基于本体的语义黄页检索研究

Towards Ontology - based Semantic Yellow Page Search Research

时念云 （中国石油大学计算机与通信工程学院 东营 257061）
杨 晨 滕良娟 （中国石油大学(华东)研究生院 东营 257061）

摘要:传统黄页检索采用的是基于关键词的检索,而缺乏对语义的表示、处理等能力,导致检索质量低下。基于本体的语义黄页检索是建立在语义网基础上的黄页检索技术,它能够提高检索的查全率和查准率。文章首先对语义网、本体以及语义黄页检索现有的一些应用系统进行了介绍,然后构造出了语义黄页检索查询过程模型,并针对语义黄页检索中的系统中核心的技术部分黄页本体、本体推理以及排序部分进行详细说明领域知识进行了说明。最后分析了语义黄页检索下一步发展方向。

关键词:本体 黄页 语义黄页检索

1 引言

黄页是国际上对按企业性质和产品类别编排的商业电话号码簿的固定称谓。随着互联网的发展,传统黄页被搬到网上演变为网上黄页这种新的商业模式。通过黄页,企业可以宣传自己的产品、服务等来提高企业知名度,终端用户可以去发现公司所提供的服务和销售的商品。

在传统的黄页检索方式下,大部分使用基于文档关键词的检索手段,由于信息资源缺少统一的语义描述,用户难以查找到与需求相关的资源,难以实现信息资源的语义共享,查询结果往往不能满足人们的需要,不能适应时代发展的需求。如何根据信息资源所具有的领域知识,实现信息资源的语义检索,提高数字化信息资源的利用率,是黄页检索领域所面临的挑战。

2 基于传统方法的黄页检索

黄页检索是用户获得黄页信息的重要途径,为用户提供了快速黄页信息获取的导航工具。黄页检索按照一定的策略在互联网中搜集和发现黄页信息,并对信息进行理解、提取、组织和处理,为用户提供检索服务,从而起到检索黄页信息的目的。

传统的黄页检索是基于关键词的检索,其优点是简单、快捷和容易实现,但其存在以下较突出的问题: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的表达其检索需求,因此表达困难也就导致了检索质量难近人意;②一词多义的现象的普遍存在导致了检索结果中包含大量的无效信息,使得查准率难以满足。例如用关键字"cook"搜索,计算机根本无法明确用户要查找的是厨师、烹饪信息、人名、企业名还是其他信息。③同义词查询的问题;例如查找的关键词为"计算机",一个黄页服务信息中必须包含"计算机"才能被检索出来。如果这个黄页服务信息中包含"计算机"的同义词"电脑",就会被漏检。④"词汇孤岛"问题。在人的大脑中,概念并不是独立存在的,它总是与其他概念之间存在各种各样的联系,在传统信息检索中,概念之间的联系是无法表达的。

造成这种问题的实质在于传统黄页检索缺乏知识处理能力和理解能力,对要检索的信息只是基于语法层面的上字、词的简单匹配,而缺乏对知识的表示、处理和理解等能力。把黄页检索从目前基于关键词层面提高到基于知识(或概念)层面,是解决问题的根本和关键。

语义 Web 技术为该问题的解决提供了技术上的支撑,语义 Web 技术可以用计算机可理解的方式标示信息,从已有知识中发现新信息,提供证据确认信息,利用发现的新知识,从而有利于信息共享与智能共享。在语义 web 中,本体(Ontology)是将语义网应用与信

息检索中的核心技术，它提供了语义交换的桥梁，能够在不同的智能主体之间达成有关术语概念的共识，这也恰恰解决了检索的语义实现问题。

3 语义网与本体介绍

1998 年，WWW 的发明者 Tim Berners－Lee 首次提出了语义 web 的概念。Tim Berners－Lee 给出了以下定义："语义 Web 是一个网，它包含了文档和文档的一些部分，描述了事物间的明显关系，并且包含语义信息，以利于机器的自动处理"[1]。语义 Web 建立在 XML 的表示基础之上，通过给网页加注"语义"信息，从而使得信息可以被机器"理解"，有利于搜索引擎进行复杂的信息查询。

本体原是一个哲学上的概念，用于描述事物的本质，在近几年作为信息抽象和知识描述的工具被计算机领域所采用。关于本体很多人给出了不同的理解，其中斯坦福大学的 Gruber 给出的定义得到众多认可，他认为"Ontology 是概念模型的明确规范的描述[3]"。总之，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和词汇间相互关系的明确定义。

4 基于本体知识的语义黄页检索

4.1 已改进的黄页检索系统

为了提高黄页检索的效率，一些国外的研究者已经在这方面进行了有益的尝试，建立了一些改进的黄页检索系统，如 Ontoseek[4]，是一个集中了在线黄页和产品目录的，基于内容检索的系统。该系统将一个本体驱动的内容匹配机制与一个具有中等表达能力的表示形式化系统相结合，尝试如何将本体和大辞典数据库相集成，为用户提供一个可以使用领域内任意词汇进行交互式语义查询的系统。系统提供给用户的都是自然语言接口，用户可以通过任意的自然语言术语来描述问题，并且可以使用词汇的概念关系图来描述关系。在 Ontoseek 系统中，概念关系图中的词汇和关系都是不受约束的，因而概念关系图可能是无效的，这就导致了查询结果也不准确。YPA[5] 系统是一个使用自然语言处理技术和信息查询技术来检索半结构化广告的系统。在 Ontoseek 和 YPA 系统中，任何自然语言集中的广告都可以被检索出来，但这也恰恰是系统的弱点，因为自然语言理解困难而且容易出错。

本文旨在克服这些缺点的基础上，利用本体技术，设计出基于本体的语义黄页检索模型。

4.2 语义黄页检索模型与过程

我们先给出一个整体的语义黄页的查询模型，如图 1 所示

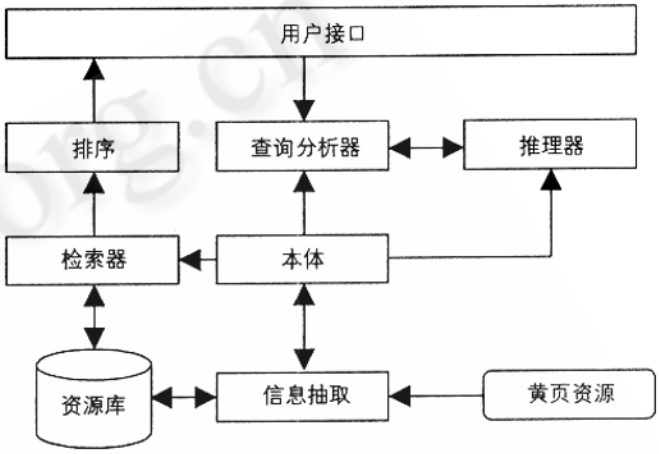

图 1 语义检索的查询框架

(1) 用户接口：用户通过用户接口提交查询请求，用户界面的主要作要是与用户交互。当用户提出检索请求时，将其提交给查询分析器。当检索信息完成后，将检索结果返回给用户。

(2) 查询分析器：分析用户的查询请求。

(3) 推理器：结合领域本体进行推理，得到扩展的查询条件。

(4) 检索器：检索器查询资源库，找出相关的所有文档，最后排序模块根据文档和用户查询请求之间的相关性对找到的文档排序，按顺序返回用户。

(5) 本体：本体库用来存储各种本体模型。同时因为本体具有知识的共享、重用、分析知识、辅助知识获取以及知识标准化的作用，查询分析器、推理器、检索器以及信息抽取也需要本体的作用才能完成。

(6) 信息抽取：把黄页资源在本体的作用下变成结构化的，机器可理解的资源库。

整个信息检索的步骤如下：用户向用户接口提交数据请求，查询分析器依据本体规范将用户请求转换为本体描述符形式的格式；分析后的用户请求提交到推理服务器，根据本体推理出相关的术语；在资源库中

搜索与相关的术语相关的资源，并把结果返回给用户。

下面我们针对系统中核心的技术部分黄页本体、本体推理以及排序部分进行详细说明。

4.3 语义黄页检索中的黄页本体

黄页主要涉及到两个个本体：行业本体、地域本体，其中行业本体描述企业信息关于行业的概念和属性，可定义三类资源对象：行业大类、行业小类、行业子类。在资源对象的基础上，可定义了四种对象属性分别描述行业大类与行业小类之间的包含关系、行业小类与行业子类之间的包含关系、行业小类与行业大类之间的从属关系、行业子类与行业小类之间的从属关系。地域本体定义类地域及地域的子类，并创建传递属性子地域属性来表达地域之间的隶属关系，那么隐含的子地域关系就可以由具有推理功能的推理机来推理得出。

我们首先建立行业本体，在这里我们建立起行业中石油产品领域的本体。如下所示：

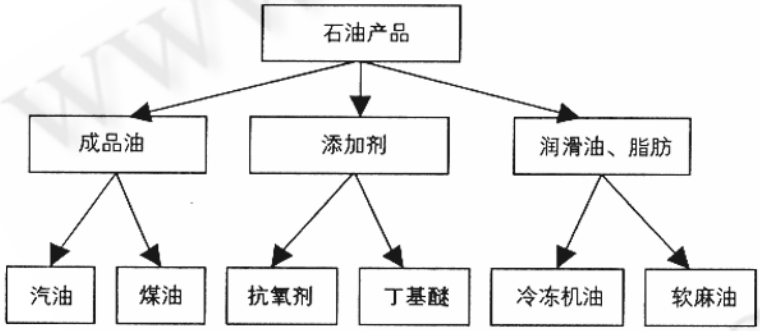

图 2　部分石油产品本体

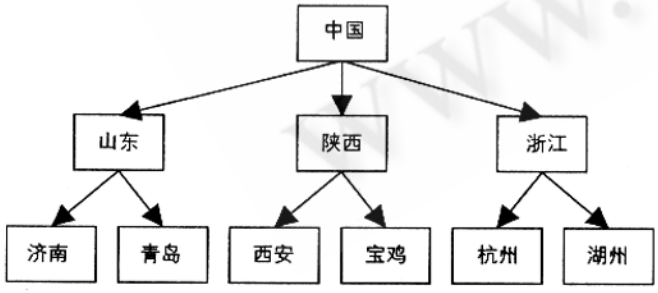

图 3　部分地域本体

已经定义了领域本体，下一步需要计算本体中的概念与一个 Web 文档之间的相关性，为了计算概念和文档之间的相关性，我们需要计算本体中词汇之间的相关性。

下面我们进行如下定义来计算本体中词汇之间的相关性：

(1) 定义 1。对于同一本体中的两个概念 C1 和 C2，若 C1⊃C2，即概念 C1 是概念 C2 的上位概念，(hypernymy)；若 C1⊂C2，即概念 C1 是概念 C2 的下位概念(hyonymy)。

(2) 定义 2。概念 C 的上位概念集合(CS：concept super)包括概念 C 的所有上位概念集合。因为 CS 中至少包含概念 C，因此 CS≠Ø。对于同一本体 O 中的 2 个概念 C_i 和 C_j 之间的语义相似度 $\theta_{C_iC_j} \in [0,1]$，$C_s$ 和 $C_{上}$ 位概念集合 C_iS 和 C_j 上位概念集合 C_jS 的交集与 C_i 上位概念集合 C_iS 和 C_j 上位概念集合 C_jS 的并集的比值为：$\forall C_i, C_j \in O$，$\theta_{C_iC_j}$ =

$$\begin{cases} 1 & C_iS \equiv C_jS \\ \dfrac{C_iS \cap C_jS}{C_iS \cup C_jS} & \varnothing \subset CS_i \cap CS_j \subset 0 \\ 0 & CS_i \cap CS_j \equiv \varnothing \end{cases} \quad \text{(公式 1)。}$$

例如石油产品本体如图 2 中，可选择特征词汇：石油产品、添加剂、丁基醚、软麻油、烷基苯料，其中烷基苯料不是本体中出现的词汇，据公式 1，可计算出 $\theta_{石油产品,添加剂}=1/2$，$\theta_{石油产品,丁基醚}=1/3$，$\theta_{石油产品,软麻油}=1/3$，$\theta_{添加济,丁基醚}=2/3$，$\theta_{软麻油,丁基醚}=1/5$，$\theta_{软麻油,添加浏}=1/4$，其具体结果如图 4 所示。而地域本体如图 3 中，可选择特征词汇：中国、陕西、西安、宝鸡、杭州。

	石油产品	添加济	丁基醚	软麻油	烷基苯料
石油产品	1	1/2	1/3	1/3	0
添加浏	1/2	1	2/3	1/4	0
丁基醚	1/3	2/3	1	1/5	0
软麻油	1/3	1/4	1/5	1	0
烷基苯料	0	0	0	0	1

图 4

	中国	陕西	西安	宝鸡	杭州
中国	1	1/2	1/3	1/3	1/3
陕西	1/2	1	2/3	2/3	1/4
西安	1/3	2/3	1	1/2	1/5
宝鸡	1/3	2/3	1/2	1	1/5
杭州	1/3	1/4	1/5	1/5	1

图 5

图4中,其中"添加剂"、"丁基醚"和"石油产品"、"添加剂",它们具有相同深度的上下位关系.而"添加剂"、"丁基醚"概念层次虽低,但它的语义相似度更高,这说明相同距离的层次关系,语义相似度随着概念层次的递减而增高。因为层次越低,概念间密度越小,带来差异的可能性也就越小。

当建立起本体中词汇之间的相关性后,本体中的概念与一个 Web 文档中特征词之间的相关性就可以明白的看出,例如文档中的特征词是石油产品、添加剂、丁基醚、软麻油、烷基苯料,本体中的概念丁基醚与这几个特征词的相关性就分别为:1/3、2/3、1、1/5、0。

4.4 语义黄页检索中的本体推理以及查询过程

在本体构建完成之后,下一步工作就是对本体进行推理,将查询的条件进行推广。在本系统,推理包括两部分:一是地域本体推理和行业本体推理,二是在上述推理的基础上进行两个本体的混合推理。第一部分推理占有较大的推理时间,而且推理的结果占用存储空间比较少,在第二部分推理时查询花费的代价比较小,因此我们将这部分进行离线推理,推理结果放入相应的数据表中。第二部分推理是对相应时间短,便于在线完成,同时如果离线完成则需要的占用较多的系统存储空间,增加检索时的查询花费。因此我们采用在线推理和离线推理想结合的方法,在离线推理部分完成地址本体推理和行业本体推理,在线部分完成两个本体的混合推理。

关于混和推理,采用两个本体对应相关性的乘积来得出最终的权值结果。

表1是地区"西安"+行业"丁基醚"推理的示例。

表 1 基于领域本体的查询条件语义推理示例

地区	行业	混和推理权值计算	权值
西安	丁基醚	1×1	1.0
西安	添加剂	1×2/3	0.67
西安	石油产品	1×1/3	0.33
西安	软麻油	1×1/5	0.2
陕西	丁基醚	2/3×1	0.67
陕西	添加剂	2/3×2/3	0.44
陕西	石油产品	2/3×1/3	0.22
陕西	软麻油	2/3×1/5	0.13
…	…	…	…

4.5 语义黄页检索中的结果控制以及排序

在检索时,我们需要首先设定一个权值的阈值来控制结果的输出,例如表1中,如果我们设定权值的阈值为0.3,则西安&软麻油、陕西&石油产品、陕西&软麻油的结果将不会出现在检索结果中。

由于信息的快速增长,往往导致信息检索返回的结果集过于庞大,超出了信息检索者所能处理的范围。因此,就提出了对检索结果进行有效排序的要求。检索结果排序的关键在于如何衡量各检索结果相对用户的重要性(相关性),本系统选用本体推理的加权值作为排列的依据,加权值越大,表示扩展后的查询条件和原始查询条件越接近,与用户的查询本意越接近,因此查询结果按照加权值降序。

5 结束语

基于本体的语义黄页检索要求特定领域内,能够正确分析并解释用户的查询条件以及资源中所包含术语之间的上下继承关系。与基于关键词检索的传统黄页检索相比较,语义黄页检索能够更好的理解用户的查询请求,从整体上分析每个资源的内容,从而得到较好的检索结果。语义黄页检索克服了关键词检索词局限于形式的固有缺陷,提高了用户的满意度,减少了不相关的返回结果,能够提高检索的精度和覆盖率。

参考文献

1 Tim Burners – Lee, James Hendler and Ora Lassila, The Semantic Web[J], Scientific American, 2001.

2 Guha, R. V., McCool, R., Miller, E.: Semantic search[J], In Proc. of the 12th Intl. WWW 2003, 700 – 709.

3 Gruber T R. A Translation Approach to Portable Ontology Specifications [J], Knowledge Acquisition, 1993;5: 199 – 220.

4 N. Guarino, C. Masolo, G. Vetere. OntoSeek: Content – Based Access to the Web. IEEE Intelligent Systems, 1999, May/June:70 – 80.

5 A. De Roeck, U. Kruschwitz, P. Neal, P. Scott, S. Steel, R. Turner, and N. Webb, YPA – an Intelligent directory enquiry assistant, BT Technology Journal[J], 1998, 16(3): 145 – 155.