

文章编号:1005-0523(2014)02-0112-07

一种关联多本体的科技奖励检索方法

曾辉,江健健,熊李艳,黄卫春

(华东交通大学信息工程学院,江西 南昌 330013)

摘要:现有的科技奖励检索都是基于关键词的匹配,忽略了对用户查询的语义理解。针对传统信息检索存在的问题以及结合当前面向实际应用的基于本体的语义检索的特点,提出一种关联多本体的科技奖励检索方法。通过对科技奖励项目的领域本体构建,对概念相似度计算方法的改进及关联多本体概念,使扩展词语更能表达用户检索意图。实验结果表明,该方法对比传统查询检索和单独本体扩展检索获得了更好的召回率和准确率。

关键词:科技奖励;关联多本体;本体构建;概念相似性

中图分类号:TP391

文献标志码:A

科技奖励检索作为整个科技奖励平台的一部分,为用户及时、方便、准确地获取相关领域的科技信息提供了保证。现有基于关键词的科技奖励检索没有考虑用户检索条件的语义理解,造成返回的查询结果与用户所需要的信息内容不匹配。本体作为特定领域内概念以及概念之间关系的集合^[1],具有良好的语义表达能力,非常适合用于概念语义的表达。

在面向实际应用的基于本体的信息检索中,大多是通过单一本体概念间相似度的计算对检索词进行扩展再送入搜索引擎中进行检索,这样,在召回率上必然大大提高,然而同时也会产生许多无关用户检索意图的结果,准确率提高不明显。针对大量的扩展概念,如何选择更能表达用户意图的概念组合成为引人关注的研究问题。基于此,提出了一种关联多本体的概念相似度计算方法来筛选扩展概念以改进科技奖励的检索。

1 相关研究

针对本体及信息检索,业界已有不少相关研究:文献[2]定义了本体的通用描述,即本体是共享概念模型的明确形式化规范说明,从内涵上来看,本体是某个领域的知识抽象表示,能够在知识层次上描述信息,为不同的主体(用户、机器等)之间的交流提供了一种语义基础。基于本体的内涵,文献[3]通过本体技术的引入,提出了一种智能检索模型的建立,从而解决网络教育资源杂乱无章的问题。文献[4]通过对交通信息进行抽象和分析,提出了一种基于Jena的城市交通领域本体推理和查询方法。文献[5]将本体应用在对政务信息进行标引、对检索条件的扩展和结构化检索条件的自动生成。文献[6]提出了一种基于领域本体的混合信息检索模型,通过建立关键词基础矩阵和语义扩展矩阵两层索引矩阵,保证一定的检索性能。文献[7]提出了一种基于语义相似度的个性化信息检索方法,有效地对用户的查询请求进行概念扩充,提高了搜索的查全率与查准率。文献[8]通过分析本体上下文结构,引入结构和实例相似度传播和快速匹配算法,设计一种复合匹配策略,用于本体映射。文献[9]通过本体模型和概念相似度的计算对检索信息进行检索意图树的构建并扩展,得到更好的检索结果。

收稿日期:2013-10-20

基金项目:国家自然科学基金项目(61363072)

作者简介:曾辉(1973—),男,副教授,研究方向为数据库技术,计算机决策支持系统。

纵观上述研究,目前信息检索中,都有考虑有效利用本体和语义信息来对文本进行语义检索。但都只是对单一本体概念进行扩展,没有将多个本体概念进行关联,需进一步完善。本文将本体引入科技奖励检索系统的主要目的在于对用户的检索输入进行语义分析,针对扩展检索词所在本体的不同情况,通过对扩展概念的相似性计算来对扩展概念进行组合、筛选,使扩展后的概念更能体现用户的查询表达,提高检索准确率。

2 领域本体构建及关联多本体检索方法

2.1 领域本体的构建

领域本体的构建是实现语义检索系统的重要基础和关键环节。同时,领域本体对检索文档的语义信息标注以及对用户检索意图的解析也都起着至关重要的作用。本文根据文献[10]提出的方法来构建领域本体。分为4个步骤:

1) 确定本体的领域和范围。模型希望构建有关科技奖励申报项目所涉及的13个领域,根据每个不同领域所涉及的知识,考虑领域知识的深度和广度以及关系的复杂程度。

2) 获取领域知识。文本构建的本体知识来源主要是科技奖励申报领域的文档信息,专家知识及可复用存在的本体。

3) 定义类及其关系,建立本体模型。本阶段对收集的知识进行分析抽象,进而建立本体模型。本体类定义要明确,不应包括全部信息,应表示类的最突出属性。一个新类通常会增加其父类不具备的新的属性,或覆盖父类属性的约束。

4) 使用 OWL (web ontology Language) 网络本体语言表示本体。为实现本体的形式化表示,我们可以使用 OWL 表示本体模型。OWL 本体包括了类、属性、个体的描述^[11]。OWL 提供了丰富的公理,不仅准确描述了知识中的类、属性、个体,还对它们之间的复杂的逻辑关系进行精确描述,为知识的推理做了良好的准备。

如图1为本文构建的农业项目领域本体的部分关系图,采用的是protégé_4.1 本体构建工具。

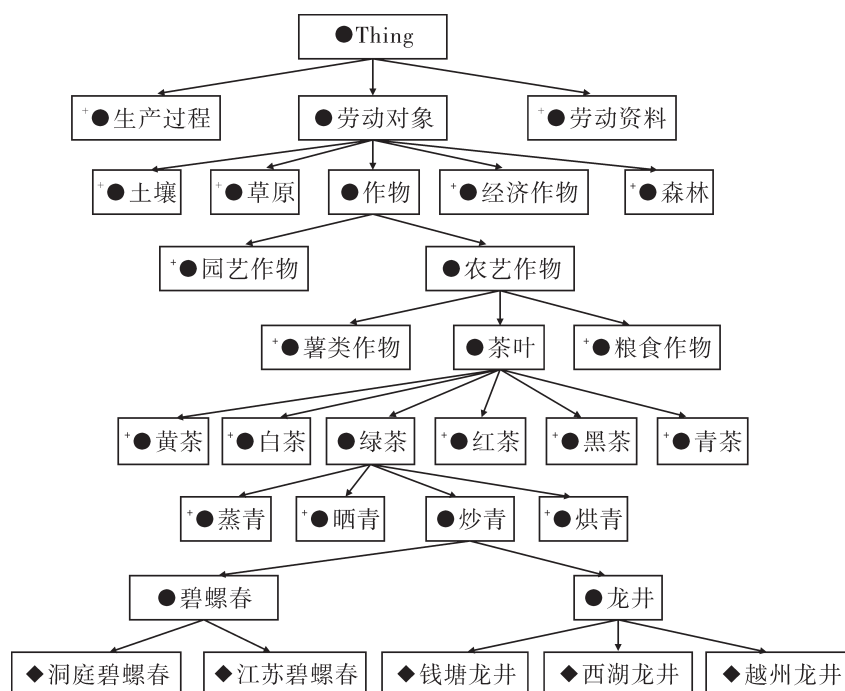


图1 农业项目领域本体部分关系图

Fig.1 Partial diagram for domain ontology of agricultural projects

2.2 用户查询处理

通常用户查询输入的方式有3种:单一关键词,多个关键词及自然语言查询。对于单一关键词,直接用本体概念相似性对检索词进行扩展后进行检索;对于自然语言查询,通过结合领域概念进行分词最终转化为多个关键词的情形。本文主要叙述多个关键词的查询,其算法流程如下:

输入:用户查询词语,领域本体

输出:查询检索结果

Begin:

- 1) 对用户输入的查询词语,判断它是否是本体中的概念,如果不是本体中的概念,不进行扩展,反之,转向步骤2。
- 2) 若其中的词语是本体中的概念,基于距离的语义相似度计算对概念词进行语义扩展,选取语义相似度排在前 M 个的作为该检索词的扩展词。
- 3) 比较不同词语的扩展词间是否有相同概念:如果有,根据文本提出的关联多本体方法来对分别扩展的概念进行相似度计算,选择相关性大的概念组合作为最终的检索扩展词;如果没有,则通过寻找术语上相等的概念,将概念进行联系,得到最终扩展词。
- 4) 基于最终扩展词与标注好的数据源进行匹配。
- 5) 从数据库中将匹配的检索结果按相似度高低排序展现给用户。

End.

2.3 关联多本体方法

概念相似度计算是进行语义扩展的重要步骤,其精度也是提高检索质量的关键。目前常用的计算本体的概念相似度方法有:基于距离的、基于内容的和基于属性的语义相似度计算。这些方法都是在单一本体中对检索词进行概念相似度计算、扩展,没有考虑两组扩展词间的相似度,对其进行的是独立扩展。如何在给定检索扩展阈值 T 的情况下,选择语义相关性更大的检索词对进行组合,本文在文献[12]提出的相似度计算方法上,提出了一种新的方法来解决概念存在多个本体中的相似性计算方法。

文献[12]提出的方法通过计算被比较的两个概念所具有的共同和非共同分类包含的数量来评估概念相似性。 C_1 和 C_2 两概念节点,定义它们之间的相似度 sim 计算公式如下:

$$\text{sim}(c_1, c_2) = -\log_2 \frac{|T(c_1) \cup (c_2)| - |T(c_1) \cap (c_2)|}{|T(c_1) \cup (c_2)|} \quad (1)$$

式中: $T(c_i)$ 为节点 c_i 的父节点及节点 c_i 本身的集合,即 $T(c_i) = \{c_j \in C | c_j > c_i\} \cup c_i$,“ $>$ ”表示节点间的父子关系, $c_j > c_i$ 表示 c_j 是 c_i 的一个父节点。

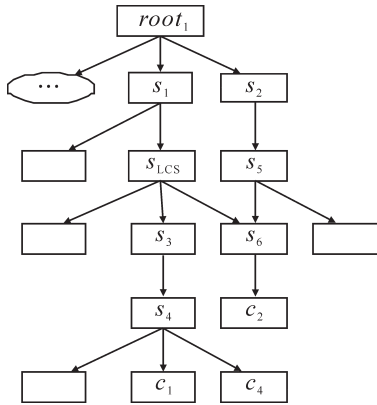
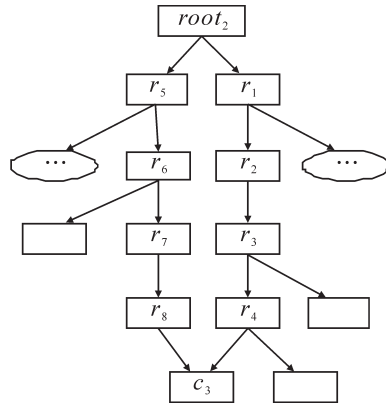
从公式(1)可以看出,该计算公式得出了一个评估分类特性的一个比值,可以用来比较独立的本体尺寸和粒度的相似性值。这个公式在多本体环境中是有相关性的,因为它能提供来自不同的本体获得结果的比较。

通常对查询概念的扩展分别是对单一本体概念的扩展,用扩展后的词语对数据源进行独立检索,没有考虑扩展词语间的关系。文本通过关联多本体概念,将处于不同本体的两个概念进行关联,挖掘该概念的深层次的关联性,使查询扩展词更能表达用户的检索意图。

目前,评估多本体间的概念相似性已有了一些解决办法,文献[13]提出的方法是通过合并不同本体成一个唯一的本体。然而这种方法有其必然的缺陷,在处理模糊重叠的概念和避免概念的不一致性时,会导致很高的计算量和人力成本。文献[14]在区别主要本体和次要本体的差异基础上,通过连接所有相等的节点把次要本体与主要本体联系在一起。

然而,由于语言歧义(同义和一词多义)和不同的知识表示过程中,术语的匹配提供了一个有限的召回率。为使这一问题减到最小,除了考虑常见的共同父节点作为那些术语的匹配,我们也考虑了他们所有的

包含,无论有或没有一个完全相同的标签。事实上,每个继承术语等效上层概念的被评估概念递归地继承了所有上层概念包含。图2、图3分别为本体 O_1 和本体 O_2 的部分概念关系图。在评估本体概念相似度时,需要明确以下术语:

图2 本体 O_1 Fig. 2 Ontology O_1 图3 本体 O_2 Fig. 3 Ontology O_2

路径长度 $Path$:指本体图中两概念节点间路径的长度。 $|Path(c_1, c_2)|$ 为路径长度。从图2可知, $Path(c_1, c_2) = \{c_1 - s_4, s_4 - s_3, s_3 - s_{LCS}, s_{LCS} - s_6, s_6 - c_2\}$ 。

得到 $|Path(c_1, c_2)| = 5$ 。路径长度与语义相似度成反比。

节点深度 $Depth$:指概念 c 到根节点的路径长度,即 $Depth(c) = |Path(c, root)|$ 。由图2可知, $Depth(c_1) = 5$ 。概念节点深度越大,表明此概念表示的意义越具体。

最近共同父节点 LCS (least common subsumer):指在词汇分类学上被比较的两概念具有最短距离的概念。其计算公式如下

$$LCS(c_1, c_2) = \arg \max_{s \in \left\{ \arg \min_{c \in Sup(c_1, c_2)} \left(|Path(c_1, c)| + |Path(c, c_2)| \right) \right\}} Depth(s) \quad (2)$$

从图2可知, $LCS(c_1, c_2) = s_{LCS}$,表明节点 s_{LCS} 是节点 c_1, c_2 的最近公共父节点。

等效上层概念 ES (equivalent superconcepts):指所属不同本体的两概念节点的父节点中存在等效的概念集合。通过计算两概念的上层节点,判断是否存在术语上等效的上层概念,其定义如下

$$ES = \{c_i \in T_{O_1}(c_1) | \exists c_j \in T_{O_2}(c_2) \wedge c_i \equiv c_j\} \quad (3)$$

式中:“ $\equiv$ ”表示两概念等效。

不同本体间等值概念之间的检测已在本体联合领域中有所研究。基于不同的原理,已经提出了几种方法来评估不同本体的概念实际上是等效的可能性。许多方法依赖于语义相似度函数来进行评估,需要由专家或来自其他知识源的计算。然而一种无监督的方法是必要的。术语的匹配方法是可行的,他们发现等效概念完全依赖于概念标签匹配^[15]。应用到我们的系统,当提到具有相同文本标签时,我们可以认为他们是等效上层概念。

概念 c_1 包含在本体 O_1 查询扩展集合中,概念 c_2 包含在本体 O_2 的查询扩展集合中。假设概念 s_3 和 r_3 是术语匹配概念,即 s_3 和 r_3 被认为是相同的概念。在这种情况下,易知, $ES = \{s_3\}$ 。通过得到术语等效概念,我们联系两本体,计算他们的共同特征节点集合。

共同特征节点集合 CS (common specificity):指由 ES 元素及其父节点所组成的节点集合。计算公式为

$$CS(c_1, c_2) = \bigcup_{c_i \in ES} (T_{O_1}(c_i) \cup T_{O_2}(c_i)) \quad (4)$$

易知, $CS(c_1, c_3) = \{s_3, root_1, s_1, s_{LCS}, root_3, r_1, r_2\}$ 。

通过找寻两本体概念的共同特征节点, 可以将看似不相关的两个概念进行关联起来。最后, 通过改进的本体概念相似度计算公式来计算所属两个不同本体的概念相似度, 公式如下

$$\text{sim}(c_1, c_2) = -\log_2 \frac{|T(c_1) \cup (c_2)| - CS(c_1, c_2)}{|T(c_1) \cup (c_2)|} \quad (5)$$

根据改进公式, 可以计算所属不同本体的两概念间的相似度, 如计算本体 O_1 中的概念 c_1 和本体 O_3 中的概念 c_3 间的相似度, $\text{sim}(c_1, c_3) = -\log_2(\frac{15-7}{15}) = 0.9$ 。很明显, 概念 c_1 和 c_3 具有很大的相关性, 可以作为检索组合进行检索, 避免语义的丢失。

3 实验结果及分析

本实验以“科技奖励推荐评审系统”中收集到的各学科研究领域的奖励申报书作为检索源, 这些文档涉及了材料、农业、药物与医疗器械、计算机技术等 13 个学科领域。由于本体构建工作的复杂性, 仅构建了农业领域的相关本体, 因此选用的是农业领域 200 篇申报书作为实验数据。模型的评价使用的是在信息检索中常用的准确率 (precision) 和召回率 (recall) 作为评价标准。

$$\text{准确率 } P = \frac{\text{检索到的相关资源数}}{\text{检索到的所有资源数}} \times 100\%$$

$$\text{召回率 } R = \frac{\text{检索到的相关资源数}}{\text{所有的相关资源数}} \times 100\%$$

实验分别采用关键词匹配、单独扩展本体和本文提出的关联多本体方法对农业项目相关领域进行检索, 图 4, 图 5, 图 6 分别为输入“杂交水稻”的 3 种不同方法的检索界面图:

通过对后台数据库的检索, 发现采用关联多本体的方法, 在检索效率方面明显要低于单独扩展本体和关键词匹配。原因是采用关联多本体方法, 需要进行多个本体间相似度计算和筛选, 需要耗费一定的时间, 但为了准确率和召回率的提高, 以满意度换效率还是可行的。相比目前最强大的搜索系统 GOOGLE, 两系统最后都是通过关键词的匹配返回结果, 在对关键词处理和后续结果显示方面存在如下区别: GOOGLE 系统将对检索词提供关键字建议及同义词扩展, 并对结果列表进行优化, 召回率差, 而且对领域相关知识的查询不令人满意; 而本系统针对的有限领域语料库, 在检索效率方面不是很好, 但在查准率和查全率方面都有很好的性能。通过人工判断与“杂交水稻”相关的资源数, 与检索结果进行对比, 得到查询性能对比表, 如表 1 所示。



图4 关键词匹配检索界面

Fig. 4 Matching keywords retrieval interface



图5 单独扩展本体检索界面

Fig.5 Separate extension ontology retrieval interface

表1 3种方法的查询性能
Tab.1 Query performance of three methods

	检索到的所有资源数/个	检索到的相关资源数/个	所有的相关资源数/个	P	R
关键词	14	9	19	0.642	0.473
单独扩展本体	23	14	19	0.610	0.737
关联多本体	20	17	19	0.850	0.895

由于检索资源库数量不大,相关资源数值小,少一两个相关资源数对计算结果数值会有很大波动。通过多次检索结果比较发现,在本次试验条件下,关联多本体检索方法可以提供更高的准确率和较好的召回率,可以给用户更直接、更满意的检索结果,具有一定的意义。

4 结束语

针对现有语义检索模型在查询扩展中存在的不足,提出了一种关联多本体的科技奖励检索方法。通过改进的关联多本体间概念的相似度计算方法来对扩展的用户检索请求进行关联、筛选,得出最能表达用户检索意图的检索词。通过实验结果表明确实提高了信息检索的准确率。同时,由于本体构建工作的复杂和评价指标的不确定性,只对农业项目的领域本体进行了构建,如果在其他领域本体完备的情况下,该方法适用于其他领域的检索。



图6 关联多本体检索界面

Fig.6 Associating multiple-ontologies retrieval interface

参考文献:

- [1] NECHES R,FIKES R E,GRUBER T R,et al.Enabling Technology for Knowledge Sharing[J].AI Magazine,1911,12(3):38-56.
- [2] WALTMAN L,VAN ECKN J. Some comments on the question whether cooccurrence data should be normalized[J]. Journal of the American Society for Information Science and Technology,2007, 58(11):1701-1703.
- [3] 马骧飞,刘淑丽,孙滨.基于 Ontology 的网络教育资源语义检索模型研究[J].计算机与数字工程,2012,40(12):79-82.
- [4] 田宏,马朋云.基于 Jena 的城市交通领域本体推理和查询方法[J]. 计算机应用与软件, 2011,28(8):57-59.
- [5] 于静,吴国全,卢焱.基于领域本体的政务信息检索系统[J].计算机应用,2010,30(6):1664-1167.
- [6] 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.
- [7] 谢文玲,潘建国.基于语义相似度的个性化信息检索方法[J].计算机应用与软件,2011,28(5):161-164.
- [8] 凌仕勇,龚锦红. 图解析方式的复合本体映射策略研究[J].华东交通大学学报,2013,30(3):82-88.
- [9] 胡川溯,符云清,钟明洋.基于领域本体的语义查询扩展[J].计算机应用,2012,21(7):83-89.
- [10] 韩韧,黄永忠,刘振林,等.OWL本体构建方法的研究[J]. 计算机工程与设计,2008,29(6):1397-1400.
- [11] MICHAEL K SMITH,CHRIS WELTY,DEBORAH L MCGUINNESS. OWL Web Ontology Language Guide[EB/OL]. (2004-02-10)[2013-10-20]. <http://www.w3.org/TR/2004/REC-owl-guide-20040210/2009>.

- [12] BATET M,SANCHEZ D,VALLS A. An ontology-based measure to compute semantic similarity in biomedicine[J].J Biomed Inform, 2011,44(1):118-125.
- [13] RODRÍGUEZ MA,EGENHOFFER MJ. Determining semantic similarity among entity classes from different ontologies[J]. IEEE Trans Knowl Data Eng, 2003,15(2):442-456.
- [14] AL-MUBAID H, NGUYEN HA. Measuring semantic similarity between biomedical concepts within multiple ontologies[J]. IEEE Trans Syst Man Cybern, 2009,39(4):389-398.
- [15] LAMBRIX P,TAN H. A tool for evaluating ontology alignment strategies[J]. J Data Semant, 2007,182:182-202.

A Science and Technology Award Retrieval Method of Associating Multiple-Ontologies

Zeng Hui, Jiang Jianjian, Xiong Liyan, Huang Weichun

(School of Information Engineering,East China Jiaotong University,Nanchang 330013,China)

Abstract: The existing science and technology award retrieval is based on keywords matching technology, which ignores the semantic understanding of user queries. To solve the problem of traditional information retrieval and combine characteristics of current practical application-oriented and ontology-based semantic retrieval, this paper proposes a science and technology award retrieval method of associating multiple-ontologies. By constructing the domain ontology for science and technology awards projects and by improving the concept similarity calculation method and associating multiple-ontologies concepts, the proposed retrieval method makes the query expansion words further express retrieval intention of users. The experimental results show the method obtains better precision and recall compared with keywords query and single-ontology extensions query.

Key words: science and technology awards; associating multiple-ontologies; ontology construction; concept similarity