基于语义的Web信息检索

胡必云,黄因生,谢荣传
(安徽大学 计算机科学与技术学院,安徽 合肥 230039)

摘 要:用户要从网络中得到所需的信息一般是通过各种搜索引擎。但是现有的搜索引擎都存在着检索相关度不高等问题。随着语义Web概念的提出及相关技术的发展,基于语义的Web信息检索逐渐成为了语义Web研究的热点。给出了传统搜索引擎存在的问题,从理论上分析了如何将语义Web技术融入Web信息检索中去,并在理论分析的基础上给出了基于语义的Web信息检索的模型。

关键词:语义Web;Web信息检索;本体
中图分类号:TP301.2 **文献标识码**:A **文章编号**:1673-629X(2006)10-0071-03

Semantics-Based Web Information Retrieval

HU Bi-yun, HUANG Yin-sheng, XIE Rong-chuan
(College of Computer Science and Technology, Anhui University, Hefei 230039, China)

Abstract: People get information from Web mainly by search engines, but always puzzled by the low relevance of them. Along with the semantic Web's proposition and the development of related technology, semantic-based Web information retrieval has gradually become the hot topic in semantic Web study. The paper analyses the problems of traditional search engines, and analyses how semantic Web technology can be used together with Web information retrieval. At last, give a semantic-based Web information retrieval model.

Key words: semantic Web; Web information retrieval; ontology

0 引 言

使用传统的搜索引擎,用户要精确地找到所需信息往往十分困难。提高搜索引擎准确度的一个主要途径是使其在某种程度上理解用户检索或信息源的内容。目前在互联网研究领域兴起的语义Web技术就是针对此应运而生的。

1 传统搜索引擎存在的问题

科学证明,搜索引擎是未知状态下发现有效信息的最有效方式[1]。但传统搜索引擎技术都是基于关键字的语法匹配和全文检索技术,主要借助于目录、索引和关键词等方法来实现。此技术的优点是简单、快捷和容易实现,但其存在以下比较突出的问题[2]:

(1)"忠实表达"问题。由于在大多数情况下用户很难通过简单的几个关键词来忠实地表达其检索需要,因此表达困难也就导致了检索质量难尽人意;

(2)无法准确揭示信息的实质内容。用题名、文摘或全文中出现的关键词标识文档的内容,常常不能充分揭示Web文档的实质内涵。

收稿日期:2006-02-19

作者简介:胡必云(1982-),女,安徽六安人,硕士研究生,研究方向为Web与数据库技术;谢荣传,副教授,硕士生导师,主要研究方向为数据库、Internet应用、多媒体技术。

2 语义Web技术与Web信息检索

语义Web技术有助于解决以上提出的问题,下面从两个方面展开讨论。

2.1 使用本体对用户查询做查询扩展

本体在语义Web前景中处于很重要的地位,本体通过对给定领域概念的一致、形式化描述使得知识可以共享、重用,以及在不同的agents(人或机器)之间达成一致理解。本体可以看作是定义了类及类之间关系,同时添加了用于推理的规则集的分类体系[3]。本体可以通过对用户检索进行领域内的概念及属性关联来扩展用户检索。同义词典也可用来对用户检索词进行查询扩展,比如用户查询"计算机",与"电脑"相关的信息也能检索出来,但应该明确的是同义词典不提供推理规则,而推理规则对于本体则很重要。本体可以通过以下3种方式对用户查询进行扩展。

(1)关联和从属的概念。

这种方式通过找出与用户提供的关键词相关的概念(通过任何谓词)和从属的概念对用户查询进行扩展。例如用户如果想要检索关于"网络"的信息,同时本体中定义了"网络"的关联概念(例如"同义关联")"因特网",这时候用户的查询就扩展为"网络"和"因特网"。如果在本体的定义中还包含概念"网络"的子概念"个人网络",这时用户查询可以扩展为"网络"、"因特网"和"个人网络"。

(2)谓词和关联的概念。

这种方式通过找出与用户提供的关键字相关的谓词(在 RDF 中指属性)及概念对用户查询进行扩展。图 1 中概念"网络"分别通过谓词"有"及"由……组成"与概念"流量"及概念"计算机"关联,则用户输入的查询"网络"可以用来形成新的查询条件"网络有流量"及"网络由计算机组成"。

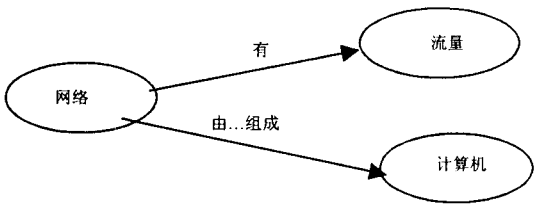

图 1 谓词关联扩展

(3)谓词特性。

这种方式通过传递、翻转或对称谓词来扩展用户查询。图 2 中概念"网络"通过传递属性"分发"与概念"数据"关联,同时概念"数据"通过属性"分发"与概念"信息"关联,虽然概念"网络"与"信息"并非直接关联,但考虑本体中传递属性的含义,通过"网络分发数据"及"数据分发信息"可以推断出"网络分发信息"。由此产生新的查询条件"网络分发数据"及"网络分发信息"。关联"网络"及"计算机"的谓词"由……组成"有翻转谓词"组成"。所以可以产生查询"计算机组成网络",它是查询"网络由计算机组成"的翻转。对称属性也可用于查询扩展。例如概念"网络"通过对称属性"使用"与概念"无线通信"关联,则可产生查询条件"网络使用无线通信"及"无线通信使用网络"。

图 2 谓词特性扩展

以上 3 种方式都是利用本体中的概念及属性来对用户查询进行扩展。扩展后的查询条件更能"忠实表达"用户的检索意图,从而提高检索结果的相关度。但在使用传递谓词及对称谓词对用户查询进行扩展的时候,有一点值得注意,即从本体中产生的概念的重要性大于用户输入的关键词,因为它们处于查询序列的前端。这对于搜索引擎分类文档相关度是需要考虑的一点。

2.2 带有语义标注的 Web 页面检索

当今 Web 页面的设计只是为了方便人的阅读,并不包含计算机能够处理的描述信息。这种设计思想限制了 Web 的用处,尤其是对那些搜索 Web 的用户来说。语义网络将 Web 网页加上表明文档语义的标签,而不仅仅是用来描述如何显示(HTML)或者语法结构(XML),从而使得计算机能够理解网页的内容,这意味着计算机能够推理网页上的数据是否满足用户的检索需求。

有两种方式可以用来对 Web 页面进行语义标注。一种方式是将语义标记直接嵌入到 HTML 页面中去,但是考虑用 DAML + OIL 或 OWL 来进行标记的时候会发现它们是用于知识表示的语言而不是用于直接嵌入到文本中去的。同时在 HTML 页面中嵌入基于 RDF 的标记与 HTML 标准不兼容,W3C 的一个工作组正在研究解决这一问题[4]。第二种方式是通过两个文件将 HTML 与语义标记进行绑定,其中一个文件包含 HTML,另一个文件包含对应的语义标记,在每个文件中分别放置用于指向另外一个文件的 URI 的指针。在检索时这两个文件将同时用于检索。虽然这种方式不易将语义标记与 HTML 的特定成分进行对应,但是用此法便于使用 HTML 标准。

当今的 Web 搜索技术不适合直接对语义标记进行索引及检索,大多数搜索引擎使用词来对文档进行索引。当对 HTML 文档进行索引的时候,嵌入在里面的标记将被大多数的搜索引擎简单地忽略掉,即使搜索引擎能够识别同时索引嵌入在 Web 文档中的标记,它也不能在搜索中有效利用语义标记的推理作用或者说以一种能区别标记与普通文本的方式来处理标记。

一种解决上述问题的方式是首先将用于文档标注的基于 RDF 的标记从 XML 名空间的简写形式转化成完整的形式,例如将三元组(http://www.cit.gu.edu.au/db, mydomain: site - owner, David Billington)转化成(http://www.cit.gu.edu.au/db, http://www.mydomain.org/site - owner, David Billington)。然后将 RDF 三元组的完整描述作为附加信息添加到 Web 文档中,允许搜索引擎对三元组的 3 个部分(subject, predicate, object)的任意组合进行索引(除去无意义的空组合)。例如对于上述三元组可以有以下 7 种索引方式:

http://www.cit.gu.edu.au/db, http://www.mydomain.org/site - owner, David Billington

http://www.cit.gu.edu.au/db, http://www.mydomain.org/site - owner

http://www.cit.gu.edu.au/db, David Billington

http://www.cit.gu.edu.au/db

http://www.mydomain.org/site - owner, David Billington

http://www.mydomain.org/site - owner

David Billington

在这一过程中,可以通过本体中的推理规则进行基于本体层次(类层次、属性层次)及本体中概念实例(实例属性)的推理,获得隐藏于文档中的信息。将通过推理得到的三元组同样经过转化后作为附加信息添加到 Web 文档中去。

3 基于语义的 Web 信息检索的模型

通过以上分析,提出了一种基于语义的 Web 信息检索模型,如图 3 所示。

在该模型中,首先由用户选择与其查询相关的领域本体,然后输入关键字。若用户输入的关键字是领域本体中的概念,则通过用户选择的领域本体中的知识来对用户输

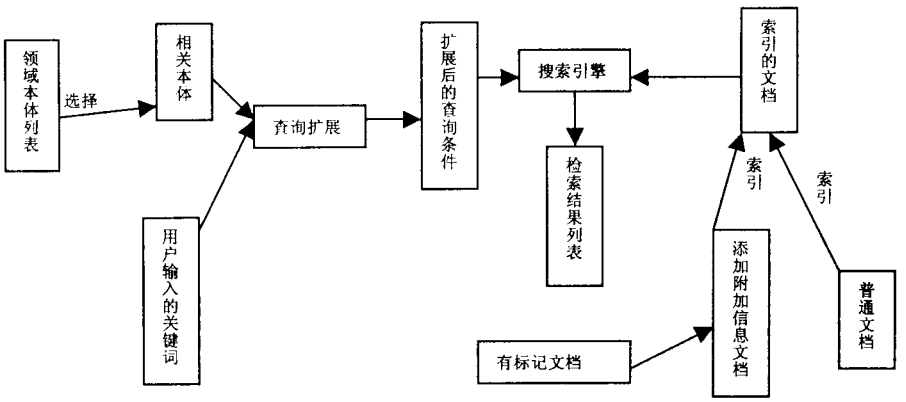

图 3　基于语义的 Web 信息检索模型

入的关键词进行文中所述的 3 种查询扩展,同时通过对扩展以后的各个查询条件的使用频度进行统计,将查询条件按其使用频度进行排序,以方便用户进行选择,搜索引擎根据用户选择的查询条件进行搜索。若用户输入的关键字不是领域本体中的概念,则不需要进行查询扩展。为了减少用户查询的响应时间,可以考虑在领域本体建立以后,即领域本体所蕴含的知识确定以后,就对本体中的概念进行查询扩展,将扩展结果保存于信息库中,这样可以使得在线处理变得相当简单,即只需从信息库中提取相应概念的查询扩展。由于考虑到目前的搜索引擎无法对包含有语义标记的文档进行索引,对带有语义标记的文档进行了文中所述的在文档中添加附加信息的处理,在处理的过程中还运用了本体的推理机制,找出了文档内的隐藏信息,将获得的隐藏信息同样经过处理后添加到文档中,使得搜索引擎可以对这些附加信息进行索引,这样使得搜索引擎对文档的索引更能反映文档的真实内容。该模型考虑了从两个方面解决传统搜索引擎用户检索返回结果相关度不高的问题,即用户查询的"忠实表达"及搜索引擎的索引能否揭示 Web 文档本质的角度。

4　相关问题

4.1　本体相关问题

面向检索的本体在内容和结构上应该符合检索系统的特点,应该考虑利用已有的主题词表、分类表或其他概念体系来半自动地生成本体,以减少重复劳动。同时随着知识的更新,会出现新的词汇,本体应该能够通过分析网络信息和用户的提问来进行自动更新。例如增加新的概念,对原有概念增加新的属性,重构概念之间的关系等。

4.2　Web 页面语义标注

对当今大量的 Web 文档进行语义标注是实现语义 Web 技术应用的前提,人工标注费时费力,所以有必要研究半自动化或自动化的语义标注方法。目前,基于本体的 Web 文档标注的项目有 SHOE(Simple HTML Ontology Extension)、WebKB(采用概念图表的本体来标注文档的人工标注方法)和采用与文档内容相关的基于知识的本体来标注文档的方法。这些方法都是通过人工处理来标注文档,虽然可以提高标注的准确度,但文档的修改和新文档的产生都需要重新标注。文献[5]采用一种类似于 OntoSeek 项目中采用的辞典,在原有本体上添加语言属性的半自动化的标注方法。

5　结　论

通过分析得出了将语义网络技术与传统搜索引擎相结合的检索模型,以解决搜索引擎返回结果相关度不高及网络过渡发展阶段多种类型 Web 文档(带有及不带有语义标记的文档)的检索的问题,对基于语义的检索系统的实现具有一定的指导意义。

参考文献:

[1] 北京奕天锐新科技有限公司 . 第 3 代搜索引擎初显锋芒[EB/OL]. http://www. 21cnbj. com/industrynews/native2003/2003-05-29-57. html. 2003.

[2] 刘遵雄. 搜索引擎的智能化发展趋势[J]. 科技情报开发与经济,2004(6):211-212.

[3] Berners-Lee T, Hendler J, Lassila O. The Semantic Web [EB/OL]. http://www. sciam. com/print_version. cfm? articleID=00048144-10D2-1C70-84A9809EC588EF21.

[4] Reagle J. RDF in XHTML[Z]. W3C Task Force Document, 2003.

[5] 贺　娇. 基于术语本体的网页标引方法[J]. 情报杂志,2004(3):28-29.

(上接第 70 页)

文中所述方法在 UUDynamics 公司的 iStar 产品的开发上得到实际的应用。

参考文献:

[1] Welling L, Thomson L. PHP and MySQL Web Development [M]. U. S. A: Pearson Education, Inc, 2005.

[2] 陈惠贞,陈俊荣. ASP. net 程序设计[M]. 北京. 中国铁道出版社,2004.

[3] 周世雄. . NET 经典范例教程[M]. 北京. 清华大学出版社,2004.

[4] Powell T, Schneider F. JavaScript 2.0 - The Complete Reference, Second Edition[M]. U. S. A: McGraw-Hill/Osborne, 2004.

[5] Dudney B, Lehr J, Willis B, et al. Mastering JavaServer Faces [M]. U. S. A: Wiley Publishing, Inc, 2004.