云南大学学报 (自然科学版), 2003, 25 (4): 324～327 CN 53 - 1045/ N ISSN 0258 - 7971
Journal of Yunnan University

基于本体论的信息检索*

郭祥文, 刘惟一, 钱 民, 张忠玉
(云南大学 计算机科学与工程系,云南 昆明 650091)

摘要:将本体论应用于信息检索,提出了基于本体论的信息检索模型. 该模型支持用户查询的导引,并按领域分类有选择地返回查询信息.

关键词:本体论;信息检索;全文检索;领域分类
中图分类号: TP 182 **文献标识码**:A **文章编号**:0258 - 7971(2003)04 - 0324 - 04

随着信息技术的发展,特别是 Internet 应用的普及,人们已从信息缺乏的时代过渡到了信息极大丰富的时代. Internet 上信息分布在位于不同位置的站点上,据统计到 1997 年夏季已经有 1.5 亿个 Web 主页分布在 65 万个站点上[1].

目前网络上的搜索引擎一般使用 2 种技术来实现信息检索:一是使用网站分类技术,即把网站进行树状的归类,登录的网站属于至少一个类别,对每个站点都有简略的描述. 雅虎采用了这种方法. 为了分类科学准确,需要有一支由各科人才组成的维护队伍. 二是使用全文检索技术. 全文检索技术处理的对象是文本,它能够对大量文档(这里是大量网页数据)建立由字(词)到文档的倒排索引,在此基础上,用户使用关键词来对文档(网页)进行查询时,系统将给用户返回含该关键词的网页.

一般来说,由于使用了专家来对网站进行归纳和分类,网站分类技术为网络信息导航带来了极大的方便,受到人们的欢迎. 但是它维护成本较高,而且对网站的描述也十分简略,其描述能力不能深入网站的内部细节,因此用户不能查询网站内部的重要信息,造成了信息丢失.

全文检索是一个很成熟的技术,它能够解决对网页细节的检索问题. 从理论上说,只要网页上出现了某个关键词,就能够使用全文检索用关键词匹配把该网页查出来,但是这又导致了它的缺陷:返回的信息太多. 更严重的是,除了综合性的搜索引擎站点有这个现象之外,现在较大的站点对自身站内信息的检索也会返回大量的网页. 传统的文本信息检索一般使用查全率(Recall)与查准率(Precision)来对检索效果进行量化评价,但是在信息海量的互联网上,信息检索用查全率与查准率来衡量检索效果不太合适. 在一些场合,高的查全率带来的成千上万的命中网页. 在网页爆炸性增长的今天,没有一个用户有时间和精力来浏览搜索引擎查到的网页. 当前的搜索引擎的缺点是不支持用户的信息导引. 本文提出了基于本体论的信息检索,支持领域的分类,并按领域分类有选择地返回网页,提高了检索的效率.

1 本体论的基本概念

本体论是对概念化对象的明确表示和描述[2,3]. Nicala Guarino 把概念化对象 C 定义为

$$C = (D, W, P),$$

其中:D 是一个领域;W 是该领域中相关的事务状态的集合;P 是领域空间$\langle D, W\rangle$上概念关系的集合.

本体论是采用某种语言 L 对概念化的描述,因此本体论依赖于所采用的语言 L. 按照表示和描述的形式化的程度不同,可以分为:完全非形式化的、半非形式化的、半形式化的和严格形式化的本体论[4]. 形式化程度越高,越有利于计算机进行自

* 收稿日期:2002 - 05 - 27
基金项目:云南省自然科学基金资助项目(69763003).
作者简介:郭祥文(1975 -),男,云南人,硕士,主要从事数据库与知识工程的研究.

动处理.

从概念化对象的定义来看,一个领域的术语、术语的定义以及各个术语之间的语义网络应是任一个领域本体论所必须包含的基本信息,同时本体论中还应包含关于同义词的描述.

2　基于本体论的信息检索模型

我们提出了基于本体论的信息检索模型,如图 1 所示.

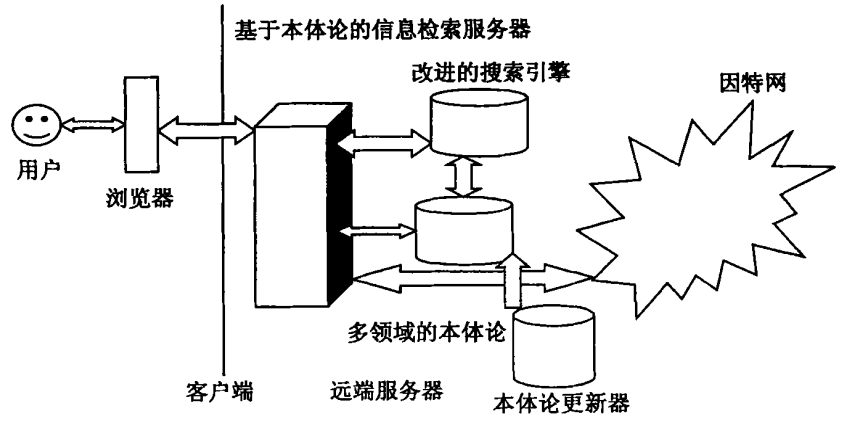

图 1　基于本体论的信息检索模型

Fig. 1　Information retrieval model based on ontologies

2.1　多领域的本体论　领域分类 $D = \{d_1, d_2 \cdots, d_n\}$,本体论 $O = \{O_1, O_2 \cdots, O_n\}$,本体论 O_i 对应于领域 d_i,由此我们建立了本体论到领域的对应关系.多领域的本体论是本体论的集合.

2.2　改进的搜索引擎　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(这里是大量网页数据)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.比如:

例　有一篇文章,如果它出现的词是比如地球、月球、太阳的词等等,那么这篇文章出现的词:“火星”,极大可能的意思应是:九大行星之一,一般不会是火中的火星的意思.

在改进的搜索引擎中,建立了从关键字到领域的索引,支持领域分类,如图 2 所示.

2.3　本体论更新器　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[5].

Nicala Guarino 对建立本体论的方法学进行了讨论[3].

本体论是世界的反映.因此它必然随着现实的发展而变化.

更新本体论的方法有 2 种方式:人工方式和系统在已有的知识上对因特网上信息学习自动更新.

搜索过程:当用户提交一个查询后,比如输入了“火星”,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,比如:这个关键是否是一个术语,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

关键字的索引表

WebLink1	HTML 文件	出现该字(词语)的片段
WebLink2	HTML 文件	出现该字(词语)的片段
…	…	…
WebLink n	HTML 文件 n	出现该字(词语)的片段 n

WebLink1 对应的领域索引表

领域 1	领域 1 领域相关度
领域 2	领域 2 领域相关度
…	…
领域 n	领域 n 领域相关度

图 2　从关键字到领域的索引表

Fig. 2　Index table from key word to domain

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.在本体论中得到这"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等.

把这些信息返回给用户,用户可以根据它关心的领域对查询结果进行过滤,这就缩小了查询的范围.也可以选择关键字的概念,由系统作概念到领域的映射.把选择的结果交给远端的基于本体论的信息检索服务器,基于本体的信息检索服务器对结果进行处理后,交给改进的搜索引擎.最后,搜索引擎把查询结果返回给用户.

3 关键字到本体论的映射

基于本体的信息检索模型中,当用户提交了一个关键字的查询后,基于本体的信息检索服务器将在本体论集中得到该关键字的信息,如:属于那个领域、同义词、定义、还有示例、语义关系等.

一个本体论可以表示为一个有向图 $G = (V, E)$,其中 V 是结点,E 是有向边,其类型有多种,比如:属于那个领域、定义、同义词、和其它概念的联系、是什么概念的子概念等.

把关键字映射到本体论集,如果本体论中出现了这个词汇,则取出领域、同义词、定义、示例等信息.

4 网页通过本体论映射到领域集合

关于这个问题,我们提出了 2 种方法,第一种方法是针对网页有关键字的情况;第二种方法针对网页没有关键字的情况.

方法 1 如果一篇文章有关键字,可以采用以下的方法[6]:

函数 Terminology(O_i) 从领域 D_i 对应的本体论中求出该领域的术语集(包括同义词);函数 Definition(O_i, Keyword) 从本体论 O_i 中求出关键字 Keyword 的定义;函数 Relation(O_i) 从本体论 O_i 中求出由概念关系构成的语义网络集.

设 $O_1, O_2, \cdots, O_v$ 分别是领域 $D_1, D_2, \cdots, D_v$ 的本体论,术语集 T_i = Terminology(O_i),其中 $(0 \leq i \leq n)$,Ks = { $\mathrm{Key}_1, \mathrm{Key}_2, \cdots, \mathrm{Key}_n$} 为被检索文档 Doc 中所给出的关键字.

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论中,则说明该文档与这一领域无关,即 $\mathrm{Ks} \cap T_i = \phi \Rightarrow \mathrm{Doc} \notin D_i$ 这里 $1 \leq i \leq n$.

经过这一步,我们可以滤掉不相关的领域,得到所有可能与该文档相关的领域,其 $\mathrm{Ds}_1, \mathrm{Ds}_2, \cdots, \mathrm{Ds}_k$,其中 $\mathrm{Ks} \cap Ts_j \neq \emptyset, S_1 \leq S_j \leq S_k$.

接下来进行近似语义网络匹配.首先求出与关键字的定义相关的术语集合.$DS = \{ dk \mid (dk \in Tsj) \wedge (dk$ 出现于 Key_i 的定义 $\mathrm{Definition}(Os_j, \mathrm{Key}_j)$ 中 $\mathrm{Key}_i \in K_S, 1 \leq i \leq m, S_1 \leq S_j \leq S_k) \}$,然后求与关键字集直接相关的术语对象集合直接相关的术语对象集合 $RO = \{ obj \mid \exists x (x \in Ks \cap (obj, x)\ \mathrm{Relation}(Os_j))), S_1 \leq S_j \leq S_k \}$.检索整个文档,统计被检索文档里出现在集合 $DS \cup Ks$ 中元素的频度 $\mathrm{freq}_i, \mathrm{freq}_i$ 体现了该文档中的术语与 Os_j 中的语义网络的近似匹配程度.我们定义 $\mathrm{Degree}(Os_j) = \mathrm{freqs}_j$,因此可以再根据 $Ds_1, Ds_2, \cdots, Ds_k$,与被检索文档的相关程度的大小 $\mathrm{Degree}(Os_j)$ 对它们进行排序.通过上述过程,可以依据本体论对文档进行按领域的分类.

方法 2 首先,对网页进行取词,得到了一个词汇集.在本体论的协助下,取出的词或概念都是具有意义的.然后,直接统计这些词汇在领域本体论出现的次数.我们定义:词汇出现的次数和这个网页的词数比称为该词的领域相关度.对领域相关度确定阈值,当领域相关度大于阈值就认为该网页与这个领域相关.于是,可得一个领域集和领域集的领域相关度.

5 结 论

本文提出了基于本体论的信息检索的方法.该方法支持领域分类,返回用户感兴趣的领域信息,提高了检索的效率.该方法处理的对象是无结构的网页,在对网页进行分类时,由于领域相关度只是一个判断是否属于该领域的值,不能说明网页一定属于这个领域.因此,返回的信息也会出现了一些与实际分类不符的情况.下一步工作就是对分类提出更好的方法来提高分类的准确性.

参考文献:

[1] BRAKE D. Lost in cyberspace[J]. New Scientist, 1997, 154(2 088):12—13.

[2] GUARINO N. Formal ontology and information systems [A]. Formal Ontology in Information System[C]. Tren-

to: IOS Press, 1998. 6—8.

[3] GUARINO N, WELTY C. A formal ontology of properties[A]. Proceedings of the ECAI-OO workshop on applications of ontologies and problem soliving methods [C], Berlin, 2000. 121—128.

[4] USCHOLD M, GRUNINGER M. Ontologies: principles methods and application[J]. Knowledge Engeering Review, 1998, 11(2): 93—155.

[5] GUARINO N, WELTY C. Ontological analysis of taxonomic relatiorr ships. Proceedings of ER-2000: the conference on conceptual modeling [EB/OL]. http://www. ladsed. pd. cnrit/infor/ontology/papers/ontology papers. html, 2003 - 04 - 11.

[6] 武成岗. 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, 38(6): 641—647.

Information retrieval model based on ontologies

GUO Xiang-wen, LIU Wei-yi, QIAN Min, ZHANG Zhong-yu
(Department of Computer Science, Yunnan University, Kunming 650091, China)

Abstract: Ontology is introduced into information retrieval. It is proposed information retrieval model based on ontologies, supporting guide for user query and selectively returning information in domain categories.

Key words: ontology; information retrieval; conceptualization; whole-document retrieval; domain classification

* *

(上接第 323 页)

[3] 乔耀军, 管克俭, 于晓映, 等. 基于 ATM 无源光网络的媒质接入控制(MAC)协议及其性能分析[J]. 高技术通讯, 2000, 5: 27—28.

[4] 原　荣. 光纤通信技术[M]. 北京: 电子工业出版社, 2000.

The analysis of access performance to an IP over WDM network using time-division switching

ZHANG Yao, YANG Ming-hua
(Department of Communication Engineering, Yunnan University, Kunming 650091, China)

Abstract: A network architecture for IP over WDM is presented. The architecture is based on time-division and wave-division multiplexing switching. The bottleneck and random access algorithm are discussed. The result appears that it is a efficient way to solve bottleneck and improve the usage of wave channels.

Key words: time division multiplexing; time slot; congestion; access control