

利用相互增强关系迭代计算本体中 概念与关系的重要性

吴 刚 张 阔 李涓子 王克宏

(清华大学计算机科学与技术系知识工程实验室 北京 100084)

摘 要 通过排序本体中概念重要性和关系权重的方式评价本体,能够辅助领域专家改进本体设计,辅助语义 Web 搜索引擎实现. 现有链接分析技术不能直接应用于对概念的排序,而且缺乏有效方法对关系赋予权重. 文中提出依据本体的图结构特点,以 Hub 值代替 Authority 值作为概念重要性,并利用本体中概念和关系相互增强的迭代方式计算概念重要性和关系权重,证明该迭代过程收敛于迭代方程组的不动点. 实验初步表明,该方法具有与 PageRank 接近的收敛速度,并能得到合理的概念重要性与关系权重的排序结果.

关键词 本体;语义 Web;排序;链接分析;收敛

中图法分类号 TP311

Ranking by Mutually Reinforcing Concepts and Relations in Ontology

WU Gang ZHANG Kuo LI Juan-Zi WANG Ke-Hong

(Knowledge Engineering Group, Department of Computer Science and Technology, Tsinghua University, Beijing 100084)

Abstract Ranking the importance of concepts and the weights of relations is an effective method for evaluating an ontology, which can improve the design of ontology for domain expert, and be used as a component of semantic Web search engine. Current link analysis ranking algorithms cannot be directly applied to rank concepts, or efficiently to assign weights to relations. According to the characteristic of ontology graph structure, an algorithm is proposed with Hub rating instead of Authority rating as importance of concepts. The algorithm mutually reinforces importance of concepts and weights of relations in the iteration process which is proved to converge to the fixpoint of equations. The experimental results show the algorithm has the similar convergence speed to PageRank but more reasonable ranking of concepts importance and relations weights.

Keywords ontology; semantic Web; ranking; link analysis; convergence

1 引 言

本体表达了语义 Web 中数据的语义,是语义 Web 的基础之一. 本体设计的好坏直接影响了语义

Web 对知识的表达. 有效地评价本体对辅助领域专家改进本体设计,辅助理解本体并构建语义 Web,辅助语义 Web 搜索引擎的实现等都具有重要实际意义.

对本体中描述的概念的重要性进行排序,并对

收稿日期:2006-06-07;最终修改稿收到日期:2007-06-05. 本课题得到国家自然科学基金(90604025)资助. 吴 刚,男,1978 年生,博士研究生,研究方向为语义 Web 数据管理等. E-mail: wug03@mails. tsinghua. edu. cn. 张 阔,男,1981 年生,博士研究生,研究方向为信息抽取、数据挖掘等. 李涓子,女,1964 年生,博士,副教授,研究方向为语义 Web 与 Web 服务、文本挖掘与知识发现. 王克宏,男,1941 年生,教授,博士生导师,研究领域为网络计算与知识处理.

概念到概念之间的关系赋予权重是一种有效的本体评价方式. 图 1 是对一个开源软件项目本体^①的有向图表示(为方便叙述,只显示了有向图的局部). 由于该本体的设计目的是描述开源软件项目,显然其中重要的概念应该包括“Project”和“Developer”,相应的重要关系应该包括“manage”和“developed_by”. 合理的概念重要性排序和关系权重设置将能够正确反映本体设计者的设计思路.

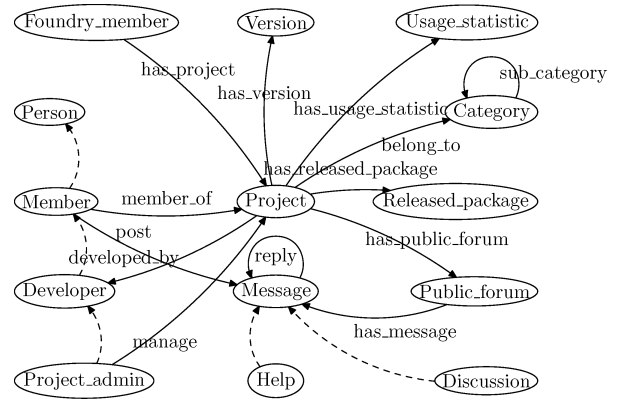


图 1 开源软件项目本体

基于链接分析的排序算法,例如,PageRank^[1], HITS^[2],SALSA^[3]等以随机过程理论为基础和收敛性分析依据,能够对有向图中的节点做出有效的评价,适用于 WWW 搜索引擎的实现. 由于不区分有向边的类型,此类算法缺乏对有向边权重计算方法的考虑.

对象级^[4]链接分析排序算法的提出扩展了上述算法,将有向标记图的边用不同的标记进行区分,着重分析对象间的关系(边)的权重对对象(节点)重要性排序的影响. PopRank^[4], ObjectRank^[5]和 RealWalk^[6]算法分别采用经验值方法、优化方法和统计方法对边的权重赋值. 然而对于事先未知的本体,或不存在可用于学习的本体实例的情况下,以上方法都无法使用. 显然对于未知的本体以上方法需要额外的权重学习时间,难于被通用搜索引擎采用.

重要性排序方法也被应用于语义 Web 检索和本体选择,典型方法包括: Swoogle^[7], OntoSelect^[8], OntoKhoj^[9]和 AKTiveRank^[10]. 对于 OntoSelect, OntoKhoj 和 AKTiveRank 3 种方法的重要性排序粒度均为本体文档级别,无法对本体内部的概念和关系给予重要性排序. 其次 OntoSelect 和 OntoKhoj 仍然采用了类似 PageRank 的算法,通过计算 Authority(或称为 Popularity)值来度量本体重要程度,即认为被多数本体引用的本体更重要. 然而,这种度量方法不适用于对本体内部的概念重要性和

关系的权重进行排序. 例如:概念“owl: Class”,由于大量在本体定义中使用,具有较高的 Authority 值,但对于特定的领域本体而言,“owl: Class”的重要性却要低于领域本体中定义的其它概念. AKTiveRank 算法考虑了 Density, Betweenness, Semantic Similarity 和 Class Match 等多种语义 Web 相关的结构信息. 但是这些度量信息的计算(分别简称为 DEM, BEM,SSM 和 CMM)部分地依赖于用户在查询本体信息时提出的查询语句. 比较而言,Swoogle 是最相关的研究工作,因为其不但可以对本体进行文档级别的重要性排序,而且可以单独地对本体中的概念重要性和概念之间的关系分别进行排序. 但是该算法仍然采用类似 PageRank 的度量方法. 此外在对概念和关系进行排序时还需要额外提供本体实例的统计信息. 在缺少这些统计信息作为先验知识的情况下(这种情况在本体设计阶段和认知阶段普遍存在),Swoogle 无法用于对本体中的概念和关系进行评价. 表 1 对比了相关工作.

表 1 相关工作对比

	概念重要性排序	关系重要性排序	采用的基本方法
OntoSelect	不支持	不支持	类似 PageRank
OntoKhoj	不支持	不支持	类似 PageRank
AKTiveRank	不支持	不支持	CMM+DEM+SSM+BEM
Swoogle	支持	支持	类似 PageRank,需统计信息

本文提出依据本体中的概念与关系所构成的链接图的特点,以 Hub 值代替 Authority 值度量概念的重要性. 利用本体中概念和关系相互增强的迭代方式计算概念重要性和关系权重. 这一方法更符合人们对本体认知的意识行为,体现本体设计者的设计意图. 可以证明该迭代过程收敛于迭代方程组的不动点. 同时实验也表明其具有与 PageRank 接近的收敛速度,并能得到合理的概念重要性与关系权重的排序结果.

本文第 2 节提出本文所采用的本体评价模型及相关定义;第 3 节描述概念和关系相互增强的迭代算法;第 4 节给出迭代算法收敛性的证明;第 5 节通过实验对算法进行评价;最后总结全文并指出今后的工作方向.

2 本体设计的意识流模型与相关定义

2.1 模型描述

通过概念(concepts)和关系(relations),本体给

① <http://keg.cs.tsinghua.edu.cn/project/software.owl>

出了一个领域内共享概念模型的明确的形式化规范说明. 概念是一个领域内相关的任何事物,在本体的有向图表示中被表示为节点;关系是领域中概念之间的交互作用,在本体的有向图表示中被表示为有向边,起点表示关系中实施作用的概念,终点表示关系中被作用的概念.

领域专家以定义本体的方式把对领域知识的理解记录下来,换言之,一个本体的定义过程反映了领域专家在设计领域本体时的意识活动. 这一现象可以通过 William James 的著名意识流^[11]理论得到解释. 该理论认为人类意识是由实在部分(substantive parts)和过渡部分(transitive parts)构成. 前者即指思想和想法,后者指想法之间的过渡过程. 该理论同时认为人们的认知意识过程就是不断地从一个想法移动到另外一个,在此过程中过渡部分控制着意识的流动方向. 在本体领域中,概念和关系分别对应上述理论中的实在部分和过渡部分. 因此本体的设计者设计本体的过程就是根据其对领域知识形成的意识流,通过某些特定的关系,将概念关联到其它概念,在此我们称之为本体设计意识流. 本体的使用者在分析该本体时也必然会无意识地遵循领域专家的这种潜在的本体设计意识流.

以图 1 为例,依据该本体作者的描述,其定义过程为:首先定义“Project”概念以及相关的描述性文本如“Version”和“Usage_statistic”;然后定义“Developer”作为对“Project”的补充,两者间通过关系“developed_by”连接;继而为“Developer”定义从“Person”到“Project_admin”的概念层次关系. 该定义过程持续下去,最终得到了完整的本体. 当使用者使用该本体的时候会比较清楚地发现“Project”和“Developer”比较重要. 这是因为这些概念要么具有大量关系指向其它概念(如“Project”),要么具有指向其它重要概念的关系(如“Developer”). 因此这些特征反映了本体设计者的设计意图,即本体设计的意识流.

通过以上分析,在本体的评价中可以采用具有如下特征的本体设计的意识流模型来表达:

特征 1. 一个概念作用于其它概念的关系越多,则该概念的重要性越高,即在有向图表示中从一个节点出发指向其它节点的有向边越多,则该节点越重要;

特征 2. 被作用的概念越重要,则作用于该概念的其它概念的重要性也越高,即在有向图中有有向边指向重要节点的那些节点也重要;

特征 3. 概念的重要性越高则作用于其它概念上的关系权重也越高,即由重要节点出发的那些有向边的权重较高.

该模型的以上 3 点特征体现了本体中概念重要性所具有的 Hubs^[2]值特征以及概念重要性和关系权重之间的相互增强特征.

2.2 相关定义

定义 1. 本体可以表示为有向图 $G=(V,E)$, 其中 V 是本体中所描述的全体概念的集合,定义在概念集合 V 之上的二元关系 E 是本体中所描述的概念到概念之间的关系的集合.

根据定义 1 以 W3C 制定的 RDF/S,OWL 等本体语言规范描述的领域本体可以用有向图的形式表示. 图 1 将“software.owl”文件中定义的开源软件项目本体以有向图(局部子图)的形式进行了表示.

定义 2. 设本体 $G=(V,E)$ 中含有 $|V|=n$ 个概念 v_1, v_2, \dots, v_n , 则本体对应有向图的邻接矩阵表达是一个 $n \times n$ 矩阵 $A=(a_{i,j})$, 其中 $1 \leq i, j \leq n$,

$$a_{ij} = \begin{cases} 1, & (i,j) \in E \\ 0, & \text{其它} \end{cases} \tag{1}$$

按照左上到右下的顺序将图 1 中的节点编号(即令“Foundry_member”为 v_1 , “Discussion”为 v_{14}), 可以构造如下的邻接矩阵表达:

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

定义 3. 设本体 $G=(V,E)$ 中含有 $|V|=n$ 个概念 v_1, v_2, \dots, v_n 以及一个边权重函数 $w(i,j)$, 则本体对应有向图的权重矩阵表达是一个 $n \times n$ 矩阵 $W=(w_{i,j})$, 其中 $1 \leq i, j \leq n$,

$$\begin{cases} 0 < w_{i,j} \leq 1, & (i,j) \in E \\ w_{i,j} = 0, & \text{其它} \end{cases} \tag{2}$$

值 $w_{i,j}=w(i,j)$ 是边 $(i,j) \in E$ 的权重. 权重值越高表示对应的关系越重要.

在本体的有向图表示中,给定一个概念 c ,记由概念 c 出发所指向的所有概念的集合为 F_c ,记指向概念 c 的所有概念的集合为 B_c . 设初始的边权重函数为 $w_{i,j}=1/|F_j|$,若存在 $(i,j) \in E$,则图 1 的初始权重矩阵表达如下:

$$\mathbf{W}_0 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{5} & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & \frac{1}{2} & 0 & 0 & 1 & \frac{1}{2} & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{5} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{5} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{5} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{5} & 0 & 0 & 0 & 0 \end{bmatrix}.$$

定义 4. 设本体 $G=(V,E)$ 中含有 $|V|=n$ 个概念 v_1, v_2, \dots, v_n , 概念 v_i 的重要性定义为 r_i , 且有 $\sum_{i=1}^n r_i=1$, 其中 $1 \leq i \leq n$.

概念重要性向量 $\mathbf{R}=(r_1, r_2, \dots, r_n)$ 和关系权重矩阵 \mathbf{W} 将被分别用于表达概念重要性排序和关系权重.

3 概念与关系相互增强的排序算法

给定一个本体的有向图表示 $G=(V,E)$, 对于任意一个概念节点 s , 其重要性 $r(s)$ 和所有的关系权重 $w(s, t_i)$ 可以根据 3.1 节中的相关定义及第 2 节所述的本体设计意识流模型, 使用下述式(3)和式(4)迭代计算得到.

$$w_{k+1}(s, t) = \frac{r_k(s)}{\sum_{t_i \in B_t} r_k(t_i)} \tag{3}$$

$$r_{k+1}(s) = \frac{1-\alpha}{|V|} + \alpha \sum_{t_i \in F_s} r_k(t_i) w_{k+1}(s, t_i) \tag{4}$$

其中 $k=0, 1, 2, \dots, s, t \in V$.

式(4)利用 Reverse PageRank^[12]有效计算节点

的全局 Hub 值. 其中参数 α 与 PageRank 算法中使用的阻尼因子相当, 取值范围为 $0 \sim 1$ (本文中设置为 0.85). $r_k(s)=r_s^k$ 为概念节点 s 在第 k 次迭代后得到的概念重要性. $w_k(s, t)=w_{s,t}^k$ 为概念节点 s 到 t 之间的关系在第 k 次迭代后的权重.

任意给定一个初始分布 $\mathbf{R}_0=(r_1^0, r_2^0, \dots, r_n^0)$, 第 4 节将证明迭代过程将最终收敛到由式(3)和式(4)组成的非线性方程组的不动点 \mathbf{R}^* . 我们认为向量 \mathbf{R}^* 包含所有概念的重要性, 与之相应的 \mathbf{W}^* 包含所有关系的权重. 由非线性数值分析理论^[13], 前后两次迭代结果的距离 $\|\mathbf{R}_k - \mathbf{R}_{k-1}\|_\infty$ 充分小时, 可取 \mathbf{R}_k 作为不动点 \mathbf{R}^* 的近似结果, 迭代停止.

如图 1 所示: 本体的有向图表示中, 设置概念重要性的初始分布为均匀分布, 则代入式(3)后的计算结果为 3.1 节所述的矩阵 \mathbf{W}_0 . 经过 62 次迭代后可以得到如下极限分布 $\mathbf{R}_{62}=(0.05844, 0.02831, 0.02831, 0.02831, 0.031694, 0.13692, 0.22061, 0.02831, 0.14469, 0.031521, 0.031521, 0.16833, 0.031521, 0.031521)$. 相应得关系权重矩阵 \mathbf{W}_{62} 如下:

$$\mathbf{W}_{62} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0.16 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.13 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0.38 & 0 & 0 & 0.52 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0.87 & 0 & 0 & 1 & 0.57 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.12 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.12 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.46 & 0 & 0.43 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.12 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.12 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

通过对向量 \mathbf{R}_{62} 中的元素排序可见, 以上结果反映出概念节点“Project”最重要, 其重要性值为 0.22061 (\mathbf{R}_{62} 中的第 7 个分量) 大于其它概念节点的重要性. 概念节点“Project_admin”和“Developer”次之. 通过将权重矩阵中某一概念节点所对应的行向量中的各个分量分别乘以对应的概念重要性, 并对非 0 值排序就可以得到以该概念为起点的关系的重要性排序. 例如, 对于概念节点“Project”而言, 按照上述方法可以得到一个行向量 $(0, 0.02831, 0.02831, 0, 0.02757, 0, 0, 0.02831, 0.08247, 0, 0.031521,$

0, 0, 0),其中各分量分别对应由概念“Project”出发到其它概念之间的关系的程度.由此可见最重要的关系是“developed_by”.

由式(3)和式(4)以及上述描述,概念与关系相互增强的排序算法描述如下.

算法 1. 概念与关系相互增强的排序算法.

```
1. Iterate(S,A)
2. //S 为初始分布, A 为有向图的邻接矩阵
3. R0 ← S, W0 ← 0, k ← 0,
4. Loop:
5. //边权重更新
6. Σ ← ARk
7. For i ← 1, 2, ..., n
8.   For j ← 1, 2, ..., n
9.     If  $\sigma_j^k \neq 0$ 
10.       $w_{i,j}^{k+1} \leftarrow \frac{r_i^k}{\sigma_j^k}$ 
11.   EndIf
12. EndFor
13. EndFor
14. //向量重要性更新
15. Rk+1 ← Wk+1Rk
16. d ← ||Rk||1 - ||Rk+1||1
17. Rk+1 ← Rk+1 + dE
18.  $\delta \leftarrow ||\mathbf{R}_{k+1} - \mathbf{R}_k||_1$ 
19. k ← k + 1
20. While  $\delta > \epsilon$ 
21. Return (Wk, Rk)
```

算法 *Iterate*(**S**,**A**)由更新边权重和更新节点重要性两部分构成. σ_j^k 是指向*i*的所有概念*j*在第*k*次迭代后的重要性之和. 参数**S**为节点重要性的初始分布,一般设置为均匀分布. 输入参数**A**为定义2所述的有向图的邻接矩阵表示. 式(4)中的参数α在算法中以向量形式**E**表示,且其1-范数||**E**||₁=α,不考虑节点间差异的情况下,一般设置*E*为均匀分布. 0<ε<1 是用于确定收敛到不动点的阈值. 当计算收敛到不动点时,算法返回当前的节点重要性分布向量**R**_{*k*}和边权重矩阵**W**_{*k*}. 当需要对结果进行排序时,只需对**R**_{*k*}中的各分量排序,或对**W**_{*k*}按列对分量排序即可.

4 迭代收敛性证明

令|*V*|=*n*,第3节中的迭代式(3)和式(4)可以整理并重写为式(5):

$$r_i^{k+1} = \frac{1-\alpha}{n} + \alpha \sum_{t_i \in F_i} r_{t_i}^k \frac{r_i^k}{\sum_{t_j \in B_{t_i}} r_{t_j}^k} \tag{5}$$

其中 $k=0,1,2,\cdots,1\leq i\leq n$.

根据非线性数值分析理论^[13],当迭代序列(5)收敛到不动点 $\mathbf{R}^*=(r_1^*,r_2^*,\cdots,r_n^*)$ 时,算法1所描述的迭代算法才有意义. 为证明收敛性,首先给出如下引理.

引理 1. 式(3)和式(4)所构成的迭代过程是一个有限的非齐次马尔可夫过程,其转移矩阵 $\mathbf{P}=(p_{i,j})\in \mathbb{R}^{n\times n}$,可以表示为

$$p_{i,j} = \begin{cases} \alpha w_{i,j} + \frac{1-\alpha}{n}, & (i,j)\in E \\ \frac{1-\alpha}{n}, & \text{其它} \end{cases}.$$

证明. 由迭代公式可以很容易得证.

我们知道,PageRank 算法中所描述的随机冲浪(Random Surf)过程是一个齐次马尔可夫过程,仅由初始分布和一个转移概率矩阵决定. 其计算过程可以看作是一个矩阵特征向量求解的过程,从而保证了计算过程的收敛性^[15]. 而对于非齐次马尔可夫链而言,转移矩阵所表示的从*k*-1时刻的状态*i*转移到*k*时刻的状态*j*的概率依赖于*k*. 因此非齐次马尔可夫过程的收敛性证明与齐次马尔可夫过程不同. 对于本文所述迭代过程,我们首先研究其遍历性,然后通过非线性方程组的不动点理论证明其收敛性.

定义 5^[15]. 一个非齐次马尔可夫链是弱遍历的,如果对于所有的*m*都有 $\lim_{k\rightarrow\infty} \sup_{f^{(0)},g^{(0)}} \|f^{(m,k)}-g^{(m,k)}\|=0$, 其中 $f^{(0)}$ 和 $g^{(0)}$ 为初始向量.

引理 2^[15]. 随机矩阵**P**的遍历系数δ(**P**)定义为 $\delta(\mathbf{P})=\sum_{j=1}^n (\min_{1\leq i\leq n} p_{i,j})$. 有限马尔可夫链是弱遍历的,如果 $\sum_{m=0}^{\infty} \delta(\mathbf{P}_m)$ 的值是发散的.

定理 1. 式(3)和式(4)所构成的迭代过程是弱遍历的.

证明. 由引理1可知迭代过程是有限的马尔可夫过程,因此有

$$\delta(\mathbf{P}) = \sum_{j=1}^n (\min_{1\leq i\leq n} p_{i,j}) \geq \sum_{j=1}^n \frac{1-\alpha}{n} = 1-\alpha,$$

可得 $\sum_{m=0}^{\infty} \delta(\mathbf{P}_m) = \infty$.

由引理2可推知结论. 证毕.

定义 6^[13]. 设映像 $G:D\subset \mathbf{R}^n\rightarrow \mathbf{R}^n$,若存在

$\alpha \in (0, 1)$, 使得对任何 $x, y \in D_0 \subset D$, 恒有 $\|G(x) - G(y)\| \leq \alpha \|x - y\|$, 则称 G 为 D_0 上的压缩映像, α 为压缩系数. 若对任意的 $x, y \in D_0$, 有 $\|G(x) - G(y)\| \leq \|x - y\|$, 则称 G 为 D_0 上的非膨胀映像. 又若上式中当 $x \neq y$ 时, 严格不等式成立, 则称 G 在 D_0 上为严格非膨胀映像.

引理 3. 若 $G: D \subset \mathbf{R}^n \rightarrow \mathbf{R}^n$ 为有界闭集 $D_0 \subset D$ 上的严格非膨胀映像, $G(D_0) \subset D_0$, 则对任意 $x^0 \in D_0$, 序列 $x^{k+1} = G(x^k)$ ($k = 0, 1, \dots$) 在 D_0 内收敛于 G 的唯一不动点.

引理 3 称为 Edelstein 定理, 参见文献[13].

定理 2. 迭代序列

$$r_t^{k+1} = \frac{1-\alpha}{n} + \alpha \sum_{t_i \in F_t} r_{t_i}^k \frac{r_t^k}{\sum_{t_j \in B_{t_i}} r_{t_j}^k},$$

$$k = 0, 1, 2, \dots, t = 1, 2, \dots, n,$$

对任意初始近似

$$\mathbf{R}_0 \in D = \{\mathbf{R} = (r_0, r_1, \dots, r_n) \mid \|\mathbf{R}\|_1 = 1, 0 \leq r_i \leq 1, i = 1, 2, \dots, n\}$$

收敛于不动点 $\mathbf{R}^* \in D$.

证明. 令

$$g_t(\mathbf{R}_k) = \frac{1-\alpha}{n} + \alpha \sum_{t_i \in F_t} r_{t_i}^k \frac{r_t^k}{\sum_{t_j \in B_{t_i}} r_{t_j}^k},$$

$$s = 1, 2, \dots, n, t = 1, 2, \dots, n.$$

则原迭代序列具有 $\mathbf{R}_{k+1} = G(\mathbf{R}_k)$ ($k = 0, 1, 2, \dots$) 的形式.

由定理 1 指出的弱遍历性, 有

$$\lim_{k \rightarrow \infty} \sup_{\mathbf{R}_0, \mathbf{R}'_0} \|\mathbf{R}_k - \mathbf{R}'_k\| = 0,$$

即对于映射 $G: D \subset \mathbf{R}^n \rightarrow \mathbf{R}^n$ 而言, 对于任意的初始分布 \mathbf{R}_0 与 \mathbf{R}'_0 , 恒有 $\|G(\mathbf{R}_0) - G(\mathbf{R}'_0)\| \leq \|\mathbf{R}_0 - \mathbf{R}'_0\|$, 且当 $\mathbf{R}_0 \neq \mathbf{R}'_0$ 时, 不等式严格成立, 因此映射 G 在 $D_0 = [0, 1]$ 上为严格非膨胀映像, 且易得 $G(D_0) \subset D_0$.

由引理 3 可知 G 所构成的迭代序列有唯一的不动点 \mathbf{R}^* . 收敛于其唯一的不动点. 定理得证. 证毕.

5 实验分析

对基于相互增强特征来排序本体中的概念与关系的算法的实验分析可以从收敛性以及排序效果方面分别进行. 实验中我们实现并分别对比了三种链接分析排序算法: 本文所提出的算法(以下以 CARRank 代称)、标准 PageRank^[1] 算法(代表 OntoSelect, OntoKhoj, Swoogle 等利用 PageRank

的算法)、Reverse PageRank^[12] 算法以及修改过的 AKTiveRank 算法(称为 AKTiveRank*, 仅考虑其中的 DEM 与 BEM 两个信息方法, 其它的度量信息依赖于用户在查询本体时提出的查询语句). 对于引言中提及的 ObjectRank^[5], PopRank^[4] 和 Real-Walk^[6] 等其它算法, 由于其无法用于本体未知或缺少学习实例的情况, 因此实验中未进行考查和对比.

实验数据来源于 SchemaWeb^①. 其中收集了 RDFS, OWL 和 DAML+OIL 本体语言格式的本体 222 个(截至 2006-03-13). 其中包括: CC/PP, Dublin Core, FOAF, Topic Maps, vCard, WordNet 等得到广泛应用的本体; 也包括: DAML+OIL, OWL, RDF/S 等 W3C 语义 Web 规范中定义的基础本体. 实验环境为一台 Intel 2.66GHz Pentium IV CPU, 1GB DDR-SDRAM 内存, 80GB IDE 硬盘, 10MB/s 以太网卡的 PC 机. 操作系统为微软 Windows 2003 Server. 各算法中所需的因子 $\alpha = 0.85$. 当前后两次迭代的误差 $\epsilon \leq 1 \times 10^{-9}$ 时, 认为迭代过程收敛到不动点.

5.1 收敛性比较

第 4 节证明了相互增强特征排序本体中的概念与关系的迭代算法具有收敛性. 与其它链接分析排序算法的对比实验同样表明该迭代算法的可行性. 图 2 为针对 SchemaWeb 中收集的 Relationship^② 本体收敛性比较实验结果. Relationship 本体的有向图表示中包含 169 个节点, 252 条有向边.

由图 2 可见 CARRank 算法与 Reverse PageRank 算法具有相似的收敛速度. 原因在于两者在计算节点的重要性方面类似, 不同之处在于 Reverse PageRank 缺少通过计算有向边的权重来修正节点重要性的计算过程, 因此收敛速度要稍快.

CARRank 和 Reverse PageRank 算法, 由于考虑 Hub 值而不是 Authority 值作为节点重要性的度量, 因此在考虑链接关系时与 PageRank 算法的计算完全相反. 其收敛速度必然存在着差异. 本体不同, 其有向图表示也不同, 因此收敛速度也不同. 换言之, 在某些有向图情况下, PageRank 的收敛速度较快, 如图 2 所示; 而在某些情况下 CARRank 的收敛速度较快, 如图 3 所示的本体 UNSPSC 的有向图表示中包含 19600 个节点, 29386 个有向边. 实验中 CARRank 和 Reverse PageRank 拥有相同的收敛速

① <http://www.schemaweb.info/>

② <http://purl.org/vocab/relationship/>

度,且都比 PageRank 先达到阈值 1×10^{-9} . 但无论哪种情况, CARRank 和 Reverse PageRank 算法的收敛速度与 PageRank 都是可以接受的. 本体有向图的类型对算法收敛的影响将在后续研究中作进一步的阐述.

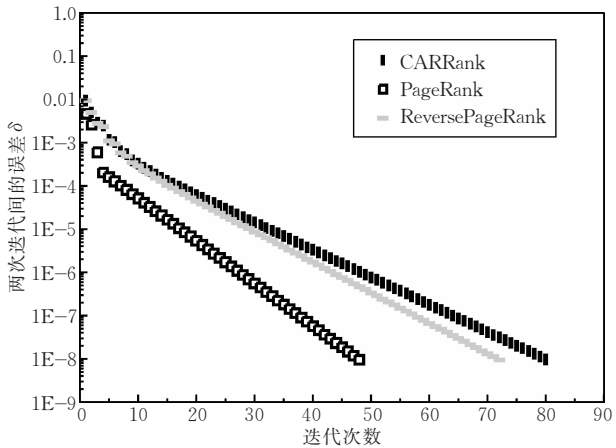


图 2 收敛性比较(本体 Relationship)

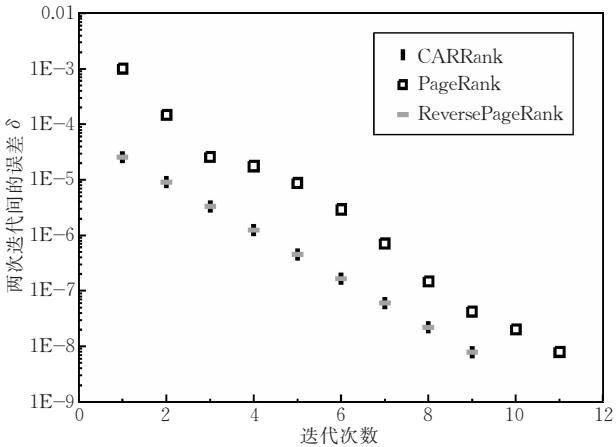


图 3 收敛性比较(本体 UNSPSC)

5.2 排序结果比较

针对 SchemaWeb 中收集的 222 个本体分别单独利用 PageRank, AKTiveRank*, Reverse PageRank 和 CARRank 4 种链接分析排序算法计算本体中概念的重要性排序. 作为对比,把本体作者对本体中概念和关系的重要性的评价作为标准答案. 另外还进行了用户实验(由 5 名研究兴趣为语义 Web 的学生参与,实验结果取平均值)进行了对比.

5.2.1 节以上述开源软件项目本体 (<http://keg.cs.tsinghua.edu.cn/project/software.owl>, 简称 Software 本体)为例展示了几种重要性排序方法的排序结果. 5.2.2 节则通过具体的评价指标来分析对比排序结果.

5.2.1 排序示例

表 2 中列出了各种算法对 Software 本体中的

概念进行重要性排序后的前 10 个概念,并与本体作者的评价以及用户实验结果进行了对比. 表 3 则列出了 CARRank 算法对 Software 本体中由概念(本体作者给出的前 5 个概念)出发的相关关系中的前 5 个重要性关系,并与本体作者的评价以及用户实验结果进行了对比. 表 2 和表 3 中,以本体作者的评价作为重要性排序的正确答案,并以黑体部分直观地表示各种方法与本体作者给出的正确答案之间的匹配程度. 注:这里 CARRank 的排序结果与第 3 节算法 1 的示例说明中的结果略有不同. 这是因为第 3 节算法 1 的示例说明所对应的图 1 仅是 Software 本体的局部. 另外,为了叙述方便第 3 节算法 1 的示例说明中仅选择本体中的类作为概念. 而实际应用中我们将本体中的类和属性均作为概念考虑.

由表 2 可见,对于 Software 本体而言,前 10 个重要的概念中,PageRank 有 5 个,AKTiveRank* 有 7 个,Reverse PageRank 有 6 个,CARRank 有 7 个,用户实验有 6 个与本体作者给出的正确答案相匹配. 此实验结果直观表明 CARRank 与 AKTiveRank* 具有相似的效果,而且要好于 PageRank, Reverse PageRank 以及用户实验的评价. 因此可以用于评价本体中的概念重要性,辅助用户理解本体作者的设计意图. 值得注意的是 CARRank 与 AKTiveRank* 都能够将概念“Project”评价为最重要的概念. 两者不同之处在于 AKTiveRank* 考虑了“Person”和“Project_Admin”,而 CARRank 考虑了“Release_Package”和“Help”. 事实上“Person”是“Member”,“Developer”和“Project_Admin”的基类,本体作者使用“Person”的意图只是为了更方便地定义“Developer”等子类,因此其重要性应该较低. PageRank 算法将“has_usage_statistics”和“statistics_bugs”等属性的重要性放在比较靠前的位置,这与本体作者的设计意图是不符的. Reverse PageRank 的评价结果要好于“PageRank”,但是它没有把“Project”概念的重要性排在最前面,由此可见使用 CARRank 算法可以通过概念与关系相互增强的方法改进 Reverse PageRank 算法,提高查准率. 由于 Software 本体的规模很小,用户可以通过较容易地阅读本体较好地理解本体作者的设计意图,因此用户实验的结果也较好,但是查准率仍然要低于 CARRank.

表 3 以本体作者的评价作为标准答案,仅对比了 CARRank 算法和用户实验结果. 其它算法因为不直接支持对关系重要性的排序功能所以没有进行

对比实验. “Project”, “Member”, “Developer”, “Category”和“Public_Forum”为本体作者给出的前 5 个重要的概念,对于每个概念 CARRank 算法和用户实验分别给出了排在前 5 位的重要的从相应概念出发的关系. 由表可见,除了 “Project”概念以外, CARRank 算法都能够比用户给出的排序结果更好地反映关系的重要性. 对于“Project”,我们分析

本体作者给出的排序结果,发现“title”, “summary”, “activity_ranking”和“project_homepage”等都是 DatatypeProperty 类型的属性. 由于 CARRank 算法的迭代过程仅作用于概念和关系,而没有考虑 DatatypeProperty 所指向的简单类型数据,所以 CARRank 无法将这些关系的重要性提高.

表 2 概念排序结果比较(本体 Software)

本体作者评价		PageRank	AKTiveRank *	Reverse PageRank	CARRank	用户实验
1	Project	Message	Project	Category	Project	Project
2	Member	has_usage_statistics	Usage_Statistics	Project	Usage_Statistics	Category
3	Developer	statistics_bugs	Developer	Usage_Statistics	Statistic_Record	Message
4	Category	statistic_record_support	Statistic_Record	Statistic_Record	Developer	Discussion
5	Public_Forum	Member	Member	Message	Category	Help
6	LastestNew	Project	Message	belong_to_category	Release_Package	Person
7	Message	Developer	Public_Forums	Help	Member	Member
8	Version	Category	Person	Public_Forums	Message	Developer
9	homepage	super_category	Category	Discussion	Help	Project_Admin
10	Usage_Statistics	page_views	Project_Admin	Developer	Public_Forums	Public_Forums

表 3 关系权重(本体 OWL)

概念	关系重要性排序			
	本体作者评价		CARRank	用户实验
Project	1	title	has_usage_statistics	project_homepage
	2	summary	developed_by	Title
	3	activity_ranking	Belong_to_category	activity_ranking
	4	project_homepage	Translations	has_public_forum
	5	project_of_statistic	Intended_audience	has_usage_statistics
Member	1	login_name	post_message	person_name
	2	publicly_displayed_name	site_member_since	
	3	email_address	login_name	
	4	user_id	email_address	
	5	site_member_since	publicly_displayed_name	
Developer	1	skills	member_of_project	person_name
	2	project_role	project_role	
	3		Skills	
	4		user_id	
	5			
Category	1	hasProject	hasProject	super_category
	2	category_name	sub_category	sub_category
	3	super_category	super_category	category_name
	4	sub_category	category_name	hasProject
	5			
Public_Forum	1	hasMessage	hasMessage	hasMessage
	2	belong_to_project		
	3	project_of_forum		
	4			
	5			

5. 2. 2 比较与分析

本节通过具体的度量指标进一步定量地对比分析 CARRank 与其它几种算法. 实验中我们选取如表 4 中的 4 个代表性的本体来作为概念与关系重要性评价的对象. 其中“OWL”是著名的本体描述语言,是一种元本体. “Software”本体是 5. 2. 1 节实验

中使用的本体,“Copyright”和“Travel”是两个比较复杂的本体. 实验过程与 5. 2. 1 节中所采用的方法相同.

我们采用使用前 20 查准率 $P@20$ 作为度量指标度量概念重要性排序结果. $P@20$ 的计算公式^[16]为

$$P@20 = \frac{n_{1\sim3} \times 20 + n_{4\sim10} \times 17 + n_{11\sim20} \times 10}{279},$$

其中, $n_{1\sim3}$ 表示排在前三位的概念中与正确答案中排在前 20 位的概念匹配的数量, 类似的 $n_{4\sim10}$ 和 $n_{11\sim20}$ 分别对应 4~10 位和 11~20 位. $P@20$ 值越高表明评价的效果越好.

表 4 实验中所用的本体

本体名称	URL
OWL	http://www.w3.org/2002/07/owl.rdf
Copyright	http://rhizomik.net/ontologies/2006/01/copyrightonto.owl
Software	http://keg.cs.tsinghua.edu.cn/project/software.owl
Travel	http://learn.tsinghua.edu.cn/homepage/2003-214945/travelontology.owl

此外, 我们设计了度量指标 P_R 作为度量从一个概念出发的所有关系的重要性的排序结果. 其公式如下:

$$P_R = \frac{\sum_{C_i} \left(\frac{\text{由 } C_i \text{ 出发的关系中的相关关系数量}}{5} \right)}{\text{相关概念的数量}},$$

P_R 值越高表明评价的效果越好.

实验中仍然以本体作者的评价为正确答案(对于“OWL”本体, 根据其作者的建议采用文献[17]中的信息作为正确答案).

图 4 显示了对于概念重要性排序的结果. 结果显示, CARRank 算法给出的排序结果在各种情况下都要好于用户实验的结果, 因此可以作为有效的评价本体中概念重要性的方法, 辅助用户理解本体作者的设计意图. 而且这种辅助效果随着本体的复杂程度增加而提高. 例如, 对于“Software”这一简单本体, CARRank 仅比用户实验的结果提高了 4%; 而对于复杂本体“Copyright”, 用户实验已经无法给出与本体作者正确答案相符的结果了, 此时 CARRank 算法的 $P@20$ 度量值仍高于 0.5. 相比之下, PageRank 算法对于大部分本体而言都不是理想的重要性排序算法. 而对于“OWL”和“Software”本体, CARRank 表现出了最好的效果. 对于“Copyright”本体, CARRank 与 AKTiveRank* 具有相同的较好效果. CARRank 算法仅在“Travel”本体中的效果差于 Reverse PageRank 算法. 其原因是本体作者在创作该本体时不仅遵从了本体设计意识流模型, 而且还采用了其它的本体设计模式, 因此无法仅通过 CARRank 算法得到最佳的概念排序结果.

表 5 是采用 P_R 度量值对比 CARRank 和用户

实验对于关系重要性评价的结果. 由于缺乏对比本体作者的评价, 仅在“Copyright”和“Software”两个本体上进行了实验.

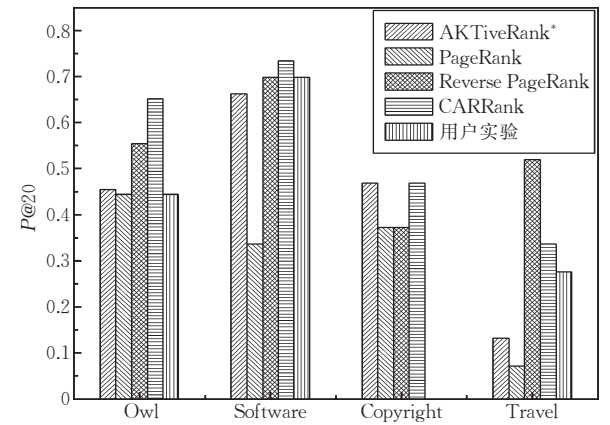


图 4 概念重要性排序比较

表 5 关系重要性排序比较

	CARRank	用户实验
Software	0.586	0.562
Copyright	0.06	0

给定一个本体, 给出由每个概念出发的一组关系中的重要性排序, 无论是对算法还是对于用户而言都是比较困难的. 实验结果显示, 对于简单本体“Software”, CARRank 算法的 P_R 值要高于用户评价结果的 P_R 值. 而对于复杂本体“Copyright”, 用户实验已经无法给出与正确答案相匹配的结果了, 但此时 CARRank 算法仍然可以给出一定正确的评价结果.

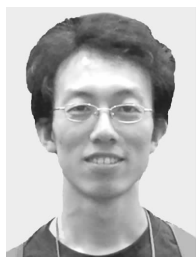
6 总 结

本体在很多领域都得到了广泛应用, 然而目前的研究中缺少能够合理并有效地对本体本身进行评价的方法. 本文通过分析本体设计的意识流模型, 提出利用概念与关系之间的相互增强特征迭代计算本体中概念的重要性和概念之间关系的权重排序. 理论证明算法收敛到迭代方程组的不动点. 收敛性实验验证了算法的可行性, 对真实世界中本体的评价结果验证了算法的有效性. 该算法不但适用于对本体本身的评价, 而且对于其它基于有向图链接分析的技术同样具有实际意义.

参 考 文 献

[1] Brin S, Page L. The anatomy of a large-scale hypertextual

- Web search engine. *Computer Networks*, 1998, 30(1-7): 107-117
- [2] Kleinberg J M. Authoritative sources in a hyperlinked environment. *Journal of ACM*, 1999, 46(5): 604-632
- [3] Lempel R, Moran S. The stochastic approach for link-structure analysis (SALSA) and the TKE effect. *Computer Networks*, 2000, 33(1-6): 387-401
- [4] Nie Z, Zhang Y, Wen J-R, Ma W-Y. Object-level ranking: Bringing order to Web objects//*Proceedings of the WWW*. Chiba, Japan, 2005: 567-574
- [5] Balmin A, Hristidis V, Papakonstantinou Y. Objectrank: Authority-based keyword search in databases//*Proceedings of the VLDB*. Toronto, Canada, 2004: 564-575
- [6] Geerts F, Mannila H, Terzi E. Relational link-based ranking//*Proceedings of the VLDB*. Toronto, Canada, 2004: 552-563
- [7] Ding L, Pan R, Finin T, Joshi A, Peng Y, Kolari P. Finding and ranking knowledge on the semantic Web//*Proceedings of the ISWC*. Galway, Ireland, 2005: 156-170
- [8] Buitelaar P, Eigner T, Declerck T. Ontoselect: A dynamic ontology library with support for ontology selection//*Proceedings of the Demo Session at the ISWC*. Hiroshima, Japan, 2004
- [9] Patel C, Supekar K, Lee Y, Park E K. Ontokhoj: A semantic Web portal for ontology searching, ranking and classification//*Proceedings of the WIDM*. New Orleans, Louisiana, USA, 2003: 58-61
- [10] Alani H, Brewster C, Shadbolt N. Ranking ontologies with Aktiverank//*Proceedings of the ISWC*. Athens, GA, USA, 2006: 1-15
- [11] James W. The principles of psychology. Harvard 1983 年版. Harvard, 1890
- [12] Fogaras D. Where to start browsing the Web? //*Proceedings of the IICS*. Leipzig, Germany, 2003: 65-79
- [13] Ortega J M, Rheinboldt W C. Iterative Solution of Nonlinear Equations in Several Variables. New York: Academic Press, 1970
- [14] Huang Xiang-Ding, Zeng Zhong-Gang, Ma Ya-Nan. The Theory and Methods for Nonlinear Numerical Analysis. Wuhan: Wuhan University Press, 2004(in Chinese)
(黄象鼎, 曾钟钢, 马亚南. 非线性数值分析的理论与方法. 武汉: 武汉大学出版社, 2004)
- [15] Shi Ren-Jie. Markov Chain and Its Application. Xi'an: Xidian University Press, 1992(in Chinese)
(施仁杰. 马尔科夫链基础及其应用. 西安: 西安电子科技大学出版社, 1992)
- [16] Leighton H V, Srivastava J. First 20 precision among world wide Web search services (search engines). *Journal of American Society for Information Science*, 1999, 50(10): 870-881
- [17] Wang T D, Parsia B, Hendler J. A survey of the Web ontology landscape//*Proceedings of the ISWC*. Athens, GA, USA, 2006: 682-694



WU Gang, born in 1978, Ph.D. candidate. His research interests include semantic Web data management.

ZHANG Kuo, born in 1980, Ph.D. candidate. His research interests focus on information extraction.

LI Juan-Zi, born in 1964, Ph.D., associate professor. Her research interests include semantic Web, Web service, text mining and knowledge discovery.

WANG Ke-Hong, born in 1941, professor, Ph.D. supervisor. His research interests include network computing and knowledge engineering.

Background

This paper discusses the importance of concepts and relations ranking in the field of the ontology evaluation. To the best of our knowledge, there is no specialized evaluation approach proposed for ranking the importance of concepts and relations. This paper attributes to the project SWARMS (Semantic Web Aiding Rich Mining System), which is supported by the National Natural Science Foundation of China (NSFC) under grant No. 90604025. The project mainly aims

to study the information extraction, ontology annotation, ontology matching, data management of semantic Web, and data mining on semantic Web. The website of the project is <http://keg.cs.tsinghua.edu.cn/project/pswmp.htm>. There are several publications in this project published on AAAI 2005 Workshop, WWW 2005 Poster track, APWEB 2006 and ASWC 2006 Demo Track.