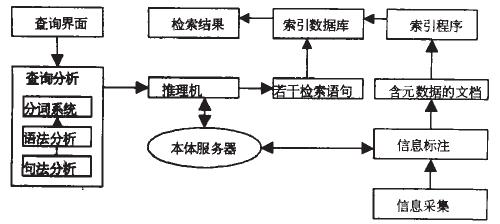


图1基于本体的搜索引擎概念模型

整个模型,从搜索引擎的功能角度来划分,我们分成五大部分:查询预处理、推理机、本体服务器、信息收集与处理、查询与显示。

4.1查询预处理

查询语句输入:用户首先进入搜索引擎的Web页面即信息查询界面,输入查询语句,设置查询条件,其中查询条件的设置主要用于解决一词多义的问题,即当用户只输入一个多义词时,还需要在查询设置条件中设置词语的相关领域,例如用户输入一个“线程",然后在查询条件中设置领域为计算机,这样在下一步的查询分析中计算机就可以分析出用户是想查找在计算机领域中“线程”这个词的信息。当然我们可以直接输入计算机线程”而不需要设置查询领域。查询分析:主要完成对查询语句的处理,其中包括对查询输入与领域设置的信息融合,以及使用常用的分词系统、语法分析、句法分析功能模块,经过这些功能模块的一一处理,最后得到若干个精确的查询语句,这些查询语句应该是推理机可以理解的形式化的语法格式,比如用OWL-QL(OWLQuerylanguage)表示的查询语句。

4.2 推理机

推理机是基于本体的搜索引擎的核心模块之一,也是区别于现今基于关键词检索搜索引擎的关键模 块,是由一组具有推理策略的程序组成的推理系统。常见的推理机有面向低层次的、通用的Jess和针对具体本体语言的推理机,如Racer、FaCT、KAON2、Pellet和 Jena 推理机等。 Jess(JavaExpertShellSystem) 是基于 Java 语言的 CLISP(commonLISP) 推理机 ,Jess(CLISP) 推理机是开 放的 , 用户提供不同的规则系统 , 就可以进行不同领 域的推理工作 , 用户可以对推理机的推理能力进行扩 展。但作为前向推理系统,Jess 用空间换时间 , 推理会 产生大量的中间数据 , 空间效率很低 ; 同时 , 由于 Jess (CLISP) 是通用推理引擎 , 不可能提供针对各种具体领 域的优化能力 , 使得这种推理机制的效率很难优化。 Racer 、 FaCT 、 KAON2 和 Pellet 等是针对本体的推 理。它们具有效率高, 使用方便的优点 ; 但不具有通用 性 , 其推理能力限定在几种具体的本体语言 ( 如 OWL 、 DAML 、 RDFS/RDF) 上 , 而且用户很难对它进行扩展 , 这些推理机采用描述逻辑作为理论基础 , 算法采用 Tableau 算法。而 Jena 是面向语义 Web 的应用开发 包 , 包含的内容比较全面 , 推理机只是其中一部分 ;Je- na 提供的推理机也是针对本体的推理 , 本质上是一种 CLISP 配本体领域产生式规则的前向推理系统 , 因此 , 运行效率不是很高。好在现在有 DIG 接口 ,DIG 有点 像数据库中的 ODBC, 允许前端挂接到后台不同的推 理引擎上。这样, 在 Jena 中 , 也可以使用 RACER 、 FaCT 、 KAON2 和 Pellet 这些更‘专业’些的推理机。 在本模型中 , 推理机就是针对本体的推理 , 首先 接受上一层的查询语句 , 这些查询语句要经过一个对 其中概念的提取过程 , 由于本体是由相关领域的概念 及概念间的关系组成 , 因此推理操作需要一系列原始 概念 , 然后根据推理规则结合本体服务器中的相关本 体 , 完成复杂的推理任务 , 最后输出精确的检索语句。

4.3 本体服务器

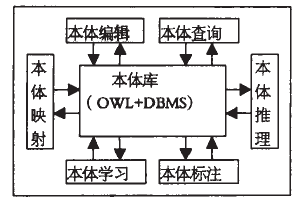
本体服务器模块是整个模型的核心部分 , 如图 2 所示。本体服务器从内部功能的角度分为三层: 本体 访问层、本体表示层和信息集成层。本体访问层负责 处理本体的使用问题 , 包括处理来自外部的本体检 索、编辑和查询等请求, 这种请求一般通过本体查询 语言表示出来 ; 模块中的本体查询将在本体表示层的 支持下解释执行本体查询语句 ; 本体表示层负责本体 的存储问题 , 包括提供集中的索引和查询辅助机制 , 本 体库中的本体语句使用标准的本体语言 ( 如 OWL) 存 储。信息集成层负责本体的获取问题, 其本体学习模块 可以通过处理来自异构数据源的各种数据形成系统内 部的本体 , 本体映射模块可以将系统以外的各种格式 的本体库中的信息转换成系统内部使用的格式 , 本体 标注模块主要负责提供一个进行语义标注的接

图2 本体服务器模块

为了重用本体服务器之外已经建立起来的本体 库 , 服务器还需要与一些已有本体库系统进行交互 , 这是通过本体访问层的本体编辑器和信息集成层的 本体映射模块来实现的 ; 它们实现了不同格式本体之 间的转换与 ( 不完全 ) 映射。在整个模块体系中, 信息 集成层为本体表示层提供信息集成的结果 ( 如 OWL 形式的本体 ) ; 本体表示层实现统一的本体存储服务 , 同时还在本体推理机的支持下进行本体的一致性检 查等必要的本体维护工作 ; 本体表示层向本体访问层 提供数据源支持 , 通过本体访问层提供的接口实现对 其访问的支持。

4.4 信息收集与处理

本模块主要完成对 Web 页面的收集、标注和索 引 , 并最终得到元数据索引数据库。首先在信息采集 模块中 , 利用一定的搜索策略进行原始数据的采集 , 例如可以使用一些抓取程序 ( Spider 、 Robot 等 ) 来获取 最原始的 Web 页面 , 然后通过本体标注模块结合本体 服务器的本体标注接口来对原始的 Web 页面完成标 注 , 标注后的页面可以为 XML/RDF 或 XML/OWL 形 式 , 索引程序采用某种索引算法对标注过的包含元数 据的文档进行索引 , 从而建立索引数据库 , 其中索引 算法的核心应该以本体中的概念实例对应的元数据 为索引关键词。

4.5 结果显示

主要完成对搜索结果的查询与显示。在索引数据库 中 , 使用推理机推理出来的精确的查询语句进行查询 , 得到一系列精确的查询结果 ; 然后按照一定的排序算 法 , 对结果进行排序 , 最后把结果显示在 Web 页面上。

5 模型的具体实现

整个模型在技术上的具体实现 , 可以采用基于 Ja- va 语言 , 并结合利用 Jena API 类库来编程完成 ; 本体 的构建目前有很多的本体编辑工具 , 像 Pretege2000, MR3 等 ; 本体描述语言可以采用 OWL DL; 对于推理机 最原始的方法可以使用 Jena API 进行编写 , 也可以采 用 Racer 、 KAON2 、 pellet 等功能比较强大的推理工具 ; 而对于本体的存储应使用关系型数据库。

6 小结

本文提出了一种基于本体的搜索引擎概念模型 , 用以解决目前基于关键词检索的搜索引擎的查全率 和准确率太低的问题 ; 由于多个本体之间的映射及本 体标注工具的研究等问题还没有很好的解决 , 所以基 于本体的搜索引擎真正应用还需时日 , 但是这种搜索 引擎的强大功能和其美好的前景已经被认可 , 以后 , 在这方面的研究定会越来越多。 创新点 : 提出了一种新的基于本体的搜索引擎概念模型。

参考文献

[1]Neches R, Fikes R E, Gruber T R, et al. Enabling Technology for Knowledge Sharing[J]. AI Magazine, 1991,12(3):36- 56.

[2] 娄雅斌 , 陶风梅 , 马垣 . 基于“本体”的异构数据源的集成方法研 究 [J]. 微计算机信息 ,2005,10- 3.

[3]D.L.Mc Guinness,R.Fikes,J.Hendler,and L.A.Stein, "DAML+OIL: An Ontology Languange for the Semantic Web"[J],IEEE Intelligent Systems,2000,pp.72- 80.

[4]D.L.Mc Guinness,and F.V.Harmelen,"OWL Web Ontology Lan- guange Overview,W3C Proposed Recommendation"[EB],http://www. w3.org/TR/2003/PR- owl- features- 20031212/,2003.

[5]T.B.Lee,and D.Connolly,"Closed World Machine"[EB/OL], http:// [www.infomesh.net/2001/cwm/](http://www.infomesh.net/2001/cwm/).

[6]T.R.Lee,J.Hendler and O.Lassila,"The Semantic Web " [J],Sci- entific American,2001.

[7]D.Beckett,"RDF/XML Syntax Specification (Revised)" [OL],W3C Recommendation,2004.