基于本体的垂直搜索引擎的研究

何凤英

(福州大学 数学与计算机科学学院,福建 福州 350002)

摘 要:分析了当前网上搜索引擎的现状及存在的问题,提出了一种结合本体的垂直搜索引擎构建思想,并阐述了垂直搜索引擎构建的关键技术,最后设计实现了一个以电子杂志为主题的垂直搜索引擎原型。

关键词:垂直搜索引擎;本体

中图分类号:TP393 **文献标识码**:A

Research of vertical search engine based on ontology

HE Feng Ying

(College of Mathematics & Computer Science, Fuzhou University, Fuzhou 350002, China)

Abstract: This paper analysis the existing problems in the current search engine, presents a construction method for vertical search engine utilizing ontology and then discusses a set of key technologies for the construction of vertical search engine. Finally a prototype of electronic magazine search engine is implemented.

Key words: vertical search engine; ontology

随着 Internet 的普及和推广,人们越来越依赖于互联网络进行各种商务活动和信息查询,因此网络信息查询已经成为人们研究和讨论的热点领域。目前主要的搜索引擎(如 Google、百度等)大多采用基于关键词或者基于内容分类的检索技术,很少具有进一步的智能化。检索结果不可避免的出现垃圾信息,查全率和查准率都存在一定的不足。

近年来,本体技术的逐步成熟为信息检索技术的发展带来了新的动力。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,为提高检索系统的查准率和查全率提供了更好的保证。

垂直搜索是针对某一个特定领域的专业搜索。它和普通网页搜索的最大区别是对网页信息进行了针对性的信息抽取和组织。由于关注的范围确定,它在特定领域的搜索效果好于一般的搜索引擎。

本文设计了一种基于本体的垂直搜索系统 SVSE (Semantic Vertical Search Engine),主要的思路是利用本体,在用户提问搜索式构造过程中增加语义推导,消除自然语言理解中的歧义,更加准确地反映用户的真实信息需求,同时加强搜索引擎的推理功能,在完成对信息源搜索的基础上,根据相关概念以及相关背景知识进行推理,挖掘出文本中的隐含信息,从而实现基于概念的智能搜索。

1 SVSE 系统总体设计架构

在设计和实现基于本体的垂直搜索引擎时,本文采用的方法是简化整个信息构建过程和检索流程,通过逐步扩展的方式提高搜索引擎的语义检索功能,主要模块采用可以更替核心算法的构建方式。SVSE 的体系结构如图 1 所示。

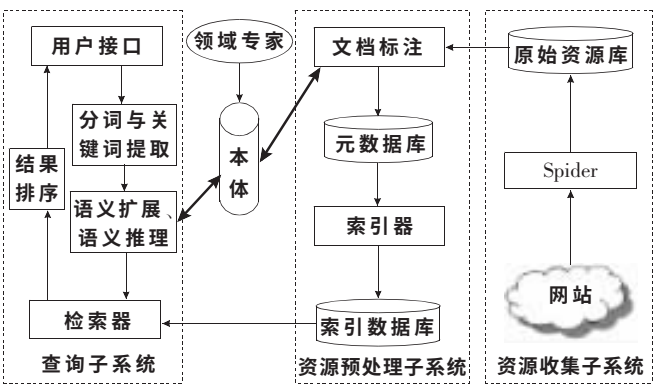

图 1 SVSE 体系结构

其中各个系统模块的主要作用如下：

(1)资源收集子系统。该模块主要负责对网站进行遍历，从而将网站上的资源抓取到本地的资源库中。对资源的收集是搜索引擎进行工作的基石，资源内容收集的好坏直接影响到系统的检索结果。

(2)资源预处理子系统。该模块主要负责对收集来的资源在本体库的指导下进行语义的标注，提取出文档的特征并对原始资源进行格式上的整理形成元数据，然后对元数据索引处理，以提供给查询子系统进行查询。

(3)查询子系统。该模块主要负责对用户的查询词进行处理，在本体的指导下对关键词进行语义扩展和语义推理，将经过扩展的查询词在系统的索引库中查询，最后将搜索结果经过一定的排序规则排序后返回给用户。

2 SVSE 系统设计的关键技术

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间相互关系的明确定义。

Hownet[1]是一个目前已被广泛认可的人工构建的本体，它提供了词汇的各种常用含义，每种含义下的同义词以及词语之间的语义网络，本文将直接采用 Hownet，而不另外构建本体。

2.1 基于本体的定题爬虫设计

资源收集子系统的目的是按照一定的搜索策略在互联网中发现新的网页信息，其关键在于定题爬虫的设计。目前定题爬虫通用的做法是根据网页中的关键词判定进行某种主题的过滤，但由于存在一词多义及一义多词的现象，这种基于关键词的判定策略已被证实精确度不高[2]，会遗漏许多相关页面或添加许多噪音页面。本文提出一种基于本体的主题相关性判定策略，利用 ontology 对领域概念及概念间关系的明确定义来提高判定精度。其框架结构如图 2 所示。

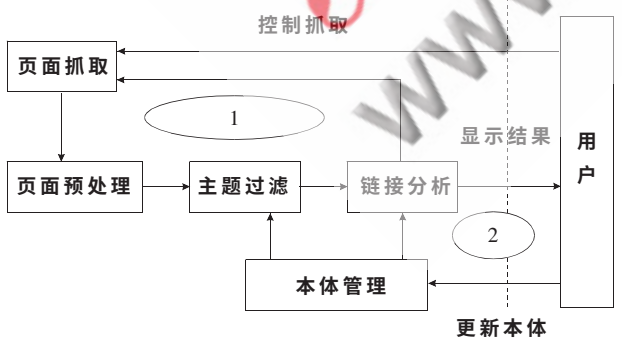

图 2 基于本体的定题爬虫框架

框架结构由 2 个循环组成：网络爬行循环和本体循环(见图 2 中的 1 和 2)。网络爬行循环从页面抓取开始，按箭头的方向形成循环 1，不停地从网络中取得主题相关信息，将结果显示给用户，同时用户也可以修改控制参数控制页面的抓取。其中进行主题过滤和链接分析时需要用到本体管理提供的本体知识作为评价依据。而本体循环从本体管理开始到主题过滤，按箭头方向形成循环 2。本体管理是用户先在领域专家的协助下明确本领域的共享概念及其概念间的关系，构建领域本体，然后根据在实际爬行过程中出现的高频率新概念进行本体的更新与维护。

2.1.1 基于本体的主题过滤

传统定题爬虫采用的是基于关键词的主题过滤，它对网页相关性的判定主要使用向量空间模型 VSM。

令 $p=\{p_1,p_2,\cdots,p_n\}$ 表示网页集合，$k=\{k_1,k_2,\cdots,k_t\}$ 表示主题特征集，k_i 为主题特征项，t 为特征项的个数，则网页 p_j 可表示成 $P_j=(w_{1,j},w_{2,j},\cdots,w_{t,j})$，其中，$w_{i,j}=(tf_{i,j}*idf_j)$，$w_{i,j}$ 表示特征项 k_i 在网页 p_j 中的权值，$tf_{i,j}$ 表示特征项 k_i 在网页 p_j 中出现的频率，idf_j 称为逆文献频率，为网页 p_j 中出现了特征项 ki 的页面数的倒数。

而主题特征向量 R 可表示成 $R=(w_{1,r},w_{2,r},\cdots,w_{t,r})$，$w_{i,r}$ 表示特征项 k_i 在主题特征向量 r 中的权值，可以在大量主题示例网页集中经过计算得出，也可以由专家给定。

在向量空间模型中，网页 p_j 的主题相关性可由其对应向量 P_j 与主题特征向量 R 的夹角余弦来计算，公式如下：

$$sim(P_j,R)=\frac{\sum_{i=1}^{t}(w_{i,j}\times w_{i,r})}{\left(\sqrt{\sum_{i=1}^{t}w_{i,j}^2}\times\sqrt{\sum_{i=1}^{t}w_{i,r}^2}\right)} \quad (1)$$

若 $sim(P_j,R)\geqslant\theta$，$\theta$ 为设定阈值，则认为页面为主题相关，保存入数据库，否则丢弃。

而基于本体的主题过滤则根据语义判定主题相关性。它将页面 p_j 中与主题词 k_i 具有相同概念的其他关键词(主要指其同义词及上、下义词)都替换成概念 k_i，这样页面 p_j 与主题特征向量 R 就能实现语义层次上的相似性判断，并且不会增加主题特征向量 R 的维数，也不存在新添加概念的权值确定问题。具体过程如下：

(1)语义化页面 p_j 为 p'_j。算法描述如下：

<输入>页面 p_j 中每一个名词 s_i；主题特征集 k；基于语义的新页面 p'_j(初始时与原页面 p_j 相同)；

<输出>每个特征项 k_i 在语义页面 p'_j 中总共出现的频率 $tf_{i,j}$；在页面中各个位置中出现的频率 $tf_{i,j,n}$，n 表示特征词出现位置(如标题，链接，加强文本等)；

将每个 s 及用 Hownet 扩展出的 s_i 的同义词及上、下义词存入集合 temp；

对 k 中的每个特征项 k_i

{If(k_i 在 s_i 的集合 temp 中)

{ 将 p'_j 中的 s_i 替换成 k_i；

$tf_{i,j}$++； //计算 k_i 出现频率

根据 s_i 在页面中位置 n，计算 k_i 在页面中

不同位置出现的频率 $tf_{i,j,n}$;} } }

(2)计算语义页面 p'_j 的特征向量 $\boldsymbol{P}'_j=(w_{1,j},w_{2,j},\cdots,w_{t,j})$。

$w_{i,j}$ 的计算公式为:

$$w_{i,j}=\lambda\times f_{i,j}\times w_{i,r} \tag{2}$$

其中,λ 表示位置加权系数,计算公式为:

$$\lambda=\sum_{n=1}^{3}[f_{(n)}\times tf_{i,j,n}] \tag{3}$$

$f_{(n)}$ 表示位置权值,本文根据实践经验并参考别人的研究成果[3],认为网页中的锚文本(anchor text)最能反映页面内容,应赋予最高权值;而标题(title)、大标题(H1、H2)、加强文本(strong)也比较能反映页面内容,赋予次高权值。$f_{(n)}$ 具体赋值如下:

$$f_{(n)}=\begin{cases}2, & n=1,\text{在 anchor text 中}\\ 1.5, & n=2,\text{在 title/H1/H2/strong 中}\\ 1, & n=3,\text{其它}\end{cases}$$

而 $f_{i,j}$ 表示页面 p_j 中的特征项 k_i 的标准化频率,计算公式为:

$$f_{i,j}=tf_{i,j}/\max_i(tf_{i,j}) \tag{4}$$

$w_{i,r}$ 表示特征项 k_i 预先给定的权值。

(3)用公式(1)计算与主题特征向量 R 的相关性。

2.1.2 基于本体的链接分析

链接是爬虫工作的基础,主题相关页面中并非所有的链接都是主题相关的,在下载所有链接的页面内容前进行一次链接预测,去除那些明显不相关的链接,爬虫的效率能得到进一步提高。

如果令 anchor_text_area anchor_textanchor_text_area 表示一个页面链接 hyperlink 的锚文本 anchor_text 及其链接附件文字 anchor_text_area,则链接相关度预测算法首先分析在 anchor-text 及 anchor-text-area 中出现的主题特征词的数目,若超过一定阈值 γ,则预测 hyperlink 主题相关,放入待处理队列 Q 中等待下载页面内容;若低于阈值 γ',并不马上放弃,因为此 hyperlink 很可能是主题页面的前驱链接[4],然后继续分析 anchor-text 及 anchor-text-area 中出现的前驱 N 层主题特征词的数目,若超过阈值 γ,则也给此 hyeperlink 继续被处理的机会,这样一定程度上能跳出定题爬虫局部搜索的通病。根据实践经验取 $N=2$,前驱两层的主题特征集 k' 由人工选择给出,anchor-text-area 取 30 个字节。算法描述如下:

<输入>:anchor_text, anchor_text_area, hyper_link, 主题特征集 k, 前驱 N 层主题特征集 k', 阈值 γ,γ';

<输出>:待处理队列 Q;

对每一个 hyperlink

{检查 anchor_text 和 anchor_text_area 中出现主题特征集 k 中特征项的次数 f;

If($f\geqslant\gamma$){将 hyperlink 放入待处理队列 Q;}

else{对每个前驱主题特征集 k'

{检查 anchor-text 和 anchor-text-area 中出现主题特征集 k' 中特征项次数 f';

If($f'\geqslant\gamma'$){将 hyperlink 放入待处理队列 Q;

return;//返回,当前 hyperlink 检查完毕,检查下一个 hyperlink}}

将 hyperlink 放入抛弃队列 Q'; //hyperlink 即非主题相关链接也非前驱链接,抛弃}}

2.2 基于本体的结构化信息抽取

Web 数据大多是无结构或半结构化的,十分不利于信息检索,资源预处理子系统的目的就是将网页中的非结构化数据按照一定的需求应用本体抽取成结构化数据以提高查询的准确率,其关键在于对文本进行语义分类。

参考文献[5]设计了 3 个函数:函数 terminology(O_i)从领域 D_i 对应的本体 O_i 中求出该领域的术语集(包括同义词);函数 definition(O_i, keyword)从本体 O_i 中求出关键字 keyword 的定义;函数 relation(O_i)从本体 O_i 中求出由概念关系构成的语义网络集。

本文借鉴参考文献[5]的思想设计了语义分类算法。设 $O_1,O_2,\cdots,O_n$ 分别是领域 $D_1,D_2,\cdots,D_n$ 的本体,术语集 T_i=terminology(O_i),其中($0\leqslant i\leqslant n$),KS={$key_1,key_2,\cdots,key_m$}为被检索文档 Doc 中所给出的关键字,则具体过程如下:

(1)过滤与文档不相关的领域

任一文档中所给出的关键字应体现该文档最核心的内容,这些最核心的内容若不出现在该领域的本体中,则说明该文档与这一领域无关,即 $KS\cap T_i=\Phi\vdash doc\notin D_i$,$1\leqslant i\leqslant n$。而所有可能与该文档相关的领域记为 $D_{s1},D_{s2},\cdots,D_{sk}$,其中 $KS\cap T_{sj}\neq\Phi$,$s_1\leqslant s_j\leqslant s_k$。

(2)近似语义网络匹配

首先求出与关键字的定义相关的术语集合 DS={dk|($dk\in T_{sj}$)∧(dk 出现于 key_i 的定义 definition(O_{sj}, key_i)中) $key_i\in$ KS, $1\leqslant i\leqslant m$, $s_1\leqslant s_j\leqslant s_k$}, 然后求与关键字集直接相关的术语对象集合 RO={obj|($\exists x(x(KS\cap T_{sj})\wedge(obj,x)\in$ relation(O_{sj})), $s_1\leqslant s_j\leqslant s_k$}。

(3)文档的语义标注

检索整个文档,统计被检索文档里出现在集合 DS∪RO 中元素的频度 $freq_{sj}$。$freq_{sj}$ 体现了该文档中的术语与 O_{sj} 中的语义网络的近似匹配程度。本文定义 degree(D_{sj})=$freq_{sj}$,根据 $D_{s1},D_{s2},\cdots,D_{sk}$ 与被检索文档的相关程度 degree(D_{sj})的大小来决定文档属于哪个领域。

2.3 基于本体的检索

查询子系统设计的关键在于对用户查询关键词进行语义扩展,从而将检索上升到本体中概念匹配的高度,以提高查全率。具体过程如下:

(1)将输入关键字集合中的每个关键字用 Hownet 扩展出它的同义词及上、下义词,得到概念集合 $A(c_1,c_2,\cdots,c_n)$。

(2)设定语义相似度阈值 η,并对概念集合进行扩充。对概念集合进行扩展时,利用相似度计算结果筛掉小于设定阈值的概念,保留超过设定阈值的那些概念,最后得到查询概念的扩充概念集。语义相似度的计算采用文献[6]的方法。算法描述如下:

①初始化空集合 $B=\{\}$;

②利用函数 terminology(O_i)获取领域本体概念集,并求出领域本体概念集与关键字概念集的差集 $C=\{x_1, x_2 \cdots, x_m\}$;

③若概念集 C 中的概念 x 与关键字原始概念集合 A 中的任意两个概念的相似度大于给定的阈值,则将概念 x 加入到概念集 B 中。

④将集合 A 与集合 B 合并作为新的扩展概念集 D。

(3)利用扩充概念集 D 中的概念形成查询向量,然后根据向量空间模型进行信息查询,找到满足特定条件的页面。

3 SVSE 系统的实现

鉴于现实世界正飞速地向电子信息化方向转变,以书籍、杂志、报刊等获取信息的方式显得尤为突出,SVSE 系统以“电子杂志”为特定主题,提供按关键字和发布日期搜索杂志的功能。

在具体的实现过程中,我们从《中国分类主题词表》中由人工精选得到主题特征集 k,并在一个约 500 KB 的示例主题页面集中计算得出各主题特征项权值 $w_{i,r}$,同时根据这些示例页面集的前驱两层页面,提取出前驱两层主题特征集 k'。对本体的操作采用 Jena 工具包,在资源信息搜集上选择以目前网上流行的规模较大且内容较全的 ZCOM 网做为测试抓取网站,抓取工具使用 Heritrix,全文索引使用 Lucene。实验中主题相关性域值 θ 取 0.3,链接相关度域值 γ,γ' 分别取 2 和 3,语义相似度域值 η 取 0.4。

系统以 PHP 语言实现,运行在一台装有 Windows XP 系统,CPU 为 P4 1.5 GHz,硬盘为 160 GB 的 PC 机上。

以“奥迪”为关键字进行搜索,系统把用户的查询信息通过语义推理模块形成本体内的表达式进行查询、判断最后把结果显示如图 3 所示。

图 3 搜索结果

从图 3 可以看到,共搜索出 20 条结果,分三页显示。显示的是电子杂志名或是摘要内容里有和关键字“奥迪”语义相关的结果,同时每条纪录还提供了杂志详细信息、收藏该杂志与下载等链接。

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查全率和查准率太低的问题,并针对其中的关键技术给出具体的实现算法。但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地解决,所以本文的垂直搜索引擎还处于原型阶段,其成熟和完善还有不少工作要做,这也是我们下一步努力的方向。

参考文献

[1] 董振东,董强.Ontology 和 HowNet[EB/OL].http://www.keenage.com/html/c-index.html.2003-08/2006-02.

[2] EHRING M, ERMAEDCHE A. ontology focused ling of Web documents[J].Proeedings of the 2003ACM Sympos i-umon Applied ComPuting,2003,1(3):624-626.

[3] CUTLER M, SHIH Y, MENG W. Using the structure of HTML documents to improve retrieval [A]. Proceedings of the USENIX Symposium on Internet Technologies and Systems Monterey [C]. California: California Press, 1997:241-251.

[4] Mdiligenti F C. Focused crawling using context graphs[A]. Proceedings of the 26th International Conference on Very Large Data Bases[C]. Cairo: Cairo Press, 2000: 527-534.

[5] 武成岗,焦文品,田启家,等.基于本体论和多主体的信息检索服务器[J].计算机研究与发展,2001(6):55-57.

[6] 何凤英.基于本体的网格服务匹配算法的研究与实现[J].计算机应用,2008(4):863-865.

(收稿日期:2009-03-02)