一个完整的基于语义网的信息搜索模型

董海凤
(上海海事大学,上海 200135)

摘 要:提出了一个完整的语义网信息搜索模型,它可以实现更快更准的网络信息搜索,以减少当前网络搜索的弊端。介绍了语义网领域发展背景和语义网的三个关键技术,论述了该模型的三个主要模块及其各自功能,详细说明了各功能的实现原理。比较传统的直接使用关键字来进行信息检索的主流网络信息检索引擎,这里使用的是在本体论基础上经过语义逻辑推理器处理后的语义索引词作为查询的输入,使查询变得准确快速而且节省网络资源。

关键词:语义网;搜索模型;本体论
中图分类号:TP311;TP393 **文献标识码**:A **文章编号**:1673-629X(2009)08-0001-03

An Integrated Information Retrieval Model Based on Semantic Web

DONG Hai-feng
(Shanghai Maritime University, Shanghai 200135, China)

Abstract: In this paper, a model of information retrieval based on semantic web is designed to realize better information retrieval and reduce the disadvantages of the current search on web. Introduces the development background of the semantic web and its three most important technologies. Also it discusses both module and their functions in detail. Compare with the traditional directly use keywords to retrieval information, here it uses the semantic indexing that dealing by semantic logic device which based on ontology as query input to make inquiries quickly and accurately.

Key words: semantic web; retrieval model; ontology

0 引 言

随着 Internet 的迅速发展,网上信息量以指数规律迅速增长。信息量增多了,但用户获取其需求信息却没有变的容易起来。现有互联网技术的局限性以及检索系统对用户提问理解不够准确等原因使得用户越来越难以在网上找到能满足需求的信息。目前万维网存在两个明显的不足:计算机不能理解网页内容的语义;网上有用信息难找,查准率也比较低,夹杂了许多用户不需要的信息垃圾[1]。存在这些问题的原因在于万维网现在采用的超文本标记语言 HTML,网页上的内容设计成专供人类浏览的,而非供计算机理解和处理的。针对这个问题。互联网的创始人 Tim Bemers-Lee 于 2000 年 12 月 18 日在 XML2000 的会议上正式提出语义网[2]。

收稿日期:2008-11-23;**修回日期**:2009-02-11
基金项目:上海市教育科研基金(教 05-31)
作者简介:董海凤(1984-),女,山东人,硕士研究生,研究方向为多媒体应用与技术;导师:张 明,教授,研究方向为多媒体信息检索、多媒体数据库、基于内容的图像检索与知识发现、信息管理与信息系统、信息隐藏与信息安全。

语义网的目标是使得 Web 上的信息具有计算机可以理解的语义,满足智能软件代理对万维网上异构和分布式信息的有效访问和搜索。

Tim Berners-Lee 对语义万维网做了如下描述:语义万维网是对当前万维网的扩展,语义万维网上的信息具有定义良好的含义,使得计算机之间以及人类能够更好地彼此合作[3]。由此,相对于传统的信息检索,基于语义网的信息检索有更高的准确性,对于网络信息检索来说具有非常重大的意义。

1 语义网的关键技术

语义网的实现需要三大关键技术的支持:XML、RDF 和 Ontology。XML(eXtensible Marked Language,即可扩展标记语言)可以让信息提供者根据需要,自行定义标记及属性名,从而使 XML 文件的结构可以复杂到任意程度。它具有良好的数据存储格式和可扩展性、高度结构化以及便于网络传输等优点,再加上其特有的 NS 机制及 XML Schema 所支持的多种数据类型与校验机制,使其成为语义网的关键技术之一。目前关于语义网关键技术的讨论主要集中在 RDF 和 On-

tology身上。

RDF是W3C组织推荐使用的用来描述资源及其之间关系的语言规范,具有简单、易扩展、开放性、易交换和易综合等特点。值得注意的是,RDF只定义了资源的描述方式,却没有定义用哪些数据描述资源。RDF由三个部分组成:RDF Data Model、RDF Schema和RDF Syntax。RDF Data Model提供了一个简单但功能强大的模型,通过资源、属性及其相应值来描述特定资源。模型定义为:

(1)它包含一系列的节点N;

(2)它包含一系列属性类P;

(3)每一属性都有一定的取值V;

(4)模型是一个三元组:{节点,属性类,节点或原始值V};

(5)每一个Data Model可以看成是由节点和弧构成的有向图。

RDF(资源描述框架),提供了用于描述元数据,特别是Web元数据的技术。它是通过三元组的方式来描述资源,包括"主语","谓词","宾语"(通过资源、属性及其相应值来描述特定资源)。描述三元组有很多方式,其中最重要的是以XML方式。

Ontology(本体或本体论),原本是一个哲学上的概念,用于研究客观世界本质。目前Ontology已经被广泛应用到包括计算机科学、电子工程、远程教育、电子商务、智能检索、数据挖掘等在内的诸多领域。它是一份正式定义名词之间关系的文档或文件。一般Web上的Ontology包括分类和一套推理规则。分类,用于定义对象的类别及其之间的关系;推理规则,则提供进一步的功能,完成语义网的关键目标即"机器可理解"。本体的最终目标是"精确地表示那些隐含(或不明确的)信息"。

当前对本体的理解仍没有形成统一的定义,如本体是共享概念模型的形式化规范说明,通过概念之间的关系来描述概念的语义;本体是对概念化对象的明确表示和描述;本体是关于领域的显式的、形式化的共享概念化规范等等。但斯坦福大学的Gruber给出的定义得到了许多同行的认可,即"本体是概念化的显示规范"。概念化(Conceptualization)被定义为:C=<D,W,R>,其中C表示概念化对象,D表示一个域,W是该领域中相关事物状态的集合,Rc是域空间上的概念关系的集合。规范(Specification)是为了形成对领域内概念、知识及概念间关系的统一的认识与理解,以利于共享与重用。

在语义网中,本体具有非常重要的地位,它是解决语义层次上万维网信息共享和交换的基础。本体是一整套对某一领域里的知识进行表述的词和术语,编制者根据该知识领域的结构将这些词和术语组成等级类目,同时规定类目的特性及其之间的关系。其与叙词表的区别在于叙词表并不能表达概念之间丰富的联系。但本体可用相应的描述语言描述概念之间的关系,为实现推理功能建立基础[4]。

一个领域本体模型给出了某领域的准确描述,通常领域本体是由该领域的专家共同制定,把它抽象为概念、概念属性及概念间关系的集合。

2 基于语义网的信息搜索模型的框架

这里基于语义网的信息搜索模型由三个模块组成,分别是:用户输入信息预处理;源信息收集、处理和索引库建立;搜索与匹配。框架图如图1所示。

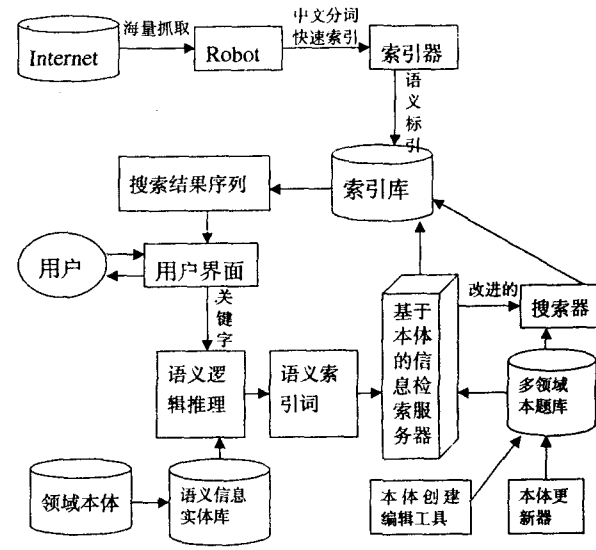

图1 基于语义网的信息检索模型

其中用户输入信息预处理负责把用户输入的关键字进行处理,生成语义索引词作为查询的输入信息;源信息收集、处理和索引库建立是指从海量的网络资源中找出相关网站通过中文分词、快速索引等技术建立索引库;搜索与匹配模块则是利用现有的语义网技术通过改进的搜索引擎寻找符合用户要求的资源。

3 模型功能的实现过程分析

3.1 用户输入信息预处理

传统网络搜索是直接使用用户输入的关键字来进行信息检索,传统的关键字检索效果不尽人意,主要原因在于用户真正的检索意图很难用几个关键字表达清楚,这也是导致现有检索系统精度不高的原因之一。解决方法是用户以自然语言的方式向系统提问,系统利用领域本体中的知识对用户的问题进行语义分析,得到用户真正的检索意图,然后再将检索请求提交给

系统的检索部分。

这里在领域本体基础上建立语义信息实体库,然后再把关键字经过语义逻辑推理器的分析找出语义索引词代替用户输入的关键字作为搜索的输入。通过语义推理,可以找出文档中隐含的语义关系,提高信息检索中的查全率和查准率。其中语义索引词被认为是用户的真正搜索意图。这样用户只需提交一次查询就可以通过不同的语义关系进行检索,获得不同的相关文档序列,不需要反复的检索和检索扩展,提高了网络资源的利用率,也可以避免因为用户描述不当而带来的搜索误差,使检索变得快速而且准确。

3.2 源信息收集、处理和索引库建立

人们目前所使用的主流搜索引擎主要由三部分构成:Robot、索引器和搜索器。搜索引擎的基本工作原理为:首先使用 Robot(也称之为 spide,一个用来采集文档的程序)来遍历网络,根据网页之间的超链接将网络上的信息资源下载到本地,然后对文档内容进行自动分析并建立索引,接着将索引后的文档替代物存放在索引库中,最后由搜索器对用户查询式进行分析处理,然后从索引库中搜索出相应的检索结果返回到客户端。

基于语义网的智能搜索引擎区别主要体现在建立索引和检索过程中,所有工作都建立在语义网络基础上。智能搜索引擎需要具备符合用户信息需求的本体知识库,在用户界面接受到查询式进行搜索时,搜索器根据已有的本体库,分析查询式,了解检索词的意义,并运用一定的描述语言,根据本体库里概念间的联系产生联想和推理,从而找全找准与用户需求最相关的文档。因此,在信息标引过程中,适当的语义知识表示是建立本体库系统的关键,只有在标引过程中深入挖掘文档的语义内容,而不仅仅局限于形式的关键词,才能提高检索效率。对 Robot 从 Internet 采集进来的文档,可以利用索引器对文档进行语义标引。文中利用这样的语义标引方法:先从文档集中抽取出特征词汇,分析特征词汇,并建立与概念集之间的联系,从而达到使用领域本体对文档进行语义标引[5],然后用这些相关信息建立网页索引数据库,以备搜索器查询。

由于本体是共享的、形式化的、抽象的概念集合,概念及概念之间关系都已被精确地描述,因此,通过这种语义标引方法,可以表示出文档中隐含的语义信息,文档的所属类别也被很好地划分,并且与其他概念之间有了明确的语义关系。这些都是实现语义推理功能和语义查询扩展的基础。

3.3 搜索与匹配

关键词经过处理,索引库也建立,最后一步就是由搜索器进行搜索匹配用户需要的信息了。当用户输入关键词转换成语义索引词之后提交给基于本体的信息搜索服务器,通过改进的搜索器分解搜索请求,由搜索系统程序从网页索引数据库中找到符合该语义词的所有相关网页。针对该语义索引词的所有相关网页的相关信息在索引库中都有记录,只需综合相关信息和网页级别形成相关度数值,然后进行排序。相关度越高,排名越靠前。最后由页面生成系统将搜索结果的链接地址和页面内容摘要等内容组织起来返回给用户。

这一步的关键技术在于本体的应用。当用户提交一个查询后,由浏览器交给了远端的基于本体论的信息检索服务器。远端的基于本体论的信息检索服务器通过查询本体论,得到这个关键字的信息,以及这个术语在不同领域的含义等等。如果不是一个术语或者说不是一个概念,这只好交给改进的搜索引擎检索,按传统的搜索引擎方法对它进行检索。如果是一个术语或者说是一个概念,则在本体论中(可能是很多领域的本体论集)有它的入口。在本体论中得到术语的信息,如:属于某个领域集合及该领域集的定义、用法示例、相关的主题、同义词,如果本体论支持多语言,还有其它语言的同义词等[6]。

其中多领域本体是本体的集合。Neches 等人将 Ontology 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"。Neches 认为:"本体定义了组成主题领域的词汇表的基本术语及其关系,以及结合这些术语和关系来定义词汇表外延的规则。"其中最著名定义是由 Gruber 提出的,"本体是概念化的明确的规范说明"[6,7]。

改进的搜索引擎:采用全文检索技术。全文检索技术处理的对象是文本,它能够对大量文档建立由字(词)到文档的倒排索引。改进的搜索引擎加上了由关键字到领域的领域索引表,支持领域分类。

本体更新器:Nicala Guarino 认为应该按照层次关系,建立不同的本体论,在建立了顶层本体论之后,就可以着手建立领域本体论了。本体论是世界的反映,因此它必然随着现实的发展而变化[6]。

4 结束语

文中提出了一种基于语义网的信息检索模型,给出了它的总体框架、基本功能和实现方法。该模型有以下新特点:

(1)功能完整,每个模块都预先处理,是一个比较完善的基于语义网的信息搜索模型。

(下转第 7 页)

定很多。

②考察最底层容器标签节点。

大多数网页一般采用 table 或 div 容器标签来布局,不失一般性以 table 标签为例说明确定正文信息块的算法,div 等其他处理方法类似。

一般正文及有用的信息多数情况下会出现在容器标签嵌套的最里层,因此先从最里层的 table 节点考察起,并对 DOM 标签树采用自底向上的顺序进行处理。即:把标签树中各个最底层的 table 节点下的所有文本节点(TextNode)合并,并统计该块的信息含量。

③判断一个节点能否成为分块节点。这是分块算法的关键步骤。

分块时,一定要尽量往父节点上靠,从而得到逻辑上的最大信息块。因为同一块内的节点标签相似度强,所以在考察时,确定当前节点的块类型,分析兄弟节点的块类型是否一致,若基本一致,则当前节点上移至父节点,继续按上述方法分析,直到发现兄弟节点类型不同时结束。此时的节点就是一个最大信息块。

4　结束语

Web信息抽取的前期研究主要集中在生成和使用包装器阶段,但该方法的缺陷是缺少灵活性和可扩展性;后期研究则基于 HTML 结构,将 Web 文档解析成语法树并对其进行分析[9],通过自动或半自动的方式产生抽取规则,然后进行信息抽取。文中首先对Web信息抽取的现有抽取技术以及现有抽取技术存在的问题进行了详细的论述。然后根据 Web 信息抽取的原理,依据软件工程的原理对 Web 信息抽取技术提出了具有指导意义的 8 条启发式规则,并在这些规则的指导下,着重阐述了 Web 信息抽取中的基于结构和内容的信息抽取。

参考文献:

[1] 陈　静,朱巧明,贡正仙.基于 Ontology 的信息抽取研究综述[J].计算机技术与发展,2007,17(10):84-86.

[2] Califf M, Mooney R. Relational Learning of Pattern-Match Rules for Information Extraction[C]//In Proceedings of the Sixteenth National Conference on Artificial Intelligence. Orlando, Florida: [s. n.], 1999.

[3] Freitag D. Machine Learning for Information Extraction in Informal Domains[D]. [s. l.]: Carnegie Mellon University, 1998.

[4] Soderland S. Learning Information Extraction Rules for Semi-structured and Free Text[J]. Machine Learning, 1999, 34(1-3):233-272.

[5] 李效东,顾毓清.基于 DOM 的 Web 信息提取[J].计算机学报,2002,25(5):526-533.

[6] Deng Cai, Yu Shipeng, Wen Jirong, et al. VIPS: a Vision-based Page Segmentation Algorithm[R]. Microsoft Technical Report, 2003.

[7] 常育红,姜　哲.基于标记树表示方法的页面结构分析[J].计算机工程与应用,2004,40(16):129-132.

[8] Kovacevic M. Recognition of Common Areas In Web Page Using Visual Information: A possible application in a page classification[C]//Proceedings of ICDM02. Maebashi, Japan: IEEE Press, 2002.

[9] 仲　华,崔志明.基于 XML 的信息抽取和多层向量空间技术研究[J].计算机技术与发展,2007,17(7):49-52.

(上接第3页)

(2)比较传统的直接使用关键字来进行信息检索的主流网络信息检索引擎,这里使用的是在本体论基础上经过语义逻辑推理器处理后的语义索引词作为查询的输入,使查询变得准确快速而且节省网络资源。

(3)在索引库的建立过程使用索引器对结果进行语义标引,形成索引库,因为适当的语义知识表示是建立本体库系统的关键,所以良好的语义标引为搜索时利用本体领域奠定基础。

(4) 基于本体的信息检索服务器是功能的核心,智能搜索引擎需要具备符合用户信息需求的本体知识库,在用户界面接受到查询式进行搜索时,搜索器根据已有的本体库,分析查询式,了解检索词的意义,并运用一定的描述语言,根据本体库里概念间的联系产生联想和推理,从而找全找准与用户需求最相关的文档。

除此之外,这个模型中还有一个改进的搜索器专门处理全文检索。

参考文献:

[1] 孙宝传.互联网未来的发展走向综述(I)[J].中国传媒科技,2003(3):61-62.

[2] Berners-Lee T, Handler J, Lassila O. The Semantic Web[J]. The Scientific American, 2001, 284(5):34-43.

[3] 朱礼军,陶　兰,黄　赤.语义万维网的概念、方法及应用[J].计算机工程与应用,2004(3):79-80.

[4] 李选如,何洁月.语义集成:本体映射方法研究[J].计算机技术与发展,2007,17(2):127-130.

[5] 何绍华,宫兆晖.基于语义网的网络信息检索相关性研究[J].情报杂志,2007(2):120-121.

[6] 王文峰,赵　莉.语义网中的本体对象研究及应用[J].枣庄学院学报,2007(2):52-53.

[7] 杨晓青.一种利用 RSF(S)建立本体论的方法[J].计算机应用研究,2002,19(4):11-15.