基于领域本体的 Web 信息检索实现机制研究*)

宋二伟　刘宗田　徐力斌　陈　光
(上海大学计算机学院　上海 200072)

摘　要　传统的基于关键词的信息检索方式，往往难以用一个或几个"关键词"表达用户真正的检索要求。针对此问题，本文提出了一种基于领域本体的信息检索机制，将用户输入的关键词，用领域本体进行理解、扩充，然后基于领域本体中概念的相关度，求出扩充后每个关键词的权值，并将之用于随后的信息检索。实验证明，本方法在基本维持查准率的同时显著提高了信息检索的查全率。

关键词　本体，概念相关度，查询扩展，关键词权值，Jena 包

The Realization Mechanism of the Web Information Retrieval Based on the Domain Ontology

SONG Er-Wei　LIU Zong-Tian　XU Li-Bin　CHEN Guang
(Department of Computer, Shanghai University, Shanghai 200072)

Abstract　The traditional way of information retrieval is to use the keywords to express the information requirement. But it is hard to faithfully express the user's retrieval purpose. So this paper brings forward an information retrieval mechanism based on the domain ontology. We use the domain ontology to comprehend and extend the original keywords inputted by the user, then assign a weighted value to each keyword based on the concept-correlation value of the domain ontology and put it into use of the later information retrieval work. The experiment results prove that this way can obviously improve the recall ratio while basically maintaining the precision ratio.

Keywords　Ontology, Concept-correlation value, Query extension, Weighted value of the keywords, Jena development kits

随着互联网的迅速发展，网上信息资源越来越丰富，网络已经成为一个全球最大的信息库。同时随着计算机和互联网在人类社会和生活各个方面的大力普及和广泛应用，如何在浩如烟海的信息海洋中找到自己所需的信息就成了很多人关注的焦点。因此，大力发展 Web 信息检索技术就成为时代发展的迫切需要。

传统的 Web 信息检索方法主要是基于关键词的全文检索，即把用户的检索请求以关键词的形式按某种检索模型(如布尔模型、经典向量模型[4]、经典概率模型[5]、推理网络模型[6]或子图匹配模型等)和 Web 文档进行匹配。这种检索方法的主要缺点是：对于用户真正的检索需求，有时很难用合适的关键词表达出来，而又由于不同用户对关键词的理解不同和不同领域关键词的含义不同，这样都使得检索结果不能准确、全面地反映用户的检索请求。

传统 Web 信息检索方法困难的实质在于：只是对用户输入的关键词和 Web 文档进行机械的匹配，而没有去理解用户的查询意图。于是，有必要去理解用户的查询意图，然后基于更全面的用户需求来对 Web 文档进行处理，进而得到更加合理的检索结果。本体(Ontology)具有良好的概念层次结构并且支持逻辑推理，因此我们可以用其对用户检索请求进行扩展。以往对用户查询扩展的研究，大多是简单的同义词扩展或单纯的上下位扩展[8,9]，而没有考虑扩展后各关键词的权值。文[12]提出了一种经本体扩展后所得关键词的权值计算和传递方法，这种方法只是机械的基于本体的层次结构而没有考虑各关键词在语料库中的实际重要程度。文[10]提出了一种 Term Similarity Tree 模型来对用户查询进行扩展，但 Term Similarity Tree 的构建和生长都很复杂。本文基于构建好的领域本体，用统计法计算领域本体中概念之间的相关度，并用改进的向量模型对用户查询请求和 Web 文档进行匹配。具体来说，第 1 节介绍了本体论的基本思想，第 2 节描述了本体中概念相关度的计算方法，第 3 节讨论了用户查询扩展模式和关键词权值的计算方法，第 4 节讨论了信息检索的具体实现方法，第 5 节描述实验系统并分析试验结果，最后总结全文并进行展望。

1　本体论

本体(Ontology)的概念最初起源于哲学领域，研究客观事物存在的本质。后来人们将本体的概念和方法应用于计算机领域，用于知识表示、知识共享和知识重用。

本体可定义为：本体是对共享概念模型的明确的形式化规范说明。要建立一个能够涵盖所有领域的通用本体是不可行的，因此比较现实的方法是抽象出某个领域内共同认可的一些概念以及概念之间的关系，用计算机可以理解的方式来构建领域本体，利用它去解决该领域的信息检索问题。

本体描述语言有多种，有代表性的有：SHOE、XOL、RDF、RDFS、OIL、DAML＋OIL 和 OWL。这些语言除了 SHOE 的语法基于 HTML 之外，其他语言的语法都基于 Web 上信息交换的标准语言——XML。

*)基金项目：本文受国家自然科学基金(60575035)资助。宋二伟　硕士，研究方向为信息检索、自然语言处理；刘宗田　博士生导师，教授，主要研究领域为人工智能和软件工程等；徐力斌　硕士，从事基于语义的知识管理系统的研究；陈　光　硕士，研究方向为 Web 数据挖掘和信息分类。

RDF(Resource Description Framework)是国际标准化组织 W3C 于 1997 年提出的一种知识表示模型,它以 XML 语法为基础,拥有较强的描述能力和一定的推理能力。

本文所用的旅行领域的本体就是用 RDF 描述的,其代码片断如下所示:

```
〈a:Class
 rdf:about="http://www.ina.fr#FiveStarHotel"〉
〈a:subClassOf
 rdf:resource="http://www.ina.fr#Hotel"/〉
〈/a:Class〉
〈a:Class
 rdf:about="http://www.ina.fr#Taxi"〉
〈a:subClassOf
 rdf:resource="http://www.ina.fr#Car"/〉
〈/a:Class〉
```

2 概念相关度计算

概念的相关度,反映的是两个概念相互关联的程度。概念相关度计算的基本思想是:如果两个概念的相关程度比较高,那么它们所在的上下文环境应该比较相似,所以,概念相关度可以用这两个概念在同一语境中共现的可能性来衡量。

假设本体中共有 n 个概念 $C_1,C_2,\cdots,C_n$。假设 C_i 在 C_j 的 *Web* 文档语料库上下文环境中出现的频率为 f_{ij},所有概念在其他概念的上下文环境中出现的频率构成表 1。

表 1

	C_1	C_2	……	C_j	……	C_n
C_1	f_{11}	f_{12}	……	f_{1j}	……	f_{1n}
C_2	f_{21}	f_{21}	……	f_{2j}	……	f_{2n}
……	……	……	……	……	……	……
C_i	f_{i1}	f_{i2}	……	f_{ij}	……	f_{in}
C_n	f_{n1}	f_{n2}	……	f_{nj}	……	f_{nn}

这样,对每个概念都可以得到用这些频率构成的向量,如 C_i 对应的向量为:

$W_i=(f_{i1},f_{i2},\cdots,f_{ij},\cdots,f_{in})$,则任意两个概念 C_i,C_j 之间的相关度可用余弦法计算:

$$Sim(C_i,C_j)=\frac{W_i\cdot W_j}{|W_i|\cdot|W_j|}$$

$$=\frac{\sum_{k=1}^{n}f_{ik}\cdot f_{jk}}{\sum_{k=1}^{n}(f_{ik})^2\cdot\sum_{k=1}^{n}(f_{jk})^2}$$

3 用户查询扩展和关键词权值计算

用户的检索请求往往用一个或几个关键词来表示,而这往往很难忠实表达用户的检索请求,本文用领域本体对用户的检索请求进行理解,并利用其良好的概念层次结构进行一些逻辑推理,这样除了用户提供的初始查询条件外,还得出和初始查询有泛化和特化关系的一些查询关键词,并利用上一节介绍的概念相关度算法求出扩展出的每个关键词的权值。

考虑到检索系统的实用性和简易性,本文只对用户检索请求进行三种扩展:上位扩展、下位扩展和平行扩展。本文假设用户对相关查询领域有一定了解,而且只针对英文 Web 文档的进行检索。如果用户提交的查询关键词没有在本体中出现的话,则先用英文同义词词典 WordNet_hyponyms 推导出其同义词,再用领域本体进行扩展。下面详细介绍一下相关定义。

定义 1(用户查询) 是指由用户提供的初始关键词构成的集合。可表示为:

$Q=\{C_1,C_2,\cdots,C_i,\cdots,C_j,\cdots,C_m\}$

$\forall i,j:1\leqslant i,j\leqslant m$,

C_i、C_j 为用户输入查询关键词且 $C_i\neq C_j$

定义 2(带权查询) 是指由初始关键词或扩展后的关键词与其对应的权值构成的二元组的集合。用此权值来表示相应的关键词在查询中的重要程度。可表示为:

$WQ=\{(C_1,W_1),(C_2,W_2),\cdots\cdots,(C_m,W_m)\}$

定义 3(上位查询扩展) 是指对用户查询中的初始关键词利用领域本体进行泛化,求出其上位关键词及其对应的权值。可表示为:

$$WQ_t=\left\{\langle C_t,W_t\rangle\ \middle|\ \begin{array}{l}\forall C_j\in Q:\\ C_j\in subclassof(C_t)\\ W_t=Sim(C_j,C_t)\end{array}\right\}$$

定义 4(下位查询扩展) 是指对用户查询中的初始关键词利用领域本体进行特化,求出其下位关键词及其对应的权值。可表示为:

$$WQ_d=\left\{\langle C_d,W_d\rangle\ \middle|\ \begin{array}{l}\forall C_i\in Q:\\ C_d\in subclassof(C_i)\\ W_d=1\end{array}\right\}$$

定义 5(平行查询扩展) 是指对用户查询中的初始关键词利用领域本体进行平行扩展,求出其兄弟关键词及其对应的权值。可表示为:

$$WQ_b=\left\{\langle C_b,W_b\rangle\ \middle|\ \begin{array}{l}\forall C_j\in Q:\\ C_t\in WQ_t\\ C_b\in subclassof(C_t)\\ C_b\neq C_j\\ W_i=Sim(C_b,C_j)\end{array}\right\}$$

定义 6(带权用户查询向量) 一般来说,用户输入的初始关键词代表了其核心查询意图;而扩展后的下位关键词由于是对初始关键词的特化,因而也是对用户核心查询意图的细化。因此,对初始关键词和下位关键词都赋予最高权值 $W=1$。

对于上位关键词和平行关键词,由于分别代表的是对初始关键词的泛化和平级扩展关系,因此,可用第 2 节求得的和初始关键词的概念相关度来表示其权值 $W=Sim(C,C')$。

至此,用户输入的初始关键词以及扩展后的关键词都有了其对应的权值,带权用户查询向量就是由所有关键词的权值构成的向量。可表示为:

$WQ=\langle W_1,W_2,\cdots,W_i,\cdots,W_j,\cdots,W_t\rangle$

4 Web 文档信息检索的实现方法

假设经领域本体扩展后得到的全部查询关键词组成的集合 Q_e 可表示为:

$Q_e=\{C_1,C_2,\cdots,C_i,\cdots,C_j,\cdots,C_t\}$

对任一 $C_i\in Q_e$,在英文同义词词典 WordNet_hyponyms 中推导出其同义词,C_i 与其同义词构成的集合记为$\{C_i\}$。如此可得出 Q_e 中的所有关键词对应的同义词集合:$\{C_1\}$,$\{C_2\}$,…,$\{C_t\}$。

假设我们的 Web 文档集 D 包含 n 篇文档,记为 $D=\{d_1,d_2,\cdots,d_j,\cdots,d_n\}$,$\{C_i\}$的所有术语出现在文档 d_j 中的频率为 f_{ij},如表 2 所示。

表 2

	d_1	d_2	… …	d_j	… …	d_n
$\{C_1\}$	f_{11}	f_{12}	… …	f_{1j}	… …	f_{1n}
$\{C_2\}$	f_{21}	f_{21}	… …	f_{2j}	… …	f_{2n}
… …	… …	… …	… …	… …	… …	… …
$\{C_i\}$	f_{i1}	f_{i2}	… …	f_{ij}	… …	f_{in}
… …	… …	… …	… …	… …	… …	… …
$\{C_t\}$	f_{t1}	f_{t2}	… …	f_{tj}	… …	f_{tn}

基于 tf-idf 算法，我们按如下方式计算$\{C_i\}$在文档 d_j 中的权重 W_{ij}：

$$tf_{ij}=\frac{1+\log(f_{ij})}{1+\log(f_{average})} \tag{1}$$

$$idf_i=1+\log\left(\frac{n}{n_i}\right) \tag{2}$$

$$W_{ij}=tf_{ij}\times idf_i \tag{3}$$

此处，n 表示 Web 文档总数，n_i 表示含有$\{C_i\}$中的关键词的文档数目。在公式(1)中，$f_{average}=\frac{1}{k}\sum_{k=1}^{t}f_{kj}$，是文档 d_j 中各术语频率的平均值，因为 tf_{ij} 不完全是$\{C_i\}$的频率 f_{ij} 对 max $\{f_1,f_{2j},\cdots,f_{ij},\cdots,f_{tj}\}$的相对值，而是对文档 d_j 中的所有术语频率值的相对值。

这样，对文档集中的每篇文档，我们都可以计算出任一$\{C_i\}$在文档中的权值，从而文档 d_j 可以由各权值组成的向量来表示：$W_j=\langle W_{1j},W_{2j},\cdots,W_{ij},\cdots,W_{tj}\rangle$

我们在第 3 节已经得出了经领域本体扩展后的用户查询向量 WQ，文档集中各文档的排序计算可由以下公式来完成：

$$Sim(W_j,WQ)=\frac{W_j\cdot WQ}{|W_j|\cdot|WQ|}$$

5 实验系统及结果分析

实验采用的领域本体是一个由 RDF 语言描述的旅行领域的本体，描述了旅行领域内的相关概念以及概念之间的关系，然后用惠普实验室提供的开源 Semantic Web Framework——Jena 对此领域本体进行相应的推理查询操作。由于本实验所用的领域本体比较简单，只描述了概念或概念属性间的层次关系，因此如果用户提供的查询关键词没有在领域本体中出现，则我们先用 jena 包对英文同义词本体 WordNet_ hyponyms 进行推理查询，找出其同义词再用 jena 到领域本体中进行扩展。

图 1

信息检索模型实用性的验证一般是使用公认的测试参考文档集（如 TREC、CACM、CF 等）来进行试验，比较不同模型的查准率和查全率。但由于这些测试文档集涵盖的领域太多，很难为之建立一个对应的本体，因此本实验只针对旅行领域，用网络爬虫爬来大约 1000 篇旅行领域的相关 Web 文档作为测试文档集。

为测试本文所述的信息检索方法的效果，基于 lucene 框架，专门设计了一个搜索引擎 OTSE（Ontology-based Traveling Search Engine），其界面如下图 1 所示。

如果用关键词“car”进行搜索，则其相关概念“rentalcar、texi、motorbike”等都会由领域本体扩展得到，扩展后的用户查询在搜索引擎中的检索结果界面如图 2 所示。

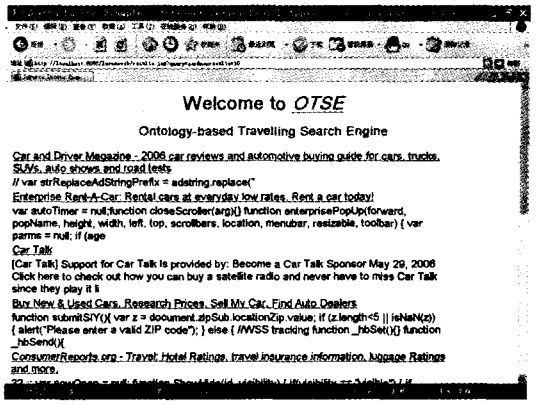

图 2

信息检索算法的评价指标主要有：查准率 P(Precision)和查全率 R(Recall)，其定义分别如下：

$$P=\frac{R_a}{C};R=\frac{R_a}{A}$$

其中，C 为满足一定条件检索出的文档数目，A 为所有相关的文档数目，R_a 为检索出的相关文档数目。

我们选 3 个查询关键词，分别用它们做 3 次检索，用本体进行查询扩展和不用本体进行查询扩展时的查准率和查全率如表 3 所示。

表 3

	hotel	car	price
不扩展	76.4%	68.6%	72.5%
	55.8%	60.5%	67.2%
扩展	73.8%	64.3%	75.7%
	69.3%	77.9%	68.2%

由表 3 可见，用领域本体对用户查询进行扩展后，在基本维持查准率的同时，查全率有了明显提升。

结论 本文用领域本体对用户的初始查询关键词进行扩展，找出与其相关的关键词，并在概念相似度的基础上赋予每个关键词相应的权值，这样在以后的检索步骤中，使用户得到的查询结果更加全面，同时又不失查询准确率。

由于本文用的领域本体还比较简单，只能对一些比较简单的属性关系、概念层次关系等进行推理，因而导致了对用户查询意图的理解和扩展能力有限，因此构建一个结构清晰功能强大的领域本体是信息检索的重要基础；另外，本文所采用的倒排文档索引结构是基于文档的全文分析的，如果在压缩文本的基础上直接进行索引和检索工作会带来更好的时间性能和更小的空间开销。以上两点是本文以后努力改进的方向。

（下转第 110 页）

者与接受者之间信息的私密性。数字签名技术提供了信息的不可抵赖性与完整性。SAML 声明包括 XML 签名，利用发送方的私钥对声明的所有内容进行数字签名。接受方获得一个声明，服务程序就会验证该声明：①声明是否被签名；②签名是否正确；③验证声明的时间戳是否有效；④签名者的身份是否可信。

4 语义策略规则的需求

在 Web 服务环境中，不可能准确地预测到何种用户在何时要访问服务，以及访问何种服务，基于策略的控制很难利用关于主体、行为、事件的精确知识。为了解决这样环境对访问控制的需求，最近的研究提出采用丰富的语义来表达策略和域知识，即利用语义 Web 语言来定义和管理策略[10]，这种方法也称为基于 Ontology 的策略定义方法。基于 Ontology 的策略定义方法能带来诸多好处，事实上，基于规则的策略定义方法，是将策略编码成逻辑编程(Logic Programming)规则。基于规则的策略语言有非常强的表达能力和较高的执行效率，但会造成推理的不可判定性(可判定性是指所有的计算都在有限的时间内完成)。基于 Ontology 的策略定义方法，主要依靠描述逻辑(Description Logic)语言的表达特点。描述逻辑是一种知识表示工具，它首先定义应用领域的相关概念，然后利用这些概念去表示应用领域中的关系或属性，达到表示出应用领域中个体和对象的目的。描述逻辑有着形式化的、基于逻辑的语义规范，可以从明显的表达推理出隐含的内容。另外，语义 Web 语言能够保证在事先互不知晓的实体之间对彼此的概念、能力、行为等有共同的理解，保证了互操作性；在抽象的高层建模的策略，可简化它们的描述，改进系统的分析能力；语义 Web 语言也包括表达力较强的查询和自动推理能力。

目前，已设计了几种基于 Ontology 的语义策略规范语言，如 KAoS[11]，Rei[12]、SWSL[13]等，但由于它们是专有的，不是通用标准语言，在实际应用中会存在互操作问题。XACML 是一个基于属性的策略语言，但并不支持语义。如何在 XACML 中注入语义，是一个值得研究的问题，也是我们今后研究的一个方向。

结束语 基于策略的访问控制 PBAC 已受到广泛的重视，并逐步应用到实际中。但在目前的实际应用中，PBAC 系统大多采用自己的方式来描述访问控制策略，这不利于策略的共享和系统间的互操作。在本文中，我们基于标准的访问控制语言 XACML 和特别适用于 Web 服务环境的访问控制模型 ABAC，提出了一种基于策略的 Web 服务访问控制框架，具有较好的互操作性、灵活性和规模性。为了更好地应用 PBAC，增加策略的语义性，提高语义互操作性，我们将在这方面做进一步的研究。

参 考 文 献

1 World Wide Web Consortium. Web service. http://www.w3.org/2002/ws

2 Juliano F S, Luciano P G, Marinho P B, et al. Policy-Based Access Control in Peer-to-Peer Grid Systems. In: Proceedings of The 6th IEEE/ACM International Workshop on Grid Computing, 2005

3 Entegrity Solutions Whitepaper. AssureAccess Policy-Based Access Control: A definition, and how it extends Role-Based Access Control. http://www.entegrity.com/products/whitepaperall.shtml. 2005

4 沈海波，洪帆. 面向 Web 服务的基于属性的访问控制研究. 计算机科学，2006，33(4)：92～96

5 OASIS Standard. eXtensible Access Control Markup Language (XACML) Version 1.0. February 2003. http://www.oasis-open.org/committees/xacml

6 Ferraiolo D F, Sandhu R, Gavrila S, et al. Proposed NIST standard for role-based access control. ACM Transactions on Information and System Security (TISSEC), 2001, 4(3)

7 OASIS Standard. Security Assertion Markup Language (SAML) V1.1, October, 2003. http://www.oasis-open.org/committees/security/docs/cs-sstc-core-01.pdf

8 Sun's XACML Implementation Programmer's Guide. http://sunxacml.sourceforge.net/guide.html, June 2006

9 Microsoft Knowledge Base. Description of the Secure Sockets Layer (SSL) Handshake, 2003. http://support.microsoft.com/default.aspx?scid=kb.

10 Nejdl W, Olmedilla D, Winslett M, et al. Ontology-based policy specification and management. In: 2nd European Semantic Web Conference (ESWC), volume 3532 of Lecture Notes in Computer Science, Heraklion, Crete, Greece, Springer, May 2005. 290～302

11 Uszok, Bradshaw J, Jeffers R, et al. KAoS Policy and Domain Services: Toward a Description-Logic Approach to Policy Representation, Deconfliction, and Enforcement. In: Proceedings of the 4th IEEE International Workshop on Policies for Distributed Systems and Network, Italy, June 2003. 93～98

12 Kagai T, Finin A, Joshi A. Policy Based Approach to Security on the Semantic Web. In: Proceedings of the 2nd International Semantic Web Conference (ISWC), LNCS, Springer, 2003, 2870

13 SWRL: A Semantic Web Rule Language Combining OWL and RuleML. Draft Version 0.7, 21 December 2004. http://www.daml.org/rules/proposal/

(上接第 106 页)

参 考 文 献

1 Greengrass Ed. Information Retrieval: A Survey. http://www.nlp.org.cn/categories/default.php?cat_id=17

2 Craven M, DiPasquo D, Freitag D, et al. Learning to Construct Knowledge Bases from the World Wide Web. Artificial Intelligence Elsevier, August 1999

3 Winkels R, Bosscher D, Boer A, Hoekstra R. Extended conceptual retrieval. Legal Knowledge and Information Systems. Jurix 2000. In: The Thirteenth Annual Conference. Amesterdam: IPS Press, 2000. 85～97

4 Raghavan V V, Wong S K M. A critical analysis of vector space model for inormation retrieval. Journal of the American Society for Information Sciences, 1986, 37(5): 279～287

5 Van Rijsbergen C J. Information Retrieval. Butterworths, 1979

6 Turtle H, Croft W B. Evaluaton of an inference network-based retrieval model. ACM Transantions on Information Systems, 1991, 9(3): 187～222

7 Swoogle. Semantic Web Search Engine. http://swoogle.umbc.edu/

8 Mandala R, Tokuanga T, Tanaka H. Combining multiple evidence from different types of thesaurus for query expansion. SIGIR, 1999

9 Miller G A, et al. Introduction to WordNet: an on-line lexical database. International Journal of Lexicography, 1990, 3(4): 235～312

10 Jin Qianli, Zhao Jun, Xu Bo. Query Expantion Based on Term Similarity Tree Model. IEEE, 2003. 400～406

11 Baeza-Yates R, Ribeiro-Netod B 等著. 现代信息检索. 王知津，贾福新，等译. 机械工业出版社，2005. 165～285

12 李飞，高济，刘柏嵩，周明健. 知识管理中语义与关键字相结合的检索方法. 计算机辅助设计与图形学学报，2004，16(12)