软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(9):2075−2088 [doi: 10.3724/SP.J.1001.2011.03876]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于语义的主题爬行策略*

叶育鑫[1,2], 欧阳丹彤[1,2+]

[1](吉林大学 计算机科学与技术学院,吉林 长春 130012)

[2](吉林大学 符号计算与知识工程教育部重点实验室,吉林 长春 130012)

Semantic-Based Focused Crawling Approach

YE Yu-Xin[1,2], OUYANG Dan-Tong[1,2+]

[1](School of Computer Science and Technology, Jilin University, Changchun 130012, China)

[2](Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China)

+ Corresponding author: E-mail: ouyd@jlu.edu.cn

Ye YX, Ouyang DT. Semantic-Based focused crawling approach. *Journal of Software*, 2011,22(9): 2075−2088.
http://www.jos.org.cn/1000-9825/3876.htm

Abstract: An approach of semantic-based focused crawling is proposed in order to use semantic resource efficiently. In this paper, a domain-ontology is used to describe the topic of Web crawling. Lexicon of the keywords list are mapped to ontology, and semantic of words are obtained through mapping. Inference services about assertion set expanding and domain-range relation are defined. The semantic relation among keywords can be inferred by inference services. At the same time, the definition of concept about Web page is given. A semantic computational model is proposed by combining inference services mentioned above. In the end, the order of URLs corresponding to their Web page is decided according to the subsumption of topic concepts. The result show that this approach is advanced in harvest-rate and crawling efficiency and is better than some classical algorithms.
Key words: ontology; semantic Web; focused crawling; Tableau calculus

摘 要: 为使主题爬行能够充分利用资源的语义信息,提出基于语义的主题爬行策略.该策略利用领域本体刻画爬行主题,将本体语义映射到关键词表.通过定义断言集一致性扩展和域值关联推理任务,推演关键词间语义关系.在定义网页主题概念的基础上,结合本体推理方案提出主题概念的语义叠加效应模型.最后,利用主题概念的语义包含关系判定 URLs 抓取顺序.实验结果表明,该语义主题爬行策略在抓取收获率和爬行效率上优于现有同类方法,该方案有效、可行.
关键词: 本体;语义 Web;主题爬行;Tableau 演算
中图法分类号: TP181 **文献标识码**: A

* 基金项目: 国家自然科学基金重大项目(60496320, 60496321); 国家自然科学基金(60873148, 60973089); 吉林省科技发展计划(20080107); 欧盟合作项目(155776-EM-1-2009-1-IT-ERAMUNDUS-ECW-L12); 符号计算与知识工程教育部重点实验室开放基金(450060326019)

收稿时间: 2009-05-13; 修改时间: 2009-08-26; 定稿时间: 2010-04-21

互联网以其独特的开放性蕴含了海量的信息资源,它允许人们在世界的任何角落共享资源.在美国有 84%的成年用户习惯于使用搜索引擎在线查找信息[1].据统计,每天有超过 6 千万用户发送 2 亿多条查询给搜索引擎请求信息检索,使 Web 搜索成为仅次于电子邮件的第二大互联网行为活动[2].网络资源的指数级增长以及环境的愈加复杂,给 Web 搜索技术带来了新的课题和挑战[3].主题爬行是对互联网中特定相关领域网页获取的关键技术.它在传统搜索引擎技术中引入文本分类、聚类以及 Web 挖掘等相关技术捕获优化特定主题的网页信息,提高现有搜索的精度,降低搜索引擎对网络资源的占用,缩短网页数据库更新的周期.

最早的主题爬行技术是由 De Bra[4]提出的 Fish-search 方法,该方法通过简单地匹配关键词来判定网页的相关性.Chakrabarti 等人[5]在其文章中说明了主题爬虫的一般体系结构和功能.McCallum 等人[6]采用贝叶斯网分析超链接,通过分析超链接结构对 URLs 排序.Rennie 等人[7]提出了使用加强学习的方法建立专门域的主题爬行策略,通过计算每个链接的奖惩值来训练网络爬虫,从而进行抓取的在线调整.Hsu 等人[8]通过结合加强学习与语境图(context graph)进一步提高抓取效果.彭涛[9]系统分析了如何有效应用粒子群算法构建主题爬虫中的分类器.这些利用经典的数据挖掘和机器学习技术来解决主题爬行问题的研究,已经有较长的历史和颇丰的成果.如何突破传统的统计分析方法研究主题爬行成为新的热点和难点.Ehrig 在他的工作中[10]首次将本体引入到主题爬行,利用本体代替传统主题爬行中的主题词库,利用本体中词汇的层次关系给出一个简单的网页的主题相关度计算方法.他的工作为解决主题爬行问题开启了新的思路.随后,Ganesh[11]在他的爬行算法中也引入过本体,但不是用于描述爬行主题,而是用于综合 PageRank 等相关链接结构分析技术计算链接所在网页的重要度.王辉等人[12]在主题爬行中计算文档特征权重时,利用本体获取根集(root set)文档的质心向量.可以说,本体丰富的语义内涵和严谨的逻辑结构是主题词库的优秀替代者.然而,如何充分利用本体的语义和逻辑结构搭建主题相关度计算模型仍然是一个开放性课题.本文在利用本体描述主题词库的基础上定义了主题爬行下的相关本体推理任务,并结合相关推理任务给出网页主题概念的语义计算模型和概念间包含关系推演,提出基于语义的主题爬行策略.Tableau 演算是关于本体推理问题的成熟解决方案,我们在语义主题爬行策略中采用 Tableau 演算处理相关推理任务.

本文首先介绍了主题爬行的研究意义,分析了现有工作的研究概况和存在问题,并针对需要解决的问题简要阐述本文的主要工作.第 1 节给出语义主题爬虫的基本框架,并说明该框架下各部分的职能和相互关系.第 2 节给出主题爬行本体的形式化定义,同时在对网页做预处理后建立网页关键词表与本体的映射关系,生成关键词映射表.第 3 节定义本体主题爬行下的推理任务并给出相关推理任务的解决方案;进一步地,结合相关推理任务给出基于语义的主题爬行策略.第 4 节给出的相关评测说明了该方法有效可行.最后是工作总结.

1 语义主题爬虫的基本框架

如何通过已下载的网页对其出链接所指向的网页主题进行准确的预判断,是主题爬行搜索策略的关键.与现有方法不同,我们通过计算语义来判断网页的主题,为此设计出基于语义的主题爬虫,其基本框架如图 1 所示.

在该语义主题爬虫的基本框架中,我们通过构建领域本体来刻画爬行主题;通过解析并统计词频获得网页的关键词表;通过语义扩展词库扩展本体词汇;通过简单的词匹配建立从关键词表到本体的映射;通过语义概念计算模型计算网页的主题;通过计算网页主题包含关系预测待抓取超链接的主题相关度并排序抓取.

为有效实现主题爬行,我们在该框架下定义了一系列结构,其中包括网页结构、本体结构、关键词表、语义扩展词库、关系词映射表、网页主题概念等.它们用于存储主题爬行过程中的相关初始信息、中间结果和最终结果.进一步地,我们将整个框架所实现的功能细化为预处理层、映射层、语义概念计算层和相关度排序层.其中:预处理层负责计算待处理网页(或初始网页)的关键词表,构建领域本体对应的语义扩展词库;映射层负责建立关键词表中的关键词与本体中的概念、属性和实例的对应关系,将关键词表中的词映射到本体中;语义概念计算层负责将映射到的概念、属性和实例集合计算成网页的主题概念;最后,相关度排序层负责计算每个主题概念的主题相关度大小并对 URLs 排序,语义主题爬虫将按 URLs 的主题相关度大小顺序去互联网爬行抓取相关网页.

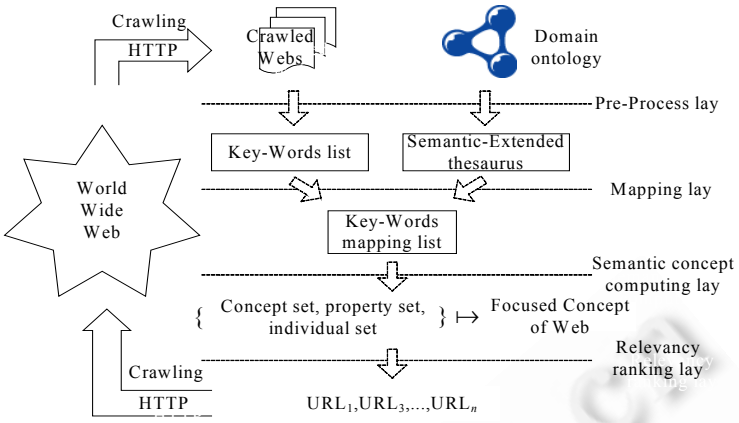


Fig.1 Framework of approach of semantic-based focused crawling

图 1 语义主题爬行策略框架

2 主题爬行本体及其映射

本节介绍刻画主题的领域本体和关键词映射表的构建.首先给出主题爬行本体的形式化定义,在此基础上,进一步阐述如何利用主题爬行本体对经过网页预处理的关键词表构建关键词映射表.

2.1 主题爬行本体

本体是现实世界某方面的一个抽象模型的形式化规范说明[13],其定义中的抽象模型包括进行领域知识建模的概念框架、互操作 agent 之间进行交流的内容明确定义以及表达的协定.领域本体是指以一个特定领域为描述对象的本体.在基于语义的主题爬行策略中,我们采用了领域本体刻画爬行主题.与以往的关键词集合描述主题不同,领域本体提供对领域知识的共识理解,确定领域内共同认可的词汇,并从不同层次进一步明确定义这些词汇间的相互关系.一般来说,一个领域本体由概念(\属性)、概念(\属性)包含关系和实例构成.其中,概念(\属性)又有原始和复杂之分,复杂概念(\属性)是通过构造算子对原始概念复合而成.本文中刻画爬行主题的本体具有同样的成分,下面给出其形式化定义.

定义 1(主题爬行本体). 主题爬行本体是解决主题爬行问题时,用于刻画爬行主题的一个领域本体.它是形如 O_{topic}:={{C},{R},{In},$TBox$,$ABox$}的五元组.

其中:C 是与主题相关概念,{C}表示以 C 为元素构成的概念集合;R 是主题相关属性,{R}表示以 R 为元素构成的属性集合;In 是主题相关实例,{In}表示以 In 为元素构成的实例集合;$TBox$ 是形如 $C\sqsubseteq D$ 的包含公理集合(其中,$C,D\in\{C\}$);$ABox$ 是形如 $R(a)$或 $R(a,b)$的断言集合(其中,$R\in\{R\}$,$a,b\in\{In\}$).

本体的表述语言并不唯一,一般采用一阶逻辑、框架逻辑等各种逻辑形式表示.语义 Web 研究中,关于本体 Web 语言标准的发布,为本体知识表示提供了新的途径.与传统的知识表示方法相比,互联网的统一资源标识(URIs)机制使得 Web 上的知识表示能够有效避免自然语言的二义性.另外,互联网的超文本链接机制暗合知识表示中的语义关联.这一切都为在互联网上形式化本体知识表示语言提供了必要条件,其语言规范 OWL[14]已经成为 W3C 组织的 Web 标记语言标准.OWL 语言兼顾了互联网资源标注和人工智能领域本体的知识表示两者的特性,将二者有机融合在一起.从知识表示的角度来看,OWL DL 是以基于 $SHOIN(D)$的描述逻辑语言为逻辑框架的一个 OWL 语言子集.与描述逻辑框架下的知识系统类似,OWL DL 语言所表示的本体一般分为断言集合($ABox$)和术语集合($TBox$)两个部分.断言集合中存在一元、二元谓词断言,术语集合存放包含公理和等价公理.为进一步丰富 Web 下本体知识的表达能力,目前,研究者们也在做本体和规则的整合[15],使得本体的知识公理系统下不仅可以有包含公理和等价公理,还可以有规则形式的公理存在.但这种整合带来了语言的不可判定性等

相关问题.本文中,我们采用 OWL DL 语言描述主题爬行本体,使用 OWL DL 语言既保证了语言的可判定性,同时又能够有效利用现有主流推理机(如 Pellet,FaCT++,Racer 等)的推理服务.

定义 2(*ABox*). 在主题爬行本体 O_{topic} 中存在互不相交的 3 个集合:概念集合{C}、属性集合{R}和实例集合{In},我们称形如 C(a)的形式为概念断言;形如 $R(a,b)$的形式为属性断言;由概念断言与属性断言构成的有限集合被称为断言集合,即 $Abox$.其中,$a,b\in\{In\}$,$C\in\{C\}$,$R\in\{R\}$.

本体的语义是通过解释来体现的.一般来说,解释 $\boldsymbol{I}$ 由解释域Δ^I和解释函数$\bullet^I$构成.映射函数$\bullet^I$映射每一个实例 $In\in\{In\}$为解释域Δ^I下的一个元素,使得不同实例对应的元素不同,记为 In^I;映射每一个概念 $C\in\{C\}$到解释域 Δ^I 上的一个子集,记为 C^I;映射每一个属性 $R\in\{R\}$到$\Delta^I\times\Delta^I$笛卡尔乘积的一个子集,记为 R^I.特别地,规定符号顶层概念⊤和底层概念⊥的解释分别为Δ^I和∅.其中需要说明的是:原始概念和原始属性通过解释函数直接映射;复杂概念和复杂属性通过对原始概念和属性的解释后,应用构造算子复合得到最终解释.

定义 3(*ABox* 的解释). 一个解释 $\boldsymbol{I}$ 满足一个概念断言 $C(a)$,当且仅当 $a^I\in C^I$;一个解释 $\boldsymbol{I}$ 满足一个属性断言 $R(a,b)$,当且仅当$(a^I,b^I)\in R^I$.$\mathcal{A}$ 是一个断言集合,我们说解释 $\boldsymbol{I}$ 是 $ABox$ $\mathcal{A}$ 的模型,当且仅当 $\boldsymbol{I}$ 满足 $\mathcal{A}$ 中的所有断言.

定义 4(*TBox*). 对主题爬行本体 O_{topic} 中的概念集合{C}内任意两个概念 C 和 D,称形如 $C\sqsubseteq D$ 的为包含公理,形如 $C\equiv D$ 的为等价公理;由包含公理和等价公理构成的有限集合为公理集合,即 $TBox$.

在定义 4 中,等价公理可以等价转换成两个包含公理,即 $C\equiv D\Leftrightarrow(C\sqsubseteq D)\sqcap(D\sqsubseteq C)$.因此,我们也可以把 $TBox$ 看做是关于包含公理的有限集合.

定义 5(*TBox* 的解释). 一个解释 $\boldsymbol{I}$ 满足一个包含公理 $C\sqsubseteq D$,当且仅当 $C^I\subseteq D^I$;一个解释 $\boldsymbol{I}$ 满足一个等价公理 $C\equiv D$,当且仅当 $C^I=D^I$;$\mathcal{T}$是一个公理集合.我们说解释 $\boldsymbol{I}$ 是 $TBox$ 的模型,当且仅当 $\boldsymbol{I}$ 满足 $\mathcal{T}$中所有公理.

2.2 建立URLs到本体的映射

实现主题爬行的根本途径是对待抓取的 URLs 进行排序预测.URLs 本身可以利用预测的信息非常有限,而这些超链接所在的网页中有大量的信息可供使用.所以,通常进行主题爬行时,要首先处理 URLs 所对应的相关网页的文本信息.

在网页预处理过程中,我们采用了广泛应用于自由文本处理的浅层文本预处理技术[16].一般认为,若一个词在该文本中出现的频率高,而相对在其他文本中出现的频率低,则该词能够很好地反映该文本主题内容.在我们的主题爬行中,将网页中拥有这种性质的词汇称为该网页的关键词.通过统计分析处理,按频度比高低确定该网页的关键词.在频度比值的限定范围内获得的关键词及其频度的集合,称为网页的关键词表.

为让我们的爬行算法能够有效识别关键词表中的关键词语义,需要建立关键词表与主题爬行本体的映射.在建立映射之前,定义了一个本体词汇的扩展词表.理论上,我们希望定义的扩展词表越全面越好,这样能够保证关键词表对爬行主题本体的最一般映射.其中,我们定义一个成功的关键词映射,如下:

定义 6(关键词映射). 关键词映射 f 是从网页关键词到主题爬行本体的对应关系,映射 f 的定义域是关键词,值域是本体中与之相匹配的概念、属性或实例.若存在一个关键词 k 同时出现在映射 f 的定义域和值域中(即同时在关键词表和本体或其语义扩展词库中),则称关键词映射 f 为一个成功映射.

定义 7(关键词映射表). 通过关键词表映射,得到的关键词与本体中概念集合、属性集合和实例对应关系的集合,称为关键词映射表.

3 语义叠加效应模型与相关度排序

首先给出语义主题爬行的核心算法——语义叠加效应模型.该模型中定义了主题爬行下的断言集一致性扩展和域值关联两项推理任务,通过执行相关推理方案实现语义叠加效应模型.在此基础上,利用主题概念包含关系判定主题相关度,实现基于语义的主题爬行.

3.1 语义叠加效应模型

通常,一个被统计的关键词在本网页中出现频率越高,相对地其他网页中出现频率较低,则能够很好地代表该网页主题.在我们的爬行策略中,关键词映射表将从一个网页中统计出的关键词对应成本体中相关的概念、属性和实例,它们依靠主题爬行本体构建了语义关联.利用主题爬行本体内在的语义关联,可以有效计算一个网页的语义主题.在我们的爬行策略中,一个网页的主题定义如下:

定义 8(网页的主题概念). 根据网页的关键词映射表,能够找出的最大限度反映该网页主题的本体概念,是该网页的主题概念,记为 C_{rel}.

为合理计算一个网页的主题概念,我们提出了语义叠加效应计算模型.基本思想是:首先通过断言集合分别建立实例到属性和概念的转换,然后进一步建立属性到概念的转换.这种转换将实例和属性对主题的效应归约成概念对主题的效应,我们将其形象地称为语义叠加效应.现有的本体推理方法——Tableau 演算就是一种成熟的推理方案,它为合理构建这种语义叠加效应模型提供了坚实的理论基础和有效的计算途径.我们应用 Tableau 演算实现实例到属性、概念以及属性到概念的映射,将一个网页的关键词映射表转化成一个带词频统计的概念集合.在语义叠加效应处理后,得到关于该网页的一个更新扩展概念词频对集合.进一步地,选择一种计算模式从这个概念集合中确定该网页的语义主题概念.这里,我们定义峰值语义效应(即选择词频最高的概念)计算该网页的主题概念.该模型的输入是概念、属性和实例的集合及其词频.模型的输出是网页主题概念 C_{rel},其具体算法如下.关于算法中的相关推理定义和证明(如:衍生规则、一致性判定等),我们将在下一节进一步给出.

算法. 主题概念语义叠加效应模型.

输入:关键词表 K_{page},主题爬行本体 O_{topic};

输出:网页主题概念 C_{rel}.

Step 1. 初始化 K_{page},由关键词名称 *name* 和频度 *frenq* 组成键值对,构成实例(\属性\概念)频度集合,分别记为 $indi:=\{name^{indi},frenq^{indi}\}$,$prop:=\{name^{prop},frenq^{prop}\}$,$cls:=\{name^{cls},frenq^{cls}\}$;根据 O_{topic} 和 K_{page} 构造断言集合 $\mathcal{A}$,通过添加主题爬行本体中的相关概念和属性更新关键词表 K_{page};

//计算实例频度集合 *indi* 的语义叠加效应

Step 2. 对一致的断言集合 $\mathcal{A}$,选择任意可用 tableau 衍生规则扩展 O_{topic} 中的断言集 $\mathcal{A}$,得到断言集 $\mathcal{A}'$;

Step 3. 令 $\mathcal{A}=\mathcal{A}'$,若对 $\mathcal{A}$ 还存在其他衍生规则可用,返回 Step 2,否则执行下一步;

Step 4. 对$\forall i.\ name_i^{indi}\in name^{indi}$,令 m=1,使得 $C_m\in\{C\}$;若 $i>|name^{indi}|$,转至 Step 6;

Step 4.1. 若 $m<|\{C\}|$,则执行下一步,否则转至 Step 5;

Step 4.2. 若 $C_m(name_i^{indi})\notin ABox'$,则令 $m=m+1$,转至 Step 4.1;

Step 4.3. 若 $C_m\in name^{cls}$,则定义函数 $add(frenq_i^{indi},frenq_{C_m}^{cls})$求和,替换关键词表 K_{page} 中概念 C_m 的频度;否则有 $C_m\notin name^{cls}$,定义函数 $create(C,frenq_i^{indi})$创建新的概念频度键值对,添加到 *cls* 中.令 $m=m+1$,并转至 Step 4.1;

Step 5. 对$\forall j.(i<j<|name^{indi}|)$,令 n=1,使得 $R_n\in\{R\}$;

Step 5.1. 若 $n<|\{R\}|$,则执行下一步,否则转至 Step 4;

Step 5.2. 若 $R_n(name_i^{indi},name_j^{indi})\notin ABox'$,则令 $n=n+1$,转至 Step 5.1;

Step 5.3. 若 $R_n\in name^{prop}$,则定义求和函数 $add(getMinFrenq(frenq_i^{indi},frenq_j^{indi}),frenq_{R_n}^{prop})$替换关键词表 K_{page} 中属性 R_n 的频度;否则有 $R_n\notin name^{prop}$ 成立,定义函数

$$create(R_n,getMinFrenq(frenq_i^{indi},frenq_j^{indi}))$$

创建新的属性频度键值对添加到 *prop* 中.令 $n=n+1$,并转至 Step 5.1;

//计算属性频度集合 *prop* 的语义叠加效应

Step 6. $\forall k.(name_k^{prop}\in name^{prop})$,若 $k\leqslant|name^{prop}|$,则任取变量 p,q,使得

$\forall p\forall q.((name_p^{cls} \in name^{cls})\wedge(name_q^{cls} \in name^{cls}))$.

其中,$1\leqslant p,q\leqslant|name^{cls}|$且 $p\neq q$.构建公理$\exists name_k^{prop}.(name_q^{cls}) \sqsubseteq name_p^{cls}$;否则,转至 Setp 9;

Step 7. 对所构建的公理取反,得$\neg(name_p^{cls} \sqsubseteq \exists name_k^{prop}.(name_q^{cls}))$;

应用 Tableau 演算,计算$\{\neg(name_p^{cls} \sqsubseteq \exists name_k^{prop}.(name_q^{cls}))\}\cup TBox$ 的一致性;

Step 8. 若术语集一致,则表明属性名 $name_k^{prop}$ 与概念 $name_p^{cls}$ 和 $name_q^{cls}$ 间存在属性约束关系,分别定义求和函数 $add(frenq_k^{prop},name_p^{cls})$ 和 $add(frenq_k^{prop},name_q^{cls})$,替换关键词表 K_{page} 中 cls_p 和 cls_q 的频度值;否则,表明它们之间不存在约束关系,返回 Step 6.

//计算概念频度集合 *cls* 的语义峰值效应

Step 9. 经过实例频度集合、属性频度集合的语义叠加效应处理,得到更新扩展后的概念频度集合 *cls*′. 初始化 l=1,C_{rel}=∅, $C_{rel}^f=0$ (其中, C_{rel}^f 定义为主题概念 C 的词频值);

Step 10. 取 cls_l' 的概念名 $name_l^{cls'}$ 赋值给 C_{rel},取 cls_l' 的词频 $frenq_l^{cls'}$ 赋值给 C_{rel}^f,即 $C_{rel}=name_l^{cls'}$, $C_{rel}^f=frenq_l^{cls'}$,同时令 $l=l+1$;

Step 11. 若 $l>|cls_l'|$,则返回 C_{rel},否则比较 C_{rel}^f 与 $frenq_l^{cls'}$ 的大小;

Step 12. 若 $C_{rel}^f < frenq_l^{cls'}$,直接返回 Step 10;否则,令 $l=l+1$,返回 Step 11.

3.2 模型的计算实例

为直观地说明如何利用语义叠加效应模型计算网页的主题概念,下面以描述“S 三星”股票和“老凤祥 B”股票的新闻网页为例阐述算法的执行过程.截取和待测网页相关的主题爬行领域本体片段如图 2 所示:在 *TBox* 中,“深圳 A 股”和“上海 B 股”分别是“普通股”和“特种股”的子类,并且“普通股”和“特种股”不相交;“普通股”和“特种股”又同是“股票”的子类;“股票”和“股份有限公司”之间有“股票所属公司”的属性关联.在 *ABox* 中,“S 三星”和“老凤祥 B”分别是概念“深圳 A 股”和“上海 B 股”的实例,“赛格三星股份有限公司”和“老凤祥股份有限公司”均是概念“股份有限公司”的实例.语义效应叠加模型就是要通过待测网页的高频词表,在主题爬行领域本体中选择最能代表该网页语义描述的概念.

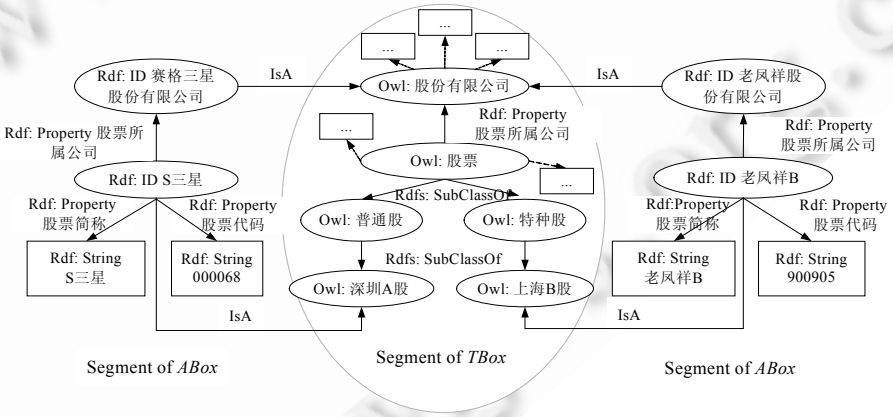


Fig.2 Segment of financial domain ontology

图 2 金融领域本体片段

为简明起见,仅取输入相关的前 5 个高频词:{S 三星,35}$_{indi}$、{老凤祥 B,31}$_{indi}$、{赛格三星股份有限公司,20}$_{indi}$、{老凤祥股份有限公司,17}$_{indi}$、{深圳 A 股,10}$_{c}$.其中,前 4 个高频词是主题爬行本体中出现的实例,最后一个词是本体中的概念(见表 1).

Table 1 Key-Words list of input stream

表 1 输入的关键词列表

Key-Words	Frequency	Type	Symbol
S 三星	35	*indi*	*a*
老凤祥 B	31	*indi*	*b*
赛格三星股份有限公司	20	*indi*	*c*
老凤祥股份有限公司	17	*indi*	*d*
深圳 A 股	10	*concept*	C_1

计算实例频度模型的叠加效应.根据初始关键词表和主题爬行本体可构造断言:

$$\mathcal{A}:=\{C_1(a);C_2(b);C_3(c);C_3(d);R_1(a,c);R_1(b,d)\},$$

其中,C_2 是“上海 B 股”,C_3 是“股份有限公司”,R_1 是“股票所属公司”.应用 Tableau 演算对 $\mathcal{A}$ 进行扩展,满足$\sqsubseteq$rule 扩展规则(定义 11),扩展断言$(\neg C_1 \sqcup C_4):a$ 更新 $\mathcal{A}$ 为 $\mathcal{A}'$,其中,C_4 代表主题爬行本体中的概念“普通股”.继续应用 $\sqcup$rule 扩展规则扩展得到 $C_4(a)$,更新 $\mathcal{A}'$.如此反复,直到没有可用规则为止,利用 Tableau 扩展 $\mathcal{A}$ 完成,补充的断言集合:$\{C_4(a);C_6(a);C_5(b);C_6(b)\}$,其中,$C_5$ 为概念“特种股”,C_6 为概念“股票”.根据算法更新关键词列表,其中:概念 C_1 是原词表中已有的概念,需要叠加词频;而 C_2~C_6 是原词表中不存在的概念,需新创建并继承对应实例的词频.根据算法继续计算实例对属性的语义效应,设 R_1 代表本体中的属性“股票所属公司”.用初始关键词列表中的实例集合对照主题爬行本体可得 $R_1(a,c)$和 $R_1(a,d)$.R_1 不存在于初始关键词列表,需要创建并取相关实例集合 $\{a,c,d\}$中词频最小的赋值给 R_1.经过实例频度叠加后,更新的关键词表见表 2.

Table 2 Updated key-words list

表 2 补充及变更词列表

Complementary word	Frequency	Type	Symbol
深圳 A 股	45	*concept*	C_1
上海 B 股	31	*concept*	C_2
股份有限公司	37	*concept*	C_3
普通股	35	*concept*	C_4
特种股	31	*concept*	C_5
股票	66	*concept*	C_6
股票所属公司	17	*property*	R_1

接下来考察断言集合 $\mathcal{A}$ 中的属性对概念的语义叠加效应.$\mathcal{A}$ 中涉及到的相关概念和属性有 C_1,C_2,C_3 和 R_1,构建形如 $C_1 \sqsubseteq \exists R_1 \cdot C_2$ 的包含公理.应用域值关联推理,在 TBox 一致性检测中发现:对于属性应该存在一个实例 x,同时满足概念 C_2 和 C_3,其中,满足 C_2 由 $R_1 \cdot C_2$ 得到,满足 C_3 由本体中定义的 R_1 值域限定得到.根据进一步利用本体中对概念 C_2 和 C_3 的定义进行 Tableau 演算扩展,可以得到 C_2 和 C_3 是两个不相交的概念.故公理$\exists R_1 \cdot C_2 \sqsubseteq C_1$ 不成立,C_1 和 C_2 之间不存在 R_1 关系.同理可判定$\exists R_1 \cdot C_1 \sqsubseteq C_2$,$\exists R_1 \cdot C_3 \sqsubseteq C_2$,$C_1 \sqsubseteq \exists R_2 \cdot C_3$ 不成立,以及$\exists R_1 \cdot C_3 \sqsubseteq C_1$,$\exists R_1 \cdot C_3 \sqsubseteq C_2$ 成立.所以有 C_1 和 C_3(即概念“深圳 A 股”和“股份有限公司”)、C_2 和 C_3(即概念“上海 B 股”和“股份有限公司”)间存在 R_1(即“股票所属公司”)的域值关联关系,需要进行属性的语义效应叠加.经过属性效应叠加后,得到的所有概念词频表见表 3.

最后比较词表中所有概念的词频,选择其中词频最高的概念“股票”作为该网页的主题概念.如果将初始输入的词表中的“老凤祥 B”变为“深圳 A 股”下的一个股票实例“中水渔业”,其他信息,包括“中水渔业”的词频也仍然是 31 保持不变.重复上述计算过程,在实例效应叠加后“深圳 A 股”的词频变更为 35+31+10=76;属性效应叠加后“深圳 A 股”的词频变更为(76+17=93)>66.此时,网页的主题概念变为“深圳 A 股”.可见,我们的模型在高频词既有普通股又有特种股时,能够计算出网页主题为“股票”;而当主题全是普通股时,能够定位到更精确的主题,即“股票”的子概念——“深圳 A 股”.

Table 3 Concept-Frequency words list

表 3 概念频度集合表

Candidate topic word	Frequency	Type	Symbol
深圳 A 股	52	*concept*	C_1
上海 B 股	48	*concept*	C_2
股份有限公司	37	*concept*	C_3
普通股	35	*concept*	C_4
特种股	31	*concept*	C_5
股票	66	*concept*	C_6

3.3 模型中的相关推理

本节对语义效应模型中的相关推理任务进行定义和命题证明.首先给出实例语义叠加效应中断言集一致性扩展的推理任务定义及其相关定理证明,接下来在属性语义效应叠加中定义了一个新的推理任务——域值关联推理,它将属性语义效应叠加转换为本体下的域值关联推理任务的推演.下面分别介绍断言集显示化推理任务和域值关联推理任务.

3.3.1 一致性扩展推理

本文将本体中断言集的一致性扩展推理任务应用于主题爬行领域.关于断言集的一致性扩展,Martin 等人[17]的工作有类似的处理.随后,Levy 等人[18]在他们的 CARIN 系统中定义了一个约束系统,该约束系统实现的功能亦是关于断言集显示化的推理任务.我们从实例语义叠加效应的预处理需求角度重新定义了断言集显示化推理任务,并进一步给出相关定理及其证明,保证该推理任务的有效执行.首先给出断言集 $\mathcal{A}$ 一致性的定义.

定义 9(一致性). 如果一个断言集合 $\mathcal{A}$ 存在一个解释 $\boldsymbol{I}$ 是 $\mathcal{A}$ 的模型,则称断言集合 $\mathcal{A}$ 是一致的.

定义 10(显示化预处理). 在语义主题爬行的实例语义叠加效应中,如果能够将本体 O_{topic} 下断言集合 $\mathcal{A}$ 的所有隐含知识显示化,并保证扩展后得到的断言集合 $\mathcal{A}'$是一致的,我们称这样的推理过程为实例语义叠加的显示化预处理.

定义 11(衍生规则). 在显示化预处理中,为推出所有隐含知识而应用的规则被称为衍生规则.

一般为保证衍生结果的正确性,通常采用Tableau演算中的规则作为扩展处理的手段.在以描述逻辑语言表述的本体下,Tableau 演算中规则的定义和数目取决于语言的表述能力.在描述逻辑语言家族中,相对简单的语言 $\mathcal{ALCQ}$对应有 4 条 Tableau 演算规则,而目前最复杂的 $\mathcal{SHOIQ}$(D)语言对应有 13 条规则.为了能够简洁阐述下面定理的证明,我们在此以 $\mathcal{ALCQ}$语言为例示意性给出对应的4条规则.其他语言框架下的规则相关的应用和证明大体类似,可进一步参考文献[19].$\mathcal{ALCQ}$中的 4 条规则分别如下:

(1) 如果$(C_1\sqcap C_2)(a)\in\mathcal{A}$,并且同时满足 $C_1(a)\notin\mathcal{A}$,$C_2(a)\notin\mathcal{A}$,则 $\mathcal{A}\to\sqcap\{C_1(a),C_2(a)\}\cup\mathcal{A}$;

(2) 如果$(C_1\sqcup C_2)(a)\in\mathcal{A}$,并且同时满足 $C_1(a)\notin\mathcal{A}$,$C_2(a)\notin\mathcal{A}$,则 $\mathcal{A}\to\sqcup\{C_1(a)\}\cup\mathcal{A}$ 或 $\mathcal{A}\to\sqcup\{C_2(a)\}\cup\mathcal{A}$;

(3) 如果$(\geqslant n\ R\cdot C)(a)\in\mathcal{A}$,$R(a,b)\in\mathcal{A}$,同时,$C(b)\notin\mathcal{A}$,$\neg C(b)\notin\mathcal{A}$,则 $\mathcal{A}\to\geqslant\{C(b)\}\cup\mathcal{A}$ 或 $\mathcal{A}\to\geqslant\{\neg C(b)\}\cup\mathcal{A}$;

(4) 如果$(\leqslant n\ R\cdot C)(a)\in\mathcal{A}$,$R(a,b)\in\mathcal{A}$,同时,$C(b)\notin\mathcal{A}$,$\neg C(b)\notin\mathcal{A}$,则 $\mathcal{A}\to\leqslant\{C(b)\}\cup\mathcal{A}$ 或 $\mathcal{A}\to\leqslant\{\neg C(b)\}\cup\mathcal{A}$.

定理 1(扩展一致性). 应用 Tableau 演算中衍生规则产生的新的断言集合 $\mathcal{A}'$是一致的,当且仅当断言集 $\mathcal{A}$ 是一致的.

证明:关于⊓衍生规则的证明,可以从对⊓构造算子与⊔构造算子的解释定义来看.对$\forall C_1,C_2\in\{C\}$,一个解释 $\boldsymbol{I}$ 满足$(C_1\sqcap C_2)(a)$,当且仅当 $C_1(a)\sqcap C_2(a)$.相应的,若存在一个解释使得包含$(C_1\sqcap C_2)(a)$的断言集是可满足的(一致的),则该解释也同样满足包含 $C_1(a)$和 $C_2(a)$的断言集合.反之亦然,即若 $\mathcal{A}'$是一致的,则 $\mathcal{A}$ 是一致的.应用解释同理可证⊔衍生规则、≥衍生规则和≤衍生规则. □

除以上给出的 4 条衍生规则的证明外,对于其他复杂语言下的衍生规则可以进行类似定理 1 的证明.根据定理 1,可以在初始断言集 $\mathcal{A}$ 一致的前提下反复运用衍生规则.扩展的终止条件是:反复扩展直至不再存在任何一条衍生规则使得扩展后的断言集 $\mathcal{A}'$是一致的.其中,断言集的一致性判定是本体推理的基本任务,通常采用

Tableau 演算方法计算.目前已有实现的一致性判定源代码,并提供 API 接口供开发者使用.因此,它不作为本文的重点进行细致的介绍和说明,可以参考文献[20,21].

3.3.2 域值关联推理

在概念语义计算模型中的属性语义叠加效应中,我们希望能够将关键词映射表中出现的属性对该网页的语义效应转换为概念对网页的语义效应.一般情况下,若一个属性的定义域和值域所对应的概念若同时出现在关键词映射词表中,我们将该属性对网页主题的语义效应用它的定义域和值域对应的概念来代替.为此,我们在本体下定义新的推理任务——域值关联推理,即计算词表中任意两个概念是否与指定属性存在关联.

定义 12(域值关联). 在属性语义叠加效应中,取关键词表内的任意两个概念 C_1 和 C_2,并将它们分别指定为待测属性 R 的定义域和值域,并构建公理 $C_1\sqsubseteq\exists R\cdot C_2$.若术语集 $\mathcal{T}\models(C_1\sqsubseteq\exists R\cdot C_2)$成立,则称概念 C_1,C_2 关于属性 R 关联,否则不关联.

定义关键词映射表中概念与属性关联关系的目的在于:若它们之间存在语义关联,则可将属性对整个网页的语义效应等效为与之关联的概念.在描述逻辑语言框架下,我们考虑绝大多数能够表达这种关联的公理表达式,如$\forall R\cdot C_2\sqsubseteq C_1,\geqslant n\ R\cdot C_2\sqsubseteq C_1,\leqslant n\ R\cdot C_2\sqsubseteq C_1,(\geqslant n\ R)\sqsubseteq C_1,(\leqslant n\ R)\sqsubseteq C_1$ 等等.随着描述逻辑语言的不断发展,总会有新的操作算子不断加入,以扩充其逻辑表达能力.这里,我们不保证已考虑到逻辑框架下域值相关的所有情况,只给出将上述几种常见逻辑形式转化为存在公理进行判定的非形式化证明.

证明:对于公理$(\geqslant n\ R)\sqsubseteq C_1$ 和$(\leqslant n\ R)\sqsubseteq C_1$,可以将 R 的值域看作顶层概念$\top$,又因为 $C_1\sqsubseteq\top$,所以有:若$\exists R\cdot C_2\sqsubseteq C_1$ 成立,则$(\geqslant n\ R)\sqsubseteq C_1$ 或$(\leqslant n\ R)\sqsubseteq C_1$ 成立.同理,若$\exists R\cdot C_2\sqsubseteq C_1$ 成立,从公理的语义解释显然知$\geqslant n\ R\cdot C_2\sqsubseteq C_1$ 或$\leqslant n\ R\cdot C_2\sqsubseteq C_1$ 成立.对于公理$\forall R\cdot C_2\sqsubseteq C_1$ 来说,若$\exists R\cdot C_2\sqsubseteq C_1$ 成立,则说明至少存在一个解释使得 C_1 和 C_2 关于属性 R 关联.若$\forall R\cdot C_2\sqsubseteq C_1$ 成立,则该解释亦为公理$\forall R\cdot C_2\sqsubseteq C_1$ 的解释.通过上述归纳证明可知:$\exists R\cdot C_2\sqsubseteq C_1$ 是较其他域值关联表达的更一般形式,我们可以通过判定公理$\exists R\cdot C_2\sqsubseteq C_1$ 的可满足性来反映域值关联的一般情况. □

定理 2(域值关联判定). 设 C_R^{domain} 是属性 R 的定义域,C_R^{range} 是属性 R 的值域.域值关联推理任务中,若概念$(C_R^{domain}\sqcap\neg C_1)$和$(C_2\sqcap\neg C_R^{range})$均是基于主题爬行本体中的术语集 $\mathcal{T}$ 不可满足的,则判定概念 C_1,C_2 关于属性 R 关联.

证明:若概念 $\neg(C_R^{domain}\sqcap\neg C_1)$是不可满足的,则有 $C_R^{domain}\sqcap\neg C_1\sqsubseteq\bot$;进一步有 $C_R^{domain}\sqsubseteq C_1$.

同理,概念 $\neg(C_2\sqcap\neg C_R^{range})$是不可满足的,则有 $C_2\sqcap\neg C_R^{range}\sqsubseteq\bot$;进一步有 $C_2\sqsubseteq C_R^{range}$.

因为 C_R^{domain} 和 C_R^{range} 是属性 R 的定义域和值域,所以$\exists R\cdot C_R^{range}\sqsubseteq C_R^{domain}$ 成立.

又因为 $C_R^{domain}\sqsubseteq C_1$,所以有$\exists R\cdot C_R^{range}\sqsubseteq C_1$.

因为$(\exists R\cdot C_R^{range})^I=\{a\in\Delta^I|\exists b\cdot(a,b)\in R^I\wedge b\in(C_R^{range})^I\}$,$(\exists R\cdot C_2)^I=\{a\in\Delta^I|\exists b\cdot(a,b)\in R^I\wedge b\in(C_2)^I\}$;

又因为 $C_2\sqsubseteq C_R^{range}$,即 $(C_2)^I\subseteq(C_R^{range})^I$,所以 $(\exists R\cdot C_2)^I\subseteq(\exists R\cdot C_R^{range})^I$,则 $\exists R\cdot C_2\sqsubseteq\exists R\cdot C_R^{range}$.根据已得的结论,$\exists R\cdot C_R^{range}\sqsubseteq C_1$,进一步有$\exists R.C_2\sqsubseteq C_1$,即在 $\mathcal{T}$ 下有 $\mathcal{T}\models(\exists R\cdot C_2\sqsubseteq C_1)$.根据定义 12 有,概念和是域值关联的. □

通过上述定理,我们将域值关联问题转化为判定概念的可满足问题.定理中的符号 R 对应关键词表中的属性高频词,概念 C_1 和 C_2 对应关键词表中的任意两个不同概念.C_R^{domain} 为该属性在主题爬行本体中定义的定义域概念,C_R^{range} 为该属性在主题爬行本体中定义的值域概念.在程序实现中,通常将其进一步转化为断言集的一致性检测问题.与上节关于定理 1 的说明一样,计算一致性的问题请参见文献[20,21].

3.4 语义相关度排序

在应用语义叠加效应模型获得每个网页的主题概念的基础上,我们通过推演主题概念间的包含关系对网页的出链接排序,进而按出链接优先次序顺序抓取网页,实现主题爬行.排序的原则是被包含概念所对应出链接置后,包含概念的出链接优先.

定理 3(包含关系判定). 若存在概念 $C,D\in\{C\}\in O_{topic}$,称 $C\sqsubseteq D$ 在当前术语集 $\mathcal{T}$ 下成立,当且仅当存在一个实例 a,使得概念$(C\sqcap\neg D)$在利用 $\mathcal{T}$ 扩展后得到的断言集是不一致的.

证明:若 $C\sqsubseteq D$ 成立,则有 $C\sqcup(\neg D)\sqsubseteq D\sqcup(\neg D)$.因为 $D\sqcup(\neg D)\equiv\top$,所以有 $C\sqcup(\neg D)\sqsubseteq\top$ 成立.进一步地,公理的两侧同时取反,有$(\neg C)\sqcap D\sqsubseteq\bot$成立.从语义定义来讲,亦即概念$(\neg C)\sqcap D$ 是不可满足的.又因为由引理 1 可知:如果任意一个概念 S 在利用 $\mathcal{T}$ 扩展后的概念 $_{\mathcal{T}}S'$是不可满足的,那么概念 S 是不可满足的.所以对于概念$(\neg C)\sqcap D$,若利用 $\mathcal{T}$ 扩展后得到的概念$((\neg C)\sqcap D)'$是不可满足的,则概念$(\neg C)\sqcap D$ 是不可满足的.因为由引理 2 得可知,若概念$((\neg C)\sqcap D)'$是(不)可满足的,相应的断言集$\{((\neg C)\sqcap D)'(a)\}$是(不)一致的;反之亦然. □

由定理 3,保证了将概念包含关系问题转化为断言集的一致性检测推理问题.一般来讲,网络资源在内容语义上都是按照由抽象到具体的层次结构组织链接的.例如,一个学校网站的链接结构是由学校、院系、科室等等逐层嵌套的.我们通过一系列定义和命题得出结论 1,利用网页主题概念包含关系判定对应 URLs 抓取序列的策略,从语义层面上最大限度的实现了主题资源的宽度遍历,能够有效获取更多主题相关资源.

结论 1. 在领域爬行本体 O_{topic} 下,存在两个网页的主题概念 C^1_{rel} 和 C^2_{rel} .若判定有 $C^1_{rel}\sqsubseteq C^2_{rel}$,则认为 C^2_{rel} 所对应网页的主题相关度高于 C^1_{rel} ;反之亦然.

我们建立网页主题概念到网页出链接的映射关系,通过结论 1 判定网页主题概念间的包含关系.进而根据包含次序确定对应网页出链接的主题相关度并排序,对主题相关度高的网页出链接优先执行网络爬行,从互联网上抓回相关网页并进一步提取网页出链接.

4 相关评测

为验证语义主题爬行策略的有效性,我们选取 4 种相关的主题爬行方法进行了测试比较,分别是 Breadth-First 算法、Best-First 算法、Ontology-Focused 方法以及我们的语义主题爬行策略.

其中,Breadth-First 算法是基本的 Web 爬行算法,它采用宽度优先搜索策略进行网络爬行;Best-First 算法采用关键词集合描述主题,为使测评具有可比性,我们用本文提出算法中的本体和扩展词表作为它的关键词集合进行测试.Best-First 算法在主题爬行领域颇具代表性,很多研究者将其作为算法性能比较的基准[22];Ontology-Focused 方法是首个利用本体描述爬行主题的算法,为与我们的方法比较,我们在评测工作中实现了该方法,并利用该方法对我们使用的本体及其扩展词表测试效果;最后是本文提出的语义主题爬行策略,我们利用 protégé[23]工具创建了 OWL DL 语言描述的金融本体.参考金融领域专业书籍[24]定义诸如有价证券、股票、股票经纪人、货币、股票发行单位等概念、属性及其关联.所构建本体中的实例则参考了金泉网下的金融财经类(http://www.dir.jqw.com/dir/575/)、DMOZ 网下的中文站点(http://www.dmoz.org/World/Chinese_Simplified/)等主要人工分类网站资源,将诸如本体中上市的股票实例、股票公司实例等相关实例信息收集整理.最后,得到一个包含 403 个概念、228 个属性、3062 个实例的金融领域本体,作为本文实验用的主题爬行本体.其中,图 2 就是本体的一个局部片段描述.算法中的扩展词表参考具有丰富完整语义解释的 wordnet 词典[25],获得其中与本体中概念、属性和实例对应词条的同义词词条.除此之外,还扩充了诸如股票名的简称与股票名之间的同义词关系等其他同义词条,最终构建出我们需要的扩展词表.

我们以金融领域为主题进行主题爬行,计算 $pages_{relation}$(即相关网页)数目,见表 4:第 1 行代表爬虫抓取的网页数目;第 1 列表示不同的爬行算法:Bread 代表 Bread-First 算法;Best 代表 Best-First 算法;Onto 代表 Ontology-Focused 方法;Sema*和 Sema**分别是本文中语义爬行策略的部分实现,其中,Sema*未作 *ABox* 的一致性扩展,仅简单实现实例到概念的直接映射;Sema**较前者虽然完整实现了实例到概念的语义叠加部分,但仍未添加属性到概念的语义叠加;最后的 Sema 是本文中提出的完整的基于语义的主题爬行策略.

Table 4 Number of relational pages about economy topic

表 4 金融领域相关网页数目

Download	250	500	750	1 000	1 250	1 500	1 750	2 000	2 250	2 500	2 750	3 000
Bread	18	21	25	27	29	33	36	42	46	49	53	54
Best	66	107	153	342	350	360	367	380	382	400	412	420
Onto	63	215	345	520	587	645	700	780	900	900	907	930
Sema*	67	157	283	369	441	495	503	560	713	719	737	756
Sema**	73	242	324	558	621	645	664	682	785	842	979	1 080
Sema	85	275	360	590	687	705	717	720	832	900	1 072	1 203

关于表 4 中网页主题相关性的判定,如果能够采用人工界定当然是最准确的.但是显然,主题爬行资源的海量性使得人工判定不切实际.成熟的文本分类技术[26]提供了高效低成本的判定途径.我们首先获取分类器训练需要的两类数据源:一个是关于财经类的网页数据集,从金泉网这个国内大规模的商业分类网站的“财经证券”子目录抓取 10 000 个领域相关网页;另一个是领域不相关的网页数据集,从 DMOZ 网页分类网站的 Top/World/Chinese_Simplified 目录下获取 10 000 个其他领域的网页.利用爬虫将数据源下载之后,将其转换为对应的 TF/IDF 格式数据,生成 svm_model 文件.进一步地,在 LPU(learning from positive and unlabeled data)算法[27]下利用样本集训练财经类网页分类器,通过得到的分类器可以批量判定抓取网页是否与主题相关.接着计算收获率(harvest-rate),评价本文中的爬行策略.收获率是估算主题爬行器获取的一组网页与给定主题相关程度的性能指标,目前广泛应用于主题爬行研究的评测[22].应用如下的收获率公式计算收获率:

$$hr=\frac{|pages_{relation}|}{|pages_{download}|},\ hr\in[0,1],$$

其中,hr 代表收获率.$pages_{download}$ 代表下载的网页,$pages_{relation}$ 代表与主题相关网页,它们的数目比值即为收获率,收获率的取值范围在 0,1 之间.根据表 4 得到各种爬行策略的爬行收获率如图 3 所示.

首先对算法 Sema*,Sema**和 Sema 进行比较分析.Sema*只简单实现了实例到概念的直接映射,不能全面反映网页的主题;Sema**通过对 $ABox$ 的一致性扩展,能够发现隐含的语义关联,较好地实现了实例到概念语义效应叠加.实验结果表明,Sema**的主题爬行效率优于 Sema*.相对于 Sema**,本文中的基于语义的主题爬行策略(即 Sema)进一步考虑了网页高频词中属性对主题概念的叠加效应.实验结果表明,它能够有效地提高主题爬行的收获率.

继续比较基于语义的主题爬行策略与其他爬行策略(即 Breadth-First,Best-First,Ontology-Focused).由图 3 的实验结果可以看到,基于语义爬行策略(图中 Sema 线)普遍优于其他算法,该方法最好时候能够达到 0.6 的收获率.其中,Breadth-First 算法因其未采用任何主题爬行策略,抓取的收获率很低,几乎与 x 轴平行.可见,采用主题爬行策略对于专业领域的垂直搜索十分必要.图中的 Best-First 算法因其采用了基于关键字的主题爬行策略,相对 Breadth-First 算法主题网页的抓取效果明显提升.Ontology-Focused 方法的收获率相对前两种方法有了更明显的提高,甚至在个别点优于本算法.主要原因是它利用了本体结构的层次关系.我们给出的语义主题爬行策略是相比较算法中效果最好的,其利用语义知识进行主题爬行优势非常明显.除个别点低于 Ontology-Focused 方法(初步分析和相关网络资源的实际分布情况有关),在绝大多数情况下都有不错的收获率.而且随着抓取数目的增加,我们的语义主题爬行策略在收获率上并未见明显衰减迹象.

我们进一步比较了各种方法的爬行效率.以主题爬虫运行 20m 为限,统计相同时间内各种主题爬行方法抓取到的主题相关网页数量.其中,Breadth-First 方法在 20m 内共获取 69 个主题相关网页,Best-First 方法共获取 485 个,Ontology-Focused 方法获取了 766 个,Semantic-Based 方法获取了 980 个.各阶段获取比较如图 4 所示,其中:x 轴代表爬行耗时,单位为分钟;y 轴代表爬行过程中获取的相关网页;图中的点代表各种方法在对应时间点获得的主题相关网页数目.

由图 4 可以看出:没有应用任何主题爬行策略的 Breadth-First 方法效率极低,与其他方法没有可比性;Best-First 方法由于仅应用了词表做关键词,而没有考虑词间的语义关联,相对后两种方法而言,爬行效率仍然有较大差距;Ontology-Focused 方法在初期抓取效率略高于 Semantic-Based 方法,主要是因为 Ontology-

Focused 方法简单的语义距离计算效率高于 Semantic-Based 方法的推理策略.但随着 Ontology-Focused 方法中本体实例不断地补充,后期效率有所降低;而 Semantic-Based 方法则保持了较高的爬行效率.

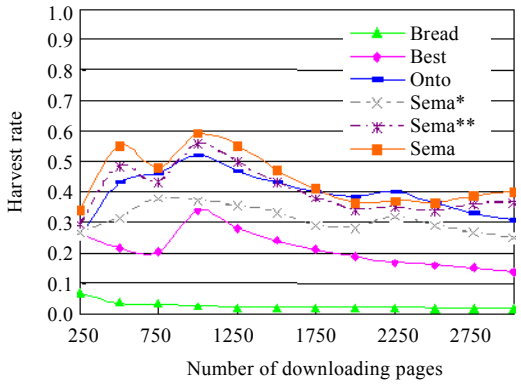


Fig.3 Result of algorithms testing

图 3 算法对比测试结果

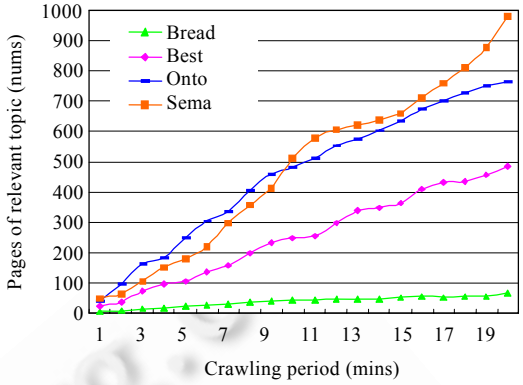


Fig.4 Comparing of crawling efficiency

图 4 爬行效率比较

5 结 论

本文在分析、比较当前主题爬行研究现状的基础上深入研究了基于本体刻画领域主题的 Web 爬行,在本体及本体推理框架下提出语义主题爬行策略.该策略通过建立映射机制赋予关键字本体语义,并利用自定义的主题爬行下推理任务(一致性扩展和域值关联推理)来架构语义叠加效应模型.对由效应模型得到的一组网页主题概念,应用概念间包含关系判定主题相关度,进而确定网页出链接排序并实现网页抓取.其中,语义叠加效应模型由实例语义效应叠加、属性语义效应叠加和峰值效应 3 部分组成.在实验分析中,我们对本文中的语义爬行策略做了进一步验证,结果表明:由于充分利用了本体的语义信息进行推演和计算,我们的方法明显优于其他相关算法,该方案有效可行.收获语义相关领域资源的目的在于为用户提供语义检索,目前主要利用本体的语义映射和匹配技术平滑用户查询和领域本体的语义差异,我们将在未来的工作中对此做进一步探讨和研究.

References:

[1] Fallows D. Search engine users: Internet searchers are confident, satisfied and trusting-but they are also unaware and naïve. Pew Internet & American Life Project. 2005. http://www.pewinternet.org/~/media/Files/Reports/2005/PIP_Searchengine_users.pdf.pdf

[2] Rainie L. Big jump in search engine use: Nearly 60 million online Americans use search engines on an average day. Pew Internet and American Life Project. 2005. http://www.pewinternet.org/Reports/2005/Big-jump-in-search-engine-use.aspx

[3] Spink A, Zimmer M. Web Search: Public Search of the Web. Berlin, Heidelberg: Springer-Verlag, 2005. 3−8. [doi: 10.1007/1-4020-2269-7]

[4] De Bra PME, Post RDJ. Searching for arbitrary information in the WWW: The fish-search for Mosac. In: Cailiau R, ed. Proc. of the Selected Papers of the 1st World Wide Web Conf. Amsterdam: Elsevier, 1994. http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/Searching/debra/article.html

[5] Chakrabarti S, van den Berg M, Dom B. Focused crawling: A new approach to topic-specific Web resource discovery. Computer Networks, 1999,31(11-16):1623−1640. [doi: 10.1016/S1389-1286(99)00052-3]

[6] Srinivasan P, Pant G, Fenczer F. Target seeking crawling and their topical performance. In: Jarvelin K, ed. Proc. of the 25th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2002.

[7] Rennie J, McCallum AK. Using reinforcement learning to spider the Web efficiently. In: Bratko I, Dzeroski S, eds. Proc. of the 16th Int'l Conf. on Machine Learning (ICML'99). San Francisco: Morgan Kaufmann Publishers, 1999. 335−343.

[8] Hsu CC, Wu F. Topic-Specific crawling on the Web with the measurements of the relevancy context graph. Information Systems, 2006,31(4):232−246. [doi: 10.1016/j.is.2005.02.007]

[9] Peng T. Research on topical Web crawling technique for topic-specific search engine [Ph.D. Thesis]. Changchun: Jilin University, 2007 (in Chinese with English abstract).

[10] Ehrig M, Maedche A. Ontology-Focused crawling of Web documents. In: Lamont BG, ed. Proc. of the 2003 ACM Symp. on Applied Computing. New York: ACM Press, 2003. 1174−1178. [doi: 10.1145/952532.952761]

[11] Ganesh S. Ontology based Web crawling—A novel approach. In: Szczepaniak PS, Kacprzyk J, Niewiadomski A, eds. Proc. of the Advances in Web Intelligence 3rd Int'l Atlantic Web Intelligence Conf. Berlin, Heidelberg: Springer-Verlag, 2005. 140−149. [doi: 10.1007/11495772_2310.1007/b137066]

[12] Wang H, Zuo WL, Yuan H. A classification method based on centroid and ontology. Journal of Computer Research and Development, 2007,44(z2):6−11 (in Chinese with English abstract).

[13] Guarino N. Formal ontology, conceptual analysis and knowledge representation. Int'l Journal of Human-Computer Studies, 1995, 43(5-6):625−640. [doi: 10.1006/ijhc.1995.1066]

[14] Bechhofer S, Harmelen FV, Hendler J, Horrocks I, McGuiness DL, Patel-Schneider PF, Stein LA. OWL Web ontology language reference. 2004. http://www.w3.org/ TR/owl-ref//

[15] Ye YX, Ouyang DT, Ling J, Zhang YG. Research and design on reasoning algorithm with ontologies and rules. Journal of Jilin University, 2009,39(5):1297−1302 (in Chinese with English abstract).

[16] Porter MF. An algorithm for suffix stripping. Program: Electronic Library and Information Systems, 2006,40(3):211−218. [doi: 10.1108/00330330610681286]

[17] Buchheit M, Donini FM, Schaerf A. Decidable reasoning in terminological knowledge representation systems. Journal of Artificial Intelligence Research, 1993,1(1):109−138. [doi: 10.1613/jair.21]

[18] Levy AY, Rousset MC. Combining horn rules and description logics in CARIN. Artificial Intelligence, 1998,104(1-2):165−209. [doi: 10.1016/S0004-3702(98)00048-4]

[19] Horrocks I, Sattler U. A tableau decision procedure for SHOIQ. Journal of Automated Reasoning, 2007,39(3):249−276. [doi: 10.1007/s10817-007-9079-9]

[20] Baclawski K, Kokar MM, Waldinger RJ, Kogut PA. Consistency checking of semantic Web ontologies. In: Horrocks I, Hendler JA, eds. Proc. of the 1st Int'l Semantic Web Conf. Hamburger: Springer-Verlag, 2002. 454−459. [doi: 10.1007/3-540-48005-6]

[21] Ye YX, Ouyang DT, Liu Y, Sun JG. Consistency checking of the *SHOIQ*(*D*)-based ontology. Computer Engineering and Science, 2009, 31(8):7−9 (in Chinese with English abstract).

[22] Menczer F, Pant G, Srinivan P. Topical Web crawlers: Evaluating adaptive algorithms. ACM Trans. on Internet Technology, 2004, 4(4):378−419. [doi: 10.1145/1031114.1031117]

[23] Knublauch H. Ontology-Driven software development in the context of the semantic Web: An example scenario with protégé/OWL. In: Frankel D, ed. Proc. of the 1st Int'l Workshop on the Model-Driven Semantic Web (MDSW 2004). Enabling Knowledge Representation and MDA® Technologies to Work Together, 2004.

[24] Huo WW. Tutorial of Financial Market. Shanghai: Publication of Fudan University, 2005. (in Chinese).

[25] Miller GA. WordNet: A lexical database for English. Communications of the ACM, 1995,38(11):39−41. [doi: 10.1145/219717.219748]

[26] Sebastinami F. Machine learning in automated text categorization. ACM Computing Surveys, 2002,34(1):1−47. [doi: 10.1145/505282.505283]

[27] Li XL, Liu B. Learning to classify texts using positive and unlabeled data. In: Gottlob G, Walsh T, eds. Proc. of the 18th Int'l Joint Conf. on Artificial Intelligence (IJCAI 2003). San Fransisco: Morgan Kaufmann Publishers, 2003. 587−594.

附中文参考文献:

[9] 彭涛.面向专业搜索引擎的主题爬行技术研究[博士学位论文].长春:吉林大学,2007.

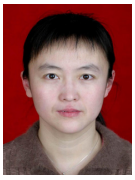
[12] 王辉,左万利,袁华.一种基于质心与本体的文本分类方法.计算机研究与发展,2007,44(z2):6−11.

[15] 叶育鑫,欧阳丹彤,领吉,张永刚.本体与规则整合的研究与设计.吉林大学学报(工学版),2009,39(5):1297−1302.

[21] 叶育鑫,欧阳丹彤,刘瑶,孙吉贵.基于 *SHOIQ*(D)的本体一致性检测.计算机工程与科学,2009,31(8):7−9.

[24] 霍文文.金融市场学教程.上海:复旦大学出版社,2005.

叶育鑫(1981−),男,吉林长春人,博士,讲师,主要研究领域为语义 Web 及其相关技术.



欧阳丹彤(1968−),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为知识工程,自动推理.