一种基于知识库的语义检索系统模型

马中杰,郑 诚,苏 喻

(安徽大学 计算机科学与技术学院,安徽 合肥 230039)

摘 要:讲述了目前检索系统存在的不足以及产生这些不足的原因,介绍了本体的概念及其在语义检索领域中的作用。在此基础上提出了一种基于知识库的语义检索系统模型,并对该模型的实现原理和关键技术进行了详细的阐述。实验结果表明,相对于传统的方法,该方法能大幅提高用户检索的查全率和查准率。

关键词:本体;知识库;语义检索

中图分类号:TP391.3 **文献标识码**:A **文章编号**:1674-7720(2010)20-0070-04

A semantic retrieval system model based on knwoledge base

MA Zhong Jie, ZHENG Cheng, SU Yu

(Department of Computer Science and Technology, Anhui University, Hefei 230039, China)

Abstract: First of all, this paper describes the deficiencies of the current system, as well as the reasons for these deficiencies. And then the concept of ontology and its role in the field of semantic retrieval were introduced. On this basis, a semantic retrieval system, based on knowledge base, was proposed, and in which the implementation principle and key techiques were described in detail. Finally, the experimental results show that the method can greatly enhance user retrieval precision and recall rate compared to traditional methods.

Key words: ontology; knowledge base; semantic retrieval

目前检索系统主要是基于关键字的全文匹配或者是按主题进行分类。但是,前者仅仅是进行字符串的匹配,不能对信息的语义进行揭示;而主题分类对信息资源揭示的效率较低、深度有限。由于以上缺陷,人们致力于寻求一种新的检索模式。本体作为一种能够在语义和知识层次上描述信息系统的概念模型建模工具,具有良好的概念层次结构和对逻辑推理的支持,于是人们便开始了基于本体的语义检索的尝试,试图利用本体的语义关系来提高检索系统的语义智能,从而使信息检索从目前基于关键字的层面提高到基于知识的层面。

根据处理网络文档方式的不同,基于本体的语义检索分为基于知识库的语义检索和基于语义网文档的语义检索。前一种指尽可能维持现有文档的内容形式,利用知识表示的强大功能来建立庞大的知识库。而后一种基于语义网,语义网文档是包含语义信息的文档,能被软件代理直接访问,这种检索方式代表着互联网的发展方向。但是,要想以可支付的代价将现有网络文档转换成语义网文档是不太现实的,所以本文主要研究基于知识库的语义检索。

1 本体论概述

1.1 本体的起源和定义

本体原本是哲学领域的一个概念,后来该概念被信息系统、知识系统等所借用,并迅速成为人们的研究热点。有关本体概念,目前比较公认的定义为“本体是共享概念模型的明确的形式化规范说明”。该定义包含了4层含义:“概念模型”指通过抽象出客观世界中一些现象的相关概念而得到的模型;“明确”指所使用的概念及其约束都有明确的定义;“形式化”指能被计算机所处理;“共享”指本体中体现的是共同认可的知识。

1.2 本体的描述语言

为了让计算机能够对信息的语义进行处理,需要一定的编码语言(例如RDF等)来表达本体的体系结构。资源描述框架RDF(Resource Description Framework)定义了一个基本的数据模型,该模型包括了三种对象类型:资源(resources)、属性(properties)、声明(statements)。资源可以是网页、多媒体等,通常用URI来命名;属性用来描述

资源的一个特定方面、特征等；一个 RDF 的声明就是一个资源和一个属性加上这个属性的取值所形成的集合。一个声明由以下三部分组成：主语(subject)、谓语(predicate)、宾语(object)。

1.3 本体在语义检索中的作用

本体在语义检索中的作用可概括为以下几点：

(1)本体为语义标注和扩展提供了标准的词汇库；

(2)检索中所进行的推理工作必须在本体中进行；

(3)本体可以明确领域假设，使领域公理得到明确描述而达成共识。

2 系统的基本框架结构

本文提出的模型其基本设计思想如下：首先在领域专家的参与下建立相应领域的本体，然后把收集的数据信息参照已建立的本体，按规定的格式存储在知识库中，当用户检索时，按照本体把查询请求转换成规定的格式，并从知识库中匹配出符合条件的文档集，排序后返回给用户。该模型的主要组成部分有用户界面、领域本体、文档集、知识库等。其结构关系如图 1 所示。其实整个系统可划分成虚线所示的三部分：基于本体的信息提取和语义标注、基于知识库的查询请求处理和检索模块以及对检索结果进行排序。

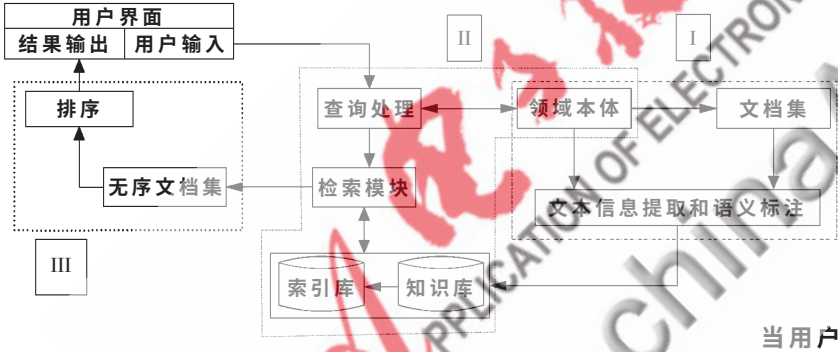

图 1 基于知识库的语义检索系统模型

2.1 知识库

在现有网络下实现真正意义上的语义检索，建立知识库是必需的。知识库是搜索代理进行推理和知识积累的关键。通常某个领域的本体提供了该领域相应的术语和概念，而知识库就是利用这些术语和概念来表达现实或者虚拟世界的正确知识。例如一个医学本体可能包含有“高血压”、“糖尿病”等术语的定义，但它并不包含某一个具体病人的诊断结果，而这正好是知识库所要表达的内容。例如王小二患有高血压，李四患有糖尿病等，在这个例子中高血压、糖尿病就是本体的概念，而各个病人的实例(王小二、李四)及其病症的描述就是知识库要表达的内容。

2.2 基于本体的信息提取和语义标注

在信息检索中为了提高检索效率，必须对网络上所存在的资源进行预处理。信息提取就是首先对文档集中的每篇文档进行词汇分析，利用禁用词表去掉文献中的虚词以及对检索作用不大的词、数字、字母、标点符号等，仅保留具有实际意义的名词、动词等，然后确定索引元素，并在本体中获得能够正确表达文档内容的概念性词或词组。

语义检索即在一个知识库中做逻辑判断并推理，检索的结果往往都是知识库中的元组，但用户需要的是提供相关文档，这就需要通过明确、无隐蔽的标注方式，把知识库中的概念、实例或者关系与那些描述它们的文档关联起来，这就是语义标注的功能。通常使用文档—实例关联表来存储文档和实例间的映射关系，这种关联表也称索引库，有了索引库之后就可以通过查询接口返回的元组实例获得相应的文档链接。该部分的流程图如图 2 所示。

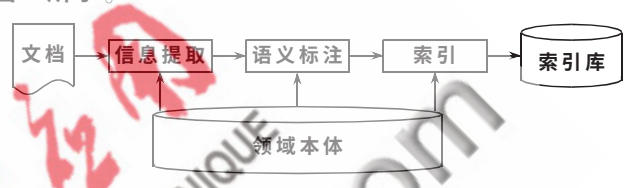

图 2 信息提取与语义标注的流程图

为了量化实例(instance)与文档之间相关性的大小，有时还需要确定标注的权重。现在通用的算法是 TF-IDF 算法，其计算公式为：

$$w_{ij}=(freq_{ij}/\max\ freq_{ij})\times\lg(D/D_w)$$

其中 w_{ij} 表示实例 I_i 在文档 D_j 中的权重。D 是全部文档数，D_w 则是包含特征词的文档数。$freq_{ij}$ 表示实例 I_i 对应的标签在文档 D_j 中出现的频率，$\max\ freq_{ij}$ 表示在文档 D_j 中出现次数最多的实例的频率。

2.3 基于知识库的查询请求处理和信息检索模块

当用户输入检索词后，查询请求处理模块对查询语句进行分析，从中提取出能正确表达查询语义的概念性词或词组。然后将其带到本体中查找相应的概念，并对概念进行语义化处理，得到一个检索式集合，再由检索代理从知识库中匹配出符合条件的元组集[1]。该部分主要包括三方面工作：(1)基于本体的语义查询扩展；(2)查询语句的规范与重构；(3)信息检索。

2.3.1 基于本体的语义查询扩展

据统计，在信息检索中，人们使用相同的词来表达同一概念的概率不到 20%，这就要求必须在用户原查询词的基础上添加与之相关的词，以解决一义多词的问题。基于本体的语义查询扩展就是借助本体的语义关系、层次结构和推理机制对用户的查询实现语义上的扩展。早在 2003 年，MAKI 等人就提出了基于本体结构进行查询扩展。2004 年张敏等又提出了基于语义关系查询扩展的文档重构方法[2]。

本文综合了基于路径和基于注释两种方法的优点，通过分析影响语义的因素，实现了一种基于语义相似度

的查询扩展。其模型如图 3 所示。

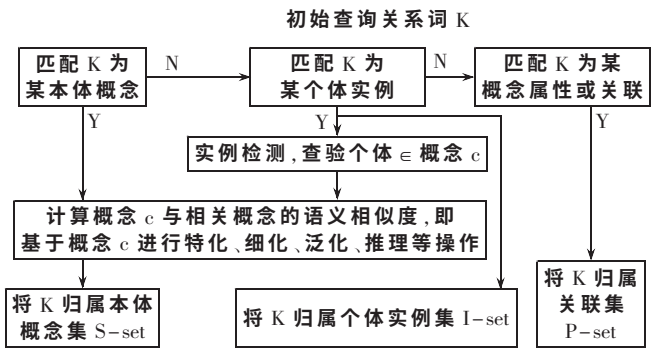

图 3 初始查询词的扩展及规范流程

语义相似度的值通常与概念间的距离、概念间的链接类型等有关。当然不同类型的连接关系，如上位、下位、同义等，对概念语义关联程度的贡献也不同[3-4]。但在实际应用中如果考虑大量的关系类型将会影响系统的性能，因此仅选取贡献较大的几种类型。本模型采用的关系类型及其权重分配方案如表 1 所示。

表 1 关系类型和关系权重

关系类型	同义	继承关系	整体与部分	其他
关系权重	1	0.9	0.6	0

综合以上各种因素，语义相似度的计算公式如下：

$$Sim(a,b)=\theta_1\times Type(a,b)+\theta_2\times[1-Dist(a,b)]+\theta_3\times Outd(a,b)$$

其中 $Type(a,b)$ 表示概念 a、b 之间的关系类型，其取值见表 1。

$$Outd(a,b)=\left(\frac{1}{2^{out\ degree(b)}}+\frac{1}{2^{out\ degree(b)-1}}+\cdots\frac{1}{2}\right)=\sum_{i=1}^{out\ degree(b)}\frac{1}{2^i},$$

其中 $out\ degree(b)$ 表示扩展概念 b 的出度，在本体的层次结构中，如果某一局部概念的密度较大，则说明对概念的细化也越大，概念也就更具代表性。

$Dist(a,b)=\sum_{i=1}^{t}w_i(c)$，表示概念 a、b 之间的语义距离，$w_i(c)=\frac{1}{2^{dep(c)}}$，$dep(c)$ 表示概念 c 在本体中的深度，$t=p_n(a,b)$，$p_n(a,b)$ 为本体中从 a 到 b 最短路径所包含的节点数。

θ_1、θ_2、θ_3 为调节因子，决定概念之间在本体层次树中深度与广度的影响，从而确定系统所需要的相似度，并且 $\theta_1+\theta_2+\theta_3=1$。

2.3.2 查询语句的规范与重构

根据图 3，扩展后的用户查询仍需进一步地规范，以判断扩展后的查询词属于三元组哪一部分并将其分别储存于相应的集合中。最后得到三个集合，分别为本体概念集 S-set、个体实例集 I-set 和属性集 P-set。这三个集合分别对应于三元组的 Subject、Object 和 Predicate，随后分析概念之间以及概念与个体之间的关系，将所有可能产生的概念关联都构建成三元组模式的查询语句提交检索模块。

2.3.3 信息检索

经过以上处理，信息检索模块接受的是具有一定检索规范的结构化查询。为了提高查全率，本模块首先根据用户提供的检索要求，基于知识库进行推理，这种推理是基于类层次和规则的，系统设计者可以根据具体需要创建适合的推理规则。之后仅需与知识库中的信息进行匹配，将满足条件的元组选出。例如，有一个服装领域的本体，对概念“服装”存在一个标签名为“价格”的属性。可以创建这样一条规则，如果价格大于 5 000 元，就认为该衣服为高档服装。所以当用户查询高档服装时，就可以根据这条规则，将知识库中满足条件的实例返回。如果是基于关键字的检索，就仅仅返回包含“高档服装”的网页，而遗漏掉那些不包含“高档服装”但满足用户需求的资源。

2.4 排序模块

通过索引库从文档集中把文档检索出来之后，得到的是一系列无序文档，在递交给用户之前需对文档进行排序。这就需要计算查询与文档之间的相关度。在语义标注时曾讲过，为了量化实例与文档之间的关联程度，通过 TF-IDF 算法来确定实例的标注权重，这样文档 D_j 就能被简化为实例的集合。令 w_{ji} 为实例 I_i 在文档中的权重，则 $d_j=(w_{j1}\cdots w_{jm})$。而查询也可在同一空间里表示成查询向量的形式（即 $q=(q_1\cdots q_m)$），利用余弦定理就可以计算得到文档与查询的相关性：$sim(D_i,Q)=d_i\times q/(|d_i|\times|q|)$。

3 实例验证与分析

为了验证系统的有效性，进行了相关的实验。根据参考文献[5]以及服饰行业专业词典，建立了一个简单的服饰领域本体。利用斯坦福大学研制开发的 Protégé 3.1.1 构建该领域本体，并通过在 Java 中调用 Protégé OWL API 中的方法，直接对建好的领域本体进行操作。由于实验中很多工作需要手工完成，考虑到工作量大，仅从网上抓取 100 篇有关服饰方面的文档组成文档集。该系统中的知识库采用一种十分简单的方法对 Web 上的资源进行标注，即在知识库中手动地添加指向 Web 上文档的 URI。

实验中，当输入“上衣”时，就可根据本体进行扩展。通过对几次试验结果的分析发现，将阈值设置为 0.9、θ_1=0.25、θ_2=0.7、θ_3=0.05 时，能够得到较好的查全率和查准率，这时查询被扩展为{上衣，背心，夹克，羽绒衣，牛仔夹克，衬衫}，如果用户的初始查询词还包括“hasBrand”以及个体实例“adidas”，则查询可构建成如表 2

表 2 查询语句的重构

S-set	P-set	I-set	可能的查询语句	检索内容
上衣	…	…	(上衣，hasBrand，adidas)	Adidas 牌上衣
背心	…	…	(夹克，hasBrand，adidas)	Adidas 牌夹克
夹克	hasBrand	adidas	(衬衫，hasBrand，adidas)	Adidas 牌衬衫
羽绒衣	…	…	…	…
牛仔夹克	…	…	…	…
衬衫	…	…	…	…

所示的 RDQL 形式的查询。

检索模块将结构化检索条件与知识库中 RDF 三元组进行匹配，并返回匹配的所有元组，通过查找实例-文档的索引库，返回无序文档集。排序模块对文档排序后返回。其实验性能如图 4 所示。

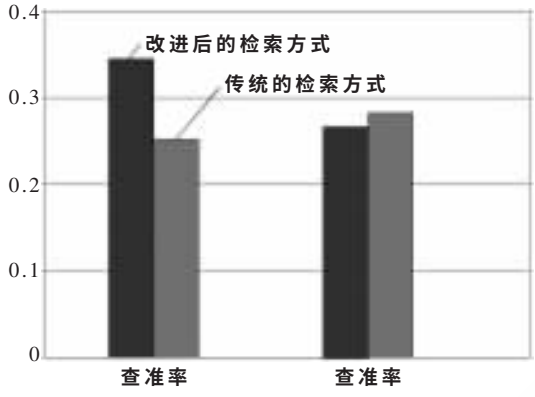

图 4 查询性能比较

性能分析：

(1)查准率。由于查询扩展和语义标注等都是基于本体进行的，这样就明确了术语的选择范围，限制了对术语可能的解释，可以很好地解决一词多义的现象。

(2)查全率。在该系统模型中，由于增加了查询语义扩展，系统可以根据用户提交的检索词推理出与原查询相近或相关的词加入查询系统，以提高检索质量。当然，必须选择合适的阈值和调节参数来控制扩展的阶数，否则在提高查全率的同时查准率将会受到影响。

通过实验可以看出，相对于传统的检索模型，该模型在查准率方面有较明显改善，查全率也几乎相当。但同时也应该注意，系统的检索性能直接取决于知识库中信息的质量及数量。当用户要查询的内容在知识库中比较丰富、完善的时候，能得到较好的检索效果。反之，该模型的检索性能便比不上基于关键字的全文检索性能，为了克服这种缺点，有时需要把基于关键字的检索整合进来，作为该模型的补充。

参考文献

[1] 张敏，宋睿华，马少平.基于语义关系查询扩展的文档重构方法[J].计算机学报，2004，27(10)：1395-1401.

[2] 郭承霞，王爱继，陈庆海.基于领域本体的智能信息检索模型研究[J].计算机科学，2009，36(4A)：101-102.

[3] 聂卉.基于本体的查询扩展与规范[J].知识组织与知识管理，2007，3(148)：35-38.

[4] 熊忠阳，李春玲，张玉芳.一种基于领域本体的混合信息检索模型[J].计算机工程，2008，34(21)：68-70.

[5] 王爱丽，朱欣娟.基于本体的服装领域语义 Web 检索方法[J].西安工程科技学院学报，2007，21(4)：489-493.

(收稿日期：2010-06-21)

作者简介：

马中杰，男，1982年生，硕士研究生，主要研究方向：数据挖掘，语义 Web。

郑诚，男，1964 年生，博士后，硕士生导师，主要研究方向：数据挖掘与 Web 技术。

苏喻，男，1984 年生，硕士研究生，主要研究方向：语义 Web，数据流分析。