一种层次聚类的 RDF 图语义检索方法研究*

刘 宁, 左凤华, 张 俊
(大连海事大学 信息科学与技术学院, 辽宁 大连 116026)

摘 要: 针对当前信息资源描述框架(RDF)检索过程中存在的内存使用过大及检索效率低等问题,提出一个 RDF 图的层次聚类语义检索模型,设计并实现了相应的检索方法。首先从 RDF 图中抽取实体数据,在本体库的指导下,通过层次聚类,将复杂的图形结构转换为适合检索的树型结构;根据在树中查找到的目标对象,确定其在 RDF 图中的位置,进行语义扩充查询。检索模型的构建缩小了检索范围,从而提高了检索效率,其语义扩充查询还可以得到较好的查全率。

关键词: RDF 图; 层次聚类; 语义检索; 向量空间模型
中图分类号: TP18 **文献标志码**: A **文章编号**: 1001-3695(2012)08-2858-04
doi:10.3969/j.issn.1001-3695.2012.08.015

Hierarchical clustering-based semantic retrieval of RDF graph

LIU Ning, ZUO Feng-hua, ZHANG Jun
(*College of Information Science & Technology*, *Dalian Maritime University*, *Dalian Liaoning* 116026, *China*)

Abstract: The current research related RDF graph retrieve exists some problems, such as low efficiency of memory usage, low search efficiency and so on. This paper proposed a hierarchical clustering semantic retrieval model on RDF graph and the method based on the model to solve aforesaid problems. That extracting entities from RDF graph and hierarchical clustering by the guidance of the ontology library made the complex graph structure into a tree structure for efficient retrieval. Orientating target object which was one of nodes in the model in RDF conducted the semantic expansion queries. Retrieval efficiency increased because retrieval scope narrow down as construction of retrieval model and recall ratio increased by the semantic expansion queries.
Key words: RDF(resource description framework) graph; hierarchical clustering; semantic retrieval; vector space model

0 引言

RDF 元数据具有很强的语义特性。在语义网的发展中,RDF 作为其实现基础表示万维网上的信息并交换知识和数据[1]。随着 RDF 数据的大量增加,普通用户直接对其进行检索的需求也在不断增加。目前已经开发出一些查询语言如 SeRQL[2] 和 SPARQL[3]。由于网络中的大部分 RDF 数据缺失模式信息的描述,并且用户必须掌握语法规则和待检索数据的模式信息才能使用查询语言,所以其检索过程对普通用户而言仍过于复杂。普通用户更倾向于简单的关键词检索方式,但是关键词检索并不适用于语义检索,往往会使 RDF 图的检索失去其语义特性,并且使 RDF 图的检索效率很低,无法得到令人满意的结果。

本文针对这些问题,结合图检索与关键词检索的优势以及 RDF 图数据自身特点,提出了一种层次聚类的 RDF 图语义检索模型,很好地解决了当前 RDF 数据检索存在的不足。

1 概述

1.1 RDF 图

RDF 的数据结构为图。RDF 图的最基本构成是以三元组(subject, predicate, object)形式存在的对资源的陈述(statement)。陈述的主语(subject)通常是指向相应资源(resource)的 URI[4],谓语(predicate)是表示某一属性(property)的 URI,而宾语(object)既可以是指向某一个资源的 URI,也可以单纯地以一个文字(literal)作为属性值(property value)。除此之外主语和宾语也可以是没有 URI 表示的匿名资源(通常这类资源没有实际意义或尚未定义),亦称空白节点(blank node)。由此,RDF 三元组可形式化地定义[5]为:(subject, predict, object) $\in (U \cup B) \times (U) \times (U \cup B \cup L)$。其中:$U$ 是 URI 的集合;B 是空白节点的集合;L 是任意可能的数据类型所对应的文字属性值。一组上述形式的三元组被定义为一个 RDF 图(RDF graph)[6]。

定义 1 RDF 图[5]。RDF 图 T 是一组三元组的集合,其有向标记图表示记为 $G=(V, E, l_V, l_E)$:

$$V=\{v_x \mid x \in \text{subject}(T) \cup \text{object}(T)\}$$

$$E=\{e_{s,p,o} \mid (s,p,o) \in T\}$$

$$l_V(v_x)=\begin{cases}(x,d_x) & \text{如果 } x \text{ 是一个文字}(d_x \text{ 是其数据类型})\\ x & \text{否则}\end{cases}$$

其中:V 是 RDF 图中节点的集合;E 是有向边的集合;l_V 和 l_E 分别是 V 和 E 上的标记函数;subject(T)表示 T 中所有三元组的主语的集合;object(T)表示 T 中所有三元组的宾语的集合。

收稿日期: 2011-12-05; **修回日期**: 2012-02-13 **基金项目**: 国家自然科学基金资助项目(61073057);中央高校基本科研业务费专项资金资助项目(2011JC007)

作者简介:刘宁(1957-),女,辽宁大连人,教授,硕士,主要研究方向为语义网、智能信息处理(liun571@126.com);左凤华(1986-),女,硕士,主要研究方向为语义网信息检索、关系数据库检索;张俊(1971-),男,副教授,博士,主要研究方向为关系数据库检索、本体、语义网.

函数 from( )和 to( )给出边的起点和终点。图 1 所示为 RDF 图的一个片段。

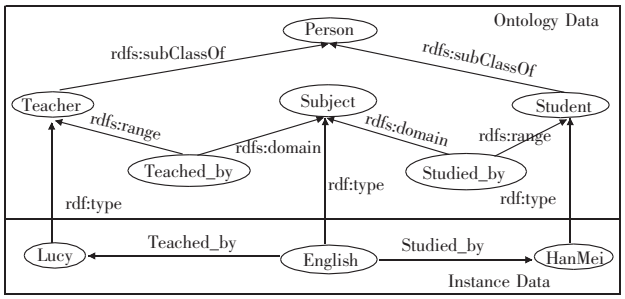

图 1　RDF 图示例

1.2　RDF 关键词检索

当前的 RDF 关键词检索相关工作按处理方式主要有两种研究,第一种方式被称为 K-A(由查询关键词直接构造查询结果),具有代表性的成果有 BLINKS[7]、EASE[8]等;第二种查询方式为 K-Q-A(由查询关键词构造形式化查询语句然后得到查询结果的方式),如 SemSearch[9]、Q-Semantic[10]、SPARK[11]等。

K-A 检索方式的查询对象为 RDF 图,其查询结果为满足某种条件的子图。检索的过程是:a)通过建立相关索引来减少查询响应时间;b)按照特定的查询算法生成查询结果集;c)对候选结果按照相应的评分机制评分,并将前 k 个结果返回给用户。直接检索方法的优点是不需要背景知识,也不需要 RDF 模式信息的支持,检索的粒度细。但这种方法目前不支持对边标签的图进行关键词检索,无法处理用户将属性或关系名作为关键词进行检索的情况。

K-Q-A 检索方式是利用 RDF 的模式信息将关键词查询变为形式化查询。首先,对 RDF 图的顶点或边按照关键词进行匹配查询;锁定目标后,在 RDF 模式信息的指导下,找到与检索关键词相关联的查询结果,确定用户的检索对象;最后,将检索对象构造成规范的形式化查询语句,返回排序的查询语句。用户通过选择查询语句向 RDF 数据库发起查询并获得最终查询结果。这种方法不需要建立大规模的结构索引,而是依赖 RDF 模式信息(RDF schema)确定查询关键词之间的关联,但目前万维网上存在的大部分 RDF 数据都没有或缺少模式信息,同时检索的响应时间开销也很大。

2　RDF 图层次聚类检索模型构建

2.1　RDF 图层次聚类检索模型提出

如图 2 所示,RDF 图层次检索模型共有四层,由下至上分别为 RDF 图层、实体层、实体聚类层、用户接口层。其中下一层是上一层的基础,上、下两层的节点之间相互对应;实体聚类层的最底层(即叶子节点)与实体层的各个实体是一一对应的。

用户接口层	用户输入关键词并获取查询结果
实体聚类层	由实体层实体通过聚类而成的树型结构
实体层	通过与上层的映射找到对应实体
RDF 图层	通过上层映射及下层语义关系找出查找内容

图 2　RDF 图层次聚类检索模型

执行查询时根据输入的关键词自顶向下进行检索,当检索到树的叶子节点时便可以进入下一层与之相对应的实体节点,然后由实体节点找到相应的 RDF 图中的三元组,通过 RDF 图的语义关系,进而可以找到更多的与检索相关的隐式数据。这样可以在很大程度上提高检索效率,同时又能充分体现出 RDF 语义特性的优势。

2.2　RDF 图层次聚类检索模型构建

本节主要工作是由 RDF 图找出实体的集合,构造出聚类检索模型的第二层,它的作用是第一层和第三层的连接枢纽。如图 1 所示,一个完整的 RDF 数据中具有模式信息,它描述了 RDF 实例数据中实例的属性、关系以及实例的类别等信息,所以通过对 RDF 模式信息的解析很容易找到 RDF 数据中存在的实体。但是目前网络中存在的大量 RDF 数据是缺少甚至没有模式信息的,所以本文针对这样的应用背景提出了一种实体抽取方案。其实现主要包括以下几个步骤:先对 RDF 图进行实体抽取,生成实体集合;然后参照本体库为每个实体进行语义标注,即确定每个实体在本体库中的类;进而构建每个类的向量空间模型,根据每个类的向量空间模型计算类之间的概念相似度,由概念相似度的值的大小进行聚类;最后生成索引库和层次聚类模型。具体的构建过程如图 3 所示。

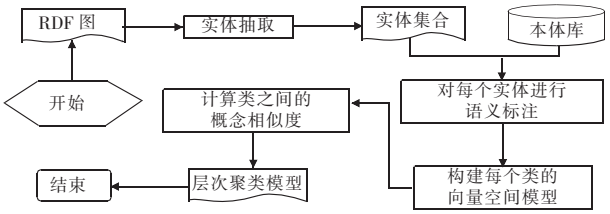

图 3　RDF 图层次聚类检索模型构建

2.2.1　RDF 图的实体抽取

实体抽取主要分两步进行,一是文档的预处理;二是实体抽取。文档的预处理主要是对 RDF 文档中标签的解析,因为标签中的内容均由单词或词组构成,所以与文本中的实体识别不同,这里可直接进行词语标注,完成对标签的解析;接下来便可以按照标签的解析结果对 RDF 节点进行命名实体识别,进而完成实体抽取工作。需要说明的是,RDF 文档的预处理操作非常重要,因为在后面的层次聚类中,它将是聚类的一个依据。其具体实施过程如图 4 所示。

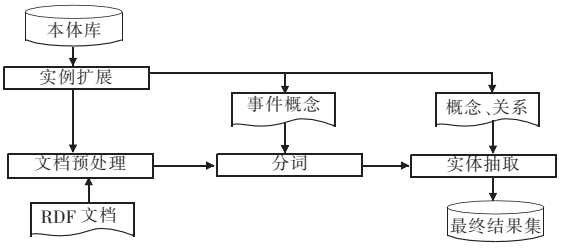

图 4　RDF 实体抽取过程

目前 Web 服务描述标准采用的是基于 XML 的 WSDL 语言[4],它能够方便地描述 Web 服务调用接口,所以将抽取得到的实体存储到实体集合中,采用 WSDL 语言将每个实体的名称及属性以及在 RDF 图中的映射关系进行存储。

2.2.2　实体集合的语义标注

引入语义标注的目的是将 2.2.1 节中抽取到的实体进行语义扩充,为每一个实体找到在本体库中所对应的本体概念。这个概念作为层次聚类中的底层概念;所以这个底层概念的确定直接关系到了聚类的效果和检索的质量。

对抽取到的实体集合进行语义标注时,考虑到网络数据应

用背景下其数据规模较大,所以不宜采用手动标注和半自动标注。进行语义标注的过程其实就是实体集合中的各个实体与本体库中本体概念相似度计算的过程。在多个实体—概念对的相似度计算值中,与该实体拥有最大相似值的概念即认为该实体为对应本体概念的实例。

因为抽取到的实体以 WSDL 的形式记录,所以对实体的标注过程就是对 WSDL 文件中元素选择最适合的本体概念。WSDL 文件的 XML schema 中隐含了大量的语义信息,其数据类型与本体描述语言 OWL 具有结构相似性,所以有之前的 WSDL 语言描述实体的基础,实体的语义标注工作变得相对简单许多。

2.2.3 实体集合的层次聚类

层次聚类是聚类算法的重要分支。它无须预先指定聚类中心或最终簇的数目,通过将数据反复地聚集形成一系列嵌套的簇,并在合适的层次终止聚类来获得期望的结果[12],更适用于信息检索领域。本文对 RDF 图进行实体聚类,其数据特点决定了相比传统的聚类而言会减少很多繁杂的工作。

在 2.2.2 节完成了 RDF 图中实体节点到本体库中概念的映射,这样的概念在本文中称为底层概念。层次模型的最底层是 RDF 图中的实体,每一个实体都有一个本体概念作为其底层概念(允许一个底层概念连接多于一个实体)。本文的层次聚类则是针对被映射到的底层概念进行的。首先将所有的底层本体概念作为基本类的集合,然后构造每个底层概念的向量空间模型(VSM),其中向量元素为每一个实体对应的本体库中的概念。根据向量空间模型来计算类之间的相似度,这里的相似度采用两个空间向量的余弦值来判断。对于已知的 v_i 和 v_j,可以根据式(1)判断其相似度。

$$\mathrm{Sim}(v_i,v_j)=\sum_{k=1}^{n}(t_{ik}\times t_{jk})/\sqrt{\sum_{k=1}^{n}(t_{ik})^2\times\sum_{k=1}^{n}(t_{jk})^2} \tag{1}$$

然而仅仅利用为底层概念构建向量空间模型计算相似度进行聚类的方法并没有从语义上解决 RDF 的聚类问题,这势必会影响 RDF 图的语义检索。为此,本文引入了语义距离这一概念。一般认为两个本体概念之间相似度可以用语义距离来衡量[13]。所以很多专家在计算语义相似度时,常引入语义距离这一概念。文献[13]中对概念边进行了划分,并给出语义距离权重表,如表 1 所示。

表 1 语义距离权重表

关系	关系		
	p	g	s
∅	0	1	1
p	0	1	1
g	1	2	3
s	1	1	2

这张概念权重表应该结合本体库中的概念间的语义关系边进行使用。其中 p 代表正关联关系的边;g 代表 generalization,即泛化,指从子节点向父节点的边;s 代表 specialization,即细化,指从父节点指向子节点的边;∅代表空操作,与列的组合表示单个关系的边。按照最近祖先节点与节点之前边的语义关系,给出的语义距离公式为

$$\mathrm{Dist}(v_i,v_j)=\sum_{k=1}^{n}\omega_{ck}+\frac{N_{vi}+N_{vj}}{N_{vi}+N_{vj}+2N_{LCA}}\times\varepsilon \tag{2}$$

其中:$\sum_{k=1}^{n}\omega_{ck}$ 表示 v_i 和 v_j 最短路径中各边距离加权之和;N_{vi}、N_{vj} 表示点 v_i 到与其最近的公共亲节点的距离;N_{LCA} 为最近公共父亲节点到根节点的距离。根据式(1)和(2)便可以确定本体概念相似度的计算公式为

$$\mathrm{SemSim}(v_i,v_j)=\frac{1}{(\mathrm{Dist}(v_i,v_j))^2+1}\times\alpha+\mathrm{Sim}(v_i,v_j)\times\beta \tag{3}$$

其中:α 与 β 是和为 1 的调节参数。当相似度的值超过阈值 q 时,则可以将两个底层概念聚类成一个概念。在聚类算法中,有一个重要的工作就是聚类标签的选择。本文与传统方法的不同之处是传统的聚类是无指导的,而本文的聚类可以在本体的指导下完成,具体按照以下几个规则:

a)若 C_i 与 C_j 为同一个类,则其中任意一个为聚类后节点的标签;

b)若 C_i 为 C_j 的祖先节点,则 C_i 为其聚类后节点的标签;

c)若 C_i 为 C_j 的子孙节点,则 C_i 为其聚类后节点的标签;

d)若 C_i 与 C_j 无继承关系,则找到同时连接 C_i 与 C_j 最近的祖先节点作为聚类后节点的标签。

需要说明的是,若是第 c)种情况,较高的相似度值就决定了 C_i 与 C_j 的公共祖先节点与它们的距离不会太远,所以这一点符合聚类标签选择的合理性。

按照上述方法,便可以直接进行后续层的聚类工作,直至将所有概念聚类成一个类,得到的层次聚类结果如定义 2 所述。

定义 2 RDF 聚类检索模型。RDF 聚类检索模型结构为树,C_i 为树中的一个节点。C_i 应该满足如下条件:

a)C_i 为一个实体(或对象),若 C_i 为叶子节点;

b)C_i 为一个类,若 C_i 为非叶子节点。

其中,若 C_i 为叶子节点,则其应该包含所有从 RDF 图中抽出的实体(或对象)。

3 RDF 图的语义检索方法实现

3.1 用户检索需求的预处理

根据 2.2 节的叙述,本文所设计的 RDF 图检索模型是通过实体检索进而得到用户需求的。但是对于普通用户,由于相关专业知识的限制,要求用户输入需要的实体或属性显然是不切实际的。搜索引擎的广泛而成功的应用带来的启示是,基于关键词的查找方式最适合大多数用户[14]。本节的主要工作是对用户需求的预处理。对用户输入关键词的处理直接影响到后面的检索工作,因为只有对关键词进行正确的扩充和定位,才能匹配到满意的检索结果。

用户输入关键词预处理主要完成关键词到领域本体的映射,由关键词找到本体库中与之相对应的概念。把本体库中的每个概念看做是由多个关键词组成的向量,关键词可以组成向量的维,将本体库看做是由概念组成的向量空间,可以表示为 $\mathrm{VSM}=\langle C_1, C_2, \cdots, C_i, \cdots, C_n\rangle$。在向量空间模型中,每一个概念可以用一个向量表示 $C_i=\langle t_{i1}, t_{i2}, \cdots, t_{ik}, \cdots, t_{in}\rangle$,其中 $1\leqslant k\leqslant n$。并且每一个 t_{ik} 应该对应这个关键词与这个概念相似度的权值 w_{ik},所有的概念又可以表示成 $C_i=\langle (t_{i1},w_{i1}), (t_{i2},w_{i2}), \cdots, (t_{ik},w_{ik}), \cdots, (t_{in},w_{in})\rangle$,其中 $1\leqslant k\leqslant n$。

根据著名的计算特征权重的方法——LTC[15] 计算权重法,计算关键词的特征权重为

$$w_{ik}=\frac{\log(f_{ik}+1)\times\log\left(\frac{N}{N_k}\right)}{\sqrt{\sum_{j=1}^{M}\left[\log(f_{ij}+1)\times\log\left(\frac{N}{N_j}\right)\right]^2}} \tag{4}$$

其中:f_{ik}是关键词 t_k 在概念向量 C_i 中出现的词频;N_k 是关键词 t_k 在概念向量空间 VSM 中出现的总次数;N 是概念向量空间 VSM 中所有概念的数量;M 是概念向量空间 VSM 中所有去除停用词后的有效关键词的总数。$w_{ik} \in [0,1]$。

在领域本体中存在关系概念 $R \in C \times C \times W$。$W$ 可以看做是领域概念的相关度。领域本体 O 中概念间每个关系都对应一个权值 w_{ck}($w_{ck} \in [0,1]$),这个权值是本体构建过程中由领域专家赋予的,它表示各种关系概念间的相关度。借鉴文献[16]中给出的计算 query-document similarity 值所用公式,给出检索结果与关键词语义相关度的计算是由相关词的特征权重与语义关系对应权重按式(5)求得

$$\mathrm{KCSim}(D_i, O_c) = \sum_{T_k} (w_{ik} \times w_{ck}) \tag{5}$$

利用式(5)中相似度值大小排序,就可以计算出用户输入关键词的最相关的本体概念或有着密切关系的概念。

3.2　实体聚类树的检索

本文中的语义检索主要是通过计算层次聚类生成的概念树中节点与检索条件的语义相似度值,将相似度值较高的节点作为检索结果。如 3.1 节所述,用户输入的关键词将会被转换为本体中的概念,这个概念与用户输入的关键词、相似度值一同构成检索条件三元组 $SC(KW, C, \mathrm{Sim})$。由定义 2 可知,一个高度为 N 层的层次聚类检索树的第 N 层到第 $N-1$ 层均为概念节点,第 1 层为实体。所以检索可以分两部分进行,一是对概念相似度的计算,二是对实体相似度的计算。

因为检索条件中的概念以及层次模型中的概念均是参照本体库扩充而得到的,所以两者的相似度计算是基于式(2)中的语义距离进行的更改。检索条件中的概念是根据用户输入关键词扩充而得到的,即检索条件中的相似度值为关键词与概念的相似度。因此在进行检索时,应考虑到检索条件中的相似度值的影响。故在式(3)的基础上加上 KCSim 因数变为

$$\mathrm{SemSim}(v_i, v_j) = \left(\frac{1}{(\mathrm{Dist}(v_i, v_j))^2 + 1} \times \alpha + \mathrm{Sim}(v_i, v_j) \times \beta\right) \times \mathrm{KCSim} \tag{6}$$

对实体相似度的计算主要包含元素为实体名称、属性及属性值的判断。因为用户输入的关键词字数有限,所以检索预处理时还原的实体中属性个数不会很多。采用向量空间方法来计算相似度,其与概念之间的计算相似,这里不再赘述。最终实体节点的相似度值为与其相连的底层概念的语义相似度值与实体相似度值两者求和。具体检索算法——entity cluster retrieval(ECR)如下:

a)将转换后的检索条件按 Sim 值由大到小排序。

b)将当前未被检索的检索条件三元组中 Sim 值最高的一个自上而下与树中节点计算语义相似性:

(a)若该节点为条件概念的祖先概念,则直接计算下一层的孩子节点,转向步骤(a);

(b)若节点标签概念与条件概念相同,则计算所有以该节点为祖先节点的第 $N-1$ 层概念节点与检索条件的相似度,然后计算叶子节点相似度,两者求和后超过阈值的节点将其放入结果集;

(c)若节点标签概念与条件概念不相关,则判断检索该节点是否为根节点,若为根节点或无未被检索的兄弟节点,则终止当前查询,进入步骤 c);若为非根节点,则转向其兄弟节点,转向步骤(a)。

c)将由当前检索条件查询到的实体放入该检索条件的结果集中。

d)判断是否有未被检索的条件,若有则转向步骤 b);否则终止此次查询。

3.3　RDF 图语义扩充检索

本节主要是利用 RDF 图的语义关系,将查找到的实体按照对应关系定位到 RDF 图中的节点,然后按照与之具有语义关系的相近节点进行语义扩充。但并不是所有与核心节点有关系的节点的重要性都是相同的,语义关系的类别不同,对检索结果的重要程度也不同。本文就常用的几种语义关系的权值作如表 2 中的定义。

表 2　语义权值定义

关系类型 R	权值 W
整体与部分关系	0.500
实例与概念关系	0.010
同义关系	0.950
上下位关系	0.500
同类关系	0.025
相关关系	0.010

按照语义权值的大小,将语义权值超过阈值的节点根据语义关系扩充到的节点追加到 3.2 节中得到的核心节点作为附加检索结果。这样在提高检索效率的同时充分利用了 RDF 的语义特性,能够很大程度地提高查全率。

4　结束语

随着网络中 RDF 数据量的迅速增长,普通用户对 RDF 的检索需求也在增加。目前存在大量的 RDF 数据缺少模式信息,并且普通用户不了解查询结构和查询语言。为了解决这些问题,本文提出的层次聚类检索模型及语义检索方法,既支持关键词查询,其检索又不依赖 RDF 模式信息。层次聚类检索模型的设计引入了语义距离的概念,将原来的 RDF 图中语义边进行量化,使得聚类更符合语义检索的需求,由图转换成树也大大地提高了检索效率。对于含有 N 个实体节点的 RDF 图层次聚类模型而言,其查找长度可以达到 $\log_2 N$。用户输入的关键词的转换和语义扩充检索不仅有良好的用户体验,而且在一定程度上提高了查全率。

该检索方法仍有许多方面需要继续完善。在 3.1 节用户需求的转换中,将关键词转换为本体中的概念,这里一定要注意概念粒度的大小得当问题。最理想的效果是所转换的概念恰好涵盖用户需求,否则将会使检索结果的数量增加许多。其次是系统实现方面,需要将文中提到的模型构建以及检索方法通过一定的数据进行有效地验证,本文将在后续工作中研究这部分的内容。

参考文献:

[1] DING Li, FININ T. Characterizing the semantic Web on the Web [C]//Proc of the 5th International Conference on Semantic Web. Berlin: Springer-Verlag, 2006: 242-257.

[2] EHRIG M, HAASE P. SeRQL: Sesame RDF query language [EB/OL]. (2003-04-09) [2011-10-28]. http://swap.semantic web.org/public/Publications/swap-d3.2.pdf.

[3] PÉREZ J, ARENAS M, GUTIERREZ C. Semantics and complexity of SPARQL[C]//Proc of the 5th International Conference on Semantic Web. Berlin: Springer-Verlag, 2006: 30-43.

[4] 吴刚. RDF 图数据管理的关键技术研究[D]. 北京:清华大学, 2010.

(下转第 2955 页)

(上接第 2861 页)

[5] 李慧颖,瞿裕忠. KREAG:基于实体三元组关联图的 RDF 数据关键词查询方法[J]. 计算机学报, 2011,34(5):825-835.

[6] W3C. Resource description framework (RDF): concepts and abstract syntax[EB/OL]. (2004-05-23)[2011-10-19]. http://www.w3.org/TR/rdf-concepts/.

[7] HE Hao, WANG Hai-xun, YANG Jun, *et al*. Blinks: ranked keyword searches on graphs[C]//Proc of ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2007:305-316.

[8] LI Guo-Liang, OOI B C, FENG Jian-hua, *et al*. EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data[C]//Proc of ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2008:903-914.

[9] LEI Yuan-gui, UREN V, MOTTA E. SemSearch: a search engine for the semantic Web[C]//Proc of the 15th International Conference on Managing Knowledge in a World of Networks. Berlin: Springer-Verlag, 2006:238-245.

[10] WANG Hao-feng, ZHANG Kang, LIU Qiao-ling, *et al*. Q2Semantic: a lightweight keyword interface to semantic search[C]//Proc of the 5th European Conference on Semantic Web. Berlin: Springer-Verlag, 2008:584-598.

[11] ZHOU Qi, WANG Cong, XIONG Miao, *et al*. SPARK: adapting keyword query to semantic search[C]//Proc of the 6th International Conference on Semantic Web and the 2nd Asian Conference Semantic Web. Berlin: Springer-Verlag, 2007:649-707.

[12] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1):48-61.

[13] CROSS V. Fuzzy semantic distance measures between ontological concepts[C]//Proc of IEEE Annual Meeting of the Fuzzy Information. Washington DC: IEEE Press, 2004:635-640.

[14] 马光志,廖家国. P2P 网络中基于 RDF/S 的数据库语义查询系统设计[J]. 计算机应用研究, 2007, 24(10):260-262, 266.

[15] AAS K, EIKVIL L. Text categorization: a survey, NR 941[R]. Oslo, Norway: Norwegian Computing Center, 1999.

[16] SALON G, BUCKLEY C. Term weighting approaches in automatic text retrieval[J]. Information Processing and Management, 1988, 24(51):513-523.

[17] LADWIG G, TRAN T. Combining query translation with query answering for efficient keyword search[C]//Proc of the 7th International Conference on Semantic Web. Berlin: Springer-Verlag, 2010: 288-303.