●王兰成 李 超 何志浩

数字图书馆都柏林核心集网页文本的知识集成与检索研究

摘 要 数字图书馆面临着进一步提高信息检索质量的问题,基于都柏林核心集的知识集成和检索能够发挥重要作用。结合 Web 网页的特点和本体知识,给出一种都柏林核心集网页文本的数据模型。词义扩展的相似匹配是当前知识检索采用的较为实用的方法,基于该模型采用词素匹配的技术并结合词义扩展的信息检索的一些技术,能使信息有效地存储集成和提高信息利用质量,能使开发的知识检索系统有更好的应用性。图 1。参考文献 7。

关键词 数字图书馆 都柏林核心集 知识集成 知识检索 领域本体

分类号 G252.7

ABSTRACT Digital libraries are faced with the problems of improving the quality of information retrieval. In this aspect, the Dublin Core-based knowledge integration and retrieval can play important roles. In this paper, the authors analyze the characteristics of web pages and ontology knowledge, propose a data model of Dublin Core web pages, and introduce some methods and techniques for efficient information storage integration and better information utilization. 1 fig. 7 refs.

KEY WORDS Digital library. Dublin Core. Knowledge integration. Knowledge retrieval. Domain ontology.

CLASS NUMBER G252.7

数字图书馆信息群的知识共享是多学科信息深度集成的重要基础,实现 Web 信息群的知识共享、集成和挖掘需要创新研究相关关键技术。国际上一些领域已具备较成熟的学科知识集成条件,目前处于研究起步阶段,但在我国由于缺少统一的学科或领域的理论和技术体系支持,该研究尚属空白。随着互联网日益成为一个巨大的信息资源库,人们上网查找所需的信息变得越来越困难,效率低甚至不可能实现。在检索系统或者搜索引擎中输入某个关键词,输出的结果记录可能有成百上千条,然而从中可能找不出想要的信息。有时人们不得不花更多的时间用于甄选和辨别。这就是目前在检索 Web 信息时可能出现的两个问题:答非所问、大海捞针。解决这两个问题需要研究元数据级网页文本的知识集成和提高信息检索的查准率,以使检索结果能真正地反映用户的需要,以更为理想的方式输出和保存结果并方便地提供用户利用。

知识集成是通过领域本体技术将数字信息群的知识按照不同的主题进行组织,形成等级类目,同时规定类目的特性及其之间的关系。它对已经存在的多个异构数据库,在尽可能少地影响本地自治性的基础上,构造具有用户所需要的某种透明性的分布式数据库,以支持对物理上分布的多个数据库的全局访问和数据库之间的互操作性。在集成架构中加入数据仓库元素,利用数据仓库对集成的知识数据进行统一视图的组织和管理,利用知识集成的元数据对跨领域数据进行描述,并在该元数据基础上建立各类视图[1]。目前许多搜索引擎的数据搜索都是位于网络表层的静态信息,无法挖掘到位于数据库里的深层数据,从而面临着 3 个难题:一是如何从数据库得到请求响应;二是如何将搜到的数据进行组织;三是如何整合这些信息并呈现出来。第二代相关性技术无法做到这一点,采用知识集成的方法则有可能达到这一目标。由于本体具有良好的概念层次结构和对逻辑推理的支持[2],我们可以在检索时进行基于概念的、语义上的匹配,在查准率和查全率上有更好的保证,从而实现知识检索。

1 都柏林核心集网页文本的数据模型

数字图书馆信息量庞大,形式和类别繁多。用户感兴趣和需要的往往是其中的某一小部分或某个领域,甚至是某个特定主题。例如,研究军事的学者所关心的往往是与军事相关的军事信息,商人关心的往往是市场行情及国家的有关政策情况,而一个关心国内外大事的市民,可能更关心时事要闻、国家远景规划等等。这就是说,用户的信息需求具有领域和主题

相关性。这就为我们基于领域本体进行 Web 信息个性化集成提供了可能。

目前的网络信息资源主要是包含大量文本信息的 HTML 网页,HTML 是一种半结构化语言,它提供的各种格式标记缺少对网页内容的描述。要对 Web 信息进行结构化获取和保存,首先需要对因特网数据建立一个模型。简单地说,需要对 HTML 网页的文本信息建立数据模型。目前已有"点—线"模型、基于 DOM 的网页数据模型、基于填充标记的网页数据模型等。"点—线"模型认为,可以用由点(网页)和线(网页间的超链接)构成的图来建立因特网上的数据模型。线将一个个点联接成一个网状结构,点和线中都有重要的信息,由此可将文本型数据源归纳为网页的标题(点)、网页的正文(点)、网页的超文本标记(点)和网页间的超链接(线)等部分。基于 DOM 的模型认为可以把网页中所要提取的有用信息作为 DOM 层次结构中的路径,并可将这些信息在 DOM 层次结构中的路径作为信息提取的"坐标",完成信息的提取工作。基于填充标记的网页数据模型则认为可以通过在 HTML 文档相关内容中填入标记完成数据模型的构建。

都柏林核心集 Dublin Core(DC)是生成一个简单的、并且在网络中为各个数字图书馆用户所接受的标准化元数据元素集,它由 15 个核心元素构成,还可被扩展或与其他元数据进行桥接。DC 元数据的表达有多种方式,一些简单的表述可以采用 DC 在 HTML、XML 和用 XML 格式的 RDF 结构中的镶嵌形式,目前已成为简单描述数字图书馆网络资源的首选元数据方案。综合分析目前的多种网页数据模型,我们提出一种基于都柏林核心集的 DC 网页文本知识数据模型,即把一个认为有效的 HTML 网页用主题(Subject)、标题(Title)、出版者(Publisher)、关系(Relation)等核心集元素来标注,然后再以 XML 格式生成结构化的模板以便于网页信息的保存和集成。这种使用 DC 元数据建立的 HTML 网页模型 Ψ 为:

Ψ(名称,标签,说明与描述) = {

(主题,Subject,从标题串中抽取表达主题概念的词);

(标题,Title,网页标题);

(作者,Author,资源内容的主要创建人或组织);

(出版者,Publisher,网站名);

(描述,Description,网页内容的文本描述);

(其他参与者,Contributor,友情链接);

(发布日期,Date,网页发布或更新日期);

(类型,Type,网页资源所属领域或种类);

(格式,Format,网页编码格式);

(标识,Identilfier,能够唯一标识资源的字符串或数字);

(关系,Relation,网页上的相关链接网址);

(来源,Source,网页所在网站的网址);

(语言,Language,网站内容的描述语言);

(时空性,Coverage,资源的空间或时间特性);

(版权,Rights,网页资源的版权)}

Ψ 网页的存储格式为:

```
<TemplateInstatnces>
    <Subject>... </Subject>
    <Titte>… </Title>
    <Author>… </Author>
    <Publisher>… </Publisher>
    <Description>… </Description>
    ...
    <Rights>… </Rights>
</TemplateInstatnces>
```

都柏林核心集的 15 个核心元素在所建的 A 网页数据模型里一对一映射。数字图书馆选用 DC 建立网页数据模型的特点表现在两个方面:首先是 DC 元数据简单易用,是一种描述网络信息资源属性的有效手段,它通过对资源内容的粗略描述,可以比较好地体现电子资源的主要属性,能够极大地提高检索结果的准确定位率;其次是 DC 元数据具有较好的通用性和兼容性,DC 并不针对某个特定的学科或领域,元数据本来就是一种用于描述 Web 信息资源的标准,对 HTML 网页数据描述有比较好的兼容性。采用 DC 建立的网页数据模型,可以使信息提取获得的数据能够以一种比较通用和标准的格式进行结构化保存,实现网页文本的知识集成与检索。

2　基于 Ψ 网页模型的知识集成与检索

基于 Ψ 网页模型的实现数字图书馆的知识集成与检索,涉及信息处理链中的知识提取、语义匹配和检索结果过滤等多方面的智能化技术。

(1)信息提取。信息提取是在定义了信息提取模式的基础上,对特定信息进行提取和保存的过程。目前已提出的信息提取方法有基于超链接和文本标记加权的方法、填充标记的方法、基于 DOM 的方法和基于实例模板的方法等[3]。HTML 文档标记,如 <title>… </title>, <B>… </B> 等等,通过分析提取相应的网页属性信息,然后填充表至模板保存。这一步工作使检索结果得以结构化地保存,极大方便了对检索结果的管理和利用。

由于数字图书馆用户通常查询的是同一主题或

领域的信息,这样每次经过处理后保存的信息必定与已有的保存结果有概念上的密切联系。如基于Ψ网页模型,用户第一次查找的是关于"美食"的信息,经过处理后得到的是与"美食"有关的结构化文档;第二次查找的就可能是关于"景点"的信息,经过处理后得到与"景点"相关的结构化文档。不难发现,这两类文档其实都与"旅游"这一主题有关,该用户可能比较关心"旅游"方面的信息,于是两份文档就可以建立相应的联系,加一个共同的父类别"旅游"信息等。通过这样不断地关联各个文档,就逐渐形成了用户关心的这一主题的结构化信息库,从而实现在领域本体知识指导下的个性化的信息集成,即知识集成。

(2)词义扩展和词素匹配。基于Ψ网页模型的知识检索系统提供词素匹配与精确匹配的方法,并具有同义扩展与相关扩展的功能。系统在用户输入关键词后,可选择词素匹配或是精确匹配,并可选择同义扩展或相关扩展。选择词素匹配时,系统将对用户提交的检索式进行切分。目前汉语自动分词主要分为基于统计的机械分词方法和基于规则的分词方法[4],机械分词方法又可分为正向最大匹配(Maximum Matching,MM)和逆向最大匹配(Reverse Maximum Matching,RMM)方法[5]。本文采用基于RMM的方法作为系统的自动分词方法,匹配切分词表,得到扩充的检索式集合,然后参与匹配过程。精确匹配则不需要对检索式进行切分。该处理流程如图1所示。

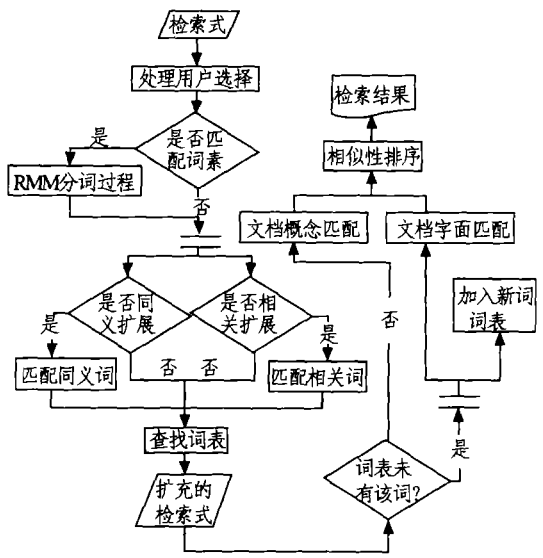

图1 有词义匹配的知识检索处理流程

由于对检索式的分词采用目前较为实用的机械分词的方法,在选择词素匹配时检索速度会受到一定的影响。精确匹配则不对检索式进行分词,因此用户在检索式表意明确的情况下,选择精确匹配能够大大提高检索速度。另一方面,同义扩展及相关扩展,可使用户的检索需求得到比较充分的扩充表达,从而提高检全率,但对检索速度会有一定的影响;同时由于检索式的扩充,可能会降低系统的检准率。

(3)输出结果的相似性排序。在与文档集进行相似匹配之后,还要根据文档与查询的匹配程度对结果进行相似度排序,以保证重要的、最贴近用户需求的文档排在前面,方便查阅与利用。本系统采用的主要排序算法是文献[6]中提供的中文同义词相似匹配的算法。该算法结合了中文词匹配方法中基于语素的字面匹配和基于词素的语义匹配两种方法,具有较高的实用性。

算法1:字面加词素概念的相似匹配算法。

输入:长度分别为m和n的A、B两个词。

输出:A、B两个词的相似度Sim(A,B)。

步骤:

a. A、B两个词的相同汉字个数为t($0 \leqslant t \leqslant Min(m,n)$),其中Min(x,y)(x,y∈N)表示正整数x,y中较小的数。

b. 求字面相似度[7]:$Sim(A,B) = \frac{1}{2}\left(\frac{t}{m} + \frac{t}{n}\right)$

在引入词位置权值后,该算法有多种改进版本,对于由纯汉字构成的词汇的相似度计算有较好的精度,但对非纯汉字组成的词汇、同一事物的学名与别名、全称与简称等情况准确率较低。

c. 定义表达度E:词的部分对整体含义所起作用的量度($0 < E \leqslant 1$)。

设待匹配词A、B的信息量总和分别为T_A,T_B,共同部分C对A的表达度为E_A,对B的表达度为E_B,即 $E_A = \frac{C}{A}$, $E_B = \frac{C}{B}$。

d. 求语义相似度:$Sim(A,B) = \frac{2E_A \times E_B}{E_A + E_B}$

本文综合上述两种相似度计算方法。从上一节的处理流程可知,当词表中没有用户输入的词时,系统通过关键字匹配的方法输出检索结果。此时结果排序就只能采用字面匹配的方法。

算法2:检索结果过滤算法。

输入:未过滤的检索结果文本OLD。

输出:过滤后的检索结果记录NEW。

步骤:

a. 对OLD文档去重和归类处理:去除重复信息,去

除文本和网页以外的多媒体信息;声音、视频等信息的提取、保存方法及数据模型不同于处理文本网页。

b. 对由 a 得到的文档的标题部分进行主题提取,获得该网页的主题;标题与网页内容的关系非常密切,通常起着概括全篇的作用。

c. 将用户输入的关键词在领域本体中进行概念匹配,获得用精确概念表述的用户信息需求。

d. c 与 b 得到的网页主题概念进行匹配。

e. 匹配成功的记录 NEW 转入下一步的信息提取流程。

通过这样一个过滤流程,基本上筛选出了真正与主题需求相关的记录,为下一步信息提取提供了良好的语料。

3　进一步的实验与结论

将通过 Google 检索获得的 80000 多条与"嵌入式系统编程"相关的记录进行处理,最后得到 1500 多篇输出的 XML 文档,这些文档的内容完全符合我们对信息的需求,都是与"嵌入式系统编程"相关的内容。我们又检索出与"移动编程"相关的 10000 多条记录,经过处理后得到的 2200 多条记录,也是基本与我们的需求相关的内容。通过上述的领域本体,基于 Ψ 网页模型生成的文档结构有如下的示例:

```
<TemplateInstatnces>
    <Rank Type="Parent"  Identifier=Computer Science. Software. ProgrammingTechnology>
    <Rank Type="Brother"  Identifier=Computer Science. Software. Programming. MobileDevelopment>
    ...
    <Subject>嵌入式编程</Subject>
    <Title>嵌入式编程的基本原理</Title>
    <Author>...</Author>
    <Publisher>...</Publisher>
    <Description>...</Description>
    <Rights>...</Rights>
    ...
</TemplateInstatnces>
```

可见通过领域本体的后端控制,文档内容间的关联已经能够体现出来并加入到结构当中。实验表明是成功的,但该集成方案还有许多有待完善的地方,如现在许多网站其网页标题是通过一些脚本程序自动生成的,而标题名可能并不能真正反映正文的内容,因此对搜索引擎输出的结果经上述过滤方案时有可能丢失一些有用的信息。另外,文中构建的网页数据模型中包含了 DC 的全部 15 项属性,这其中有的属性在一般标记的网页中很难提取出来,需要进一步构建和完善实用的领域本体。

基于 Ψ 网页模型的知识检索系统,融合了当前中文主题概念检索的包括同义扩展、相关扩展等多种新的技术。当用户输入的关键词不是系统所收录的词,增加的新词处理功能则保存至新词词表中并记录其被使用的频率。结合字面匹配和词素匹配各自的特点,综合运用了这两种方法,从而使得输出结合的排列顺序与其匹配程度尽可能一致。目前词汇的条数有限,只能在受限领域实现一定的概念检索效果;分词方法采用了逆向最大分词算法,可达到比较高的准确度,但算法的效率还需要提高。

数字图书馆面临着进一步提高信息检索质量的问题,基于都柏林核心集的知识集成和检索能够发挥重要作用。本文给出的网页文本数据模型及其词素匹配结合词义扩展的信息检索技术,对 Web 信息的有效存储集成和提高信息利用的质量,都具有重要的应用价值。

参考文献

1　Ronald Larsen, Howard Wactlar. Knowledge lost in information, Report of the NSF Workshop on rescarchdirections for Digital Libraries. June I5 - 17. 2003. Chatham. MA

2　Alexander Maedche, Steffen Staab. Mining Ontologies from Texl. AIFB. Univ. Karlsruhe. D - 76128 Karlsruhe. Germany

3　Alexander Maedche, Gunter Neumannl, Steffen Staab. Bootstrapping an Ontology-based lnformalion Extraction System. [2007-01-10]. http//www. aifb. uni-karl Sruhe. de/-sst/Research/publications/WebExplorebook. pdf

5　J. F Martinez-Trinidad. A Tool To Discover The Main Themes In A Spanish Or English Document, Expert System With Applications, 2000

6　王兰成,李超. 改进的中文同义词相似匹配方法. 中国图书馆学报,2005(3)

4,7　朱毅华等. 计算机识别汉语同义词的两种算法比较和测评. 中国图书馆学报,2002(4)

王兰成　解放军南京政治学院上海分院信息管理系教授,博士生导师。通信地址:上海市南京政治学院上海分院信息管理系。邮编 200433。

李超　何志浩　解放军南京政治学院上海分院信息管理系研究生。通信地址同上。　(来稿时间:2006-07-17)