

文章编号: 1000-5013(2012)01-0027-06

一种基于农业领域本体的语义检索模型

陈叶旺, 李海波, 余金山

(华侨大学 计算机科学与技术学院, 福建 厦门 361021)

摘要: 基于农业领域本体和词汇, 给出一个针对农业领域的语义检索改进模型. 该模型接受用户自然语言输入, 通过计算词语与本体知识之间的相关度和相似度, 来实现词汇到知识的映射. 针对农业领域本体特点, 给出相应的语义扩展规则, 把检索结果从本体知识扩展到与之相关的资源文档, 并对检索结果进行排序.

关键词: 本体; 语义扩展; 语义检索模型; 农业领域

中图分类号: TP 391.3

文献标志码: A

语义检索是提供信息服务的重要组成部分, 是当前一个研究热点^[1-9]. 然而, 目前大部分的基于本体的语检索方法只是用本体来提供一些表达能力较浅的信息空间, 或者只是在本体表达的知识中做是非判断, 存在着很大的局限性. 在过去的一段时期内, 在语义检索这个方面上取得的成果, 仅仅是在基于本体的知识系统中使用部分本体的表达能力^[10], 或是采用基于布尔检索模型^[9]. 这使得被检索的对象要么符合条件, 要么不符合条件. 对于第 1 种模型而言, 它使用部分本体的表达能力, 而本体只是用来提供一些表达能力较浅的信息空间, 本质上等同于词汇分类和词汇表. 对于第 2 种模型而言, 它存在一个很大的局限性, 即很难把目前存在的海量非结构化知识完全用本体形式化地表示^[8]. 在自然语言词汇到本体知识映射方面, 这两种模型多是以字符串匹配方式简单而直接地完成映射, 使得映射成功率相对较低. 另外, 这两个模型都没有提供一个有效的检索结果排序, 使得最终用户很难区分结果好坏. 对于用户来说, 能否方便地实现检索语句的构造, 准确地表达出自己的检索需求, 关系到能否检索到相关的结果. 采用自然语言问句的形式, 是目前表达检索意图的最好形式. 本文给出一个基于农业领域本体与词汇的改进检索模型, 并进行相应的实验.

1 用户输入自然语言处理

系统的自然语言处理, 包括用户查询预处理、问题分类、去除无意义字和词、中文分词等主要功能.

1) 预处理. 去除需求文本中不能被识别的成分, 包括空白字符、空格、TAB 和特殊字符.

2) 问题分类. 系统对每个领域知识的查询都定义了一系列查询关键词, 根据这些关键词, 可以从用户的输入中辨别出查询领域.

3) 去除无意义字和词. 属于经验处理, 主要是去除平凡词, 如“是”、“应该”、“可以”, 以及“的”、“地”、“得”等. 平凡词所在的语义表达的正确性和重要性较小, 可以去除. 平凡词列表是根据经验知识得到的, 可以不断提炼修改.

4) 中文分词. 这一步主要基于两个词库, 一个是中文基本词库, 其词汇数量有 119 850 个; 另一个是联合国粮食及农业组织(Food and Agriculture Organization, 简称 FAO)提供的中文农业词库, 其词汇数量有 37 060 个. 在分词过程中, FAO 提供中文农业词库优先级比中文基本词库高, 即先以 FAO 中文农业词库为准.

收稿日期: 2011-07-03

通信作者: 陈叶旺(1978-), 男, 讲师, 主要从事语义检索与数据挖掘的研究. E-mail: ywchen@hqu.edu.cn.

基金项目: 福建省自然科学基金资助项目(A0810013); 福建省农业科技重大项目(2010N5008); 华侨大学高层次人才科研启动项目(09BS619)

2 词汇-本体知识映射

在经过自然语言分词之后,得到的是词汇集合,须把这些词汇转化为本体知识库中对应的知识点上,这就需要所谓的映射工作.映射结果通常有如下 3 种情况:1) 所有关键词是领域本体知识库中的元素;2) 部分关键词是领域本体知识库中的元素;3) 没有关键词是领域本体知识库中的元素.

对于本体知识库中不存在的词汇,可以通过词汇的相似度计算,匹配相关的词汇形成映射关系.文中采用的映射方法,是从两个角度来考查自然词汇-本体知识之间映射的关系,即一方面从词汇-标注文档-本体知识相关程度来看词汇-本体知识的联系紧密程度;另一方面基于知识关系词典考查词汇-本体知识两者之间的相似程度.

2.1 基于词汇-标注文档-本体知识的相关关系

在词汇-标注文档-本体知识的相关关系中,一个词汇可能被包含在多个文档中,而每个文档又可被一个或多个本体知识标注.通过统计包含词汇的文档所属的本体知识,可以统计出这个词汇对不同本体知识的相关程度.这种相关程度说明了词语-本体知识间的联系紧密程度.为计算这种相关关系,通过以下几个假设来说明一个自然语言词汇对一个本体知识的相关关系.

假设 1 一个词汇 w 通过文档映射到的本体知识个数越多,它对单个本体知识的相关度越低.

假设 2 一个词汇 w 在一个受本体知识 e 标注过的文档中的词频越高, w 与 e 的相关程度越高.

假设 3 若一个词汇 w 与文档 d 相关,则标注文档 d 的本体知识 e 与文档 d 的相关度越高, w 与 e 之间的相关程度也就越高.

假设 4 一个词汇 w 在越多的受本体知识 e 标注过的文档中存在, w 与 e 之间的相关程度越高.

假设 5 出现过词汇 w 的文档与受本体知识 e 标注过的文档交集越大, w 与 e 的相关程度越高.

假设 1 从词汇在知识空间的分布情况来分析,一个词语与越多的知识关联,它对概念的区分性就越不明显,与单个知识的相关程度也就越低.假设 2 在与某个本体知识相关的文档空间中,对词汇进行词频统计.这样统计粒度细,区分性强,则可以更准确地刻画这个词对概念的所属程度.假设 5 与假设 2 的细粒度角度不同,假设 3 与假设 4 考虑的角度是粗粒度的相关文档数目.这样从粗细不同的角度能更全面地考查词汇与本体知识之间的相关度.据以上假设给出词汇 w 和本体知识 e 的相关程度计算方法.

设 $DS = \{d_1, d_2, \dots, d_m\}$ 表示一个文档库; \succ_e^d 表示知识 e 标注了文档 d , $NPRO(e, d)$ 表示知识 e 标注文档 d 的相关度,则有

$$W2E(w, e) = \text{sum Freq}(w, e) \times \log\left(\frac{1}{|\text{Rel ES}(w)|} + 1\right) \times \log\left(\frac{\sum_{d_i \in \chi} \text{NPOR}(e, d_i)}{|\lambda|} + 1\right). \quad (1)$$

式(1)中: $\text{Ann RS}(e) = \{d | \succ_e^d \wedge d \in DS\}$ 为 e 标注过的文档集合; $|\text{Ann RS}(e)|$ 表示 e 标注过的文档数量; $\text{count}(w, d)$ 表示词语 w 在文档 d 中出现的次数; $\text{len}(d)$ 表示文档 d 的长度; $\text{sum Freq}(w, e) = \sum_{i=1}^{|\text{Ann RS}(e)|} \frac{\text{count}(w, d_i)}{\text{len}(d_i)}$, $d_i \in \text{Ann RS}(e)$, 表示词语 w 通过文档映射到 e 的词频统计量; $\text{Rel RS}(w) = \{d | (\frac{\text{count}(w, d)}{\text{len}(d)} > \text{threshold}) \wedge d \in DS\}$ 表示 w 词频大于 threshold 的文档集合, $|\text{Rel RS}(w)|$ 表示这个文档集中文档的数量; $\text{Rel ES}(w) = \{e | \exists d, s, t, d \succ_e^s \wedge e \in \text{Rel RS}(w)\}$ 表示词汇 w 映射到本体知识的集合, $|\text{Rel ES}(w)|$ 表示 w 映射到的本体知识数量; $\chi = \text{Rel RS}(w) \cap \text{Ann RS}(e)$; $\lambda = \text{Rel RS}(w) \cup \text{Ann RS}(e)$.

算法 1 的词汇-文档-本体知识相关度计算伪代码,如表 1 所示.

2.2 基于知识关系词典的词汇-本体知识相似度

在语义信息缺乏的情况下,知识词典只能实现词汇到本体知识的多对一映射,从而解决自然语言中的多词同义现象.例如,将电脑和计算机都映射到本体概念“computer”上,再比如“叶子”和“叶片”都映射到本体知识的“leaf”上.然而,自然语义的一词多义问题就比较难解决了,如苹果一词可以同时表示苹果计算机和一种植物.概念映射的另一个困难是概念词典常常无法覆盖所有的用户词汇,用户可能会使用生僻的词汇来表示心目中的概念,由此可能产生无法被概念词典识别的“孤儿词汇”.

因此,基于知识关系词典的映射应该取决于如下两点:1) 字符串匹配,若词汇与领域本体中的某个实体的标签及扩展表示字符串相似,则将这个本体实体作为候选映射元素;2) 自然语言上下文与本体语义上下文之间的关系. 如果候选映射元素集合之间存在着领域本体中所定义的语义关系,那么可以认为映射成功可信度高;而如果它们之间相互孤立,则认为成功可信度低.

定义 1 知识关系词典. 一个本体实体 e 的知识关系词典 L_e 是由一个词汇集合 T_e 和一个语义环境 Context_e 组成,即 $L_e = T_e \cup \text{Context}_e$. 一个词汇 t 出现 Context_e 中,则记为 $t \angle \text{Context}_e$.

定义 2 词汇-知识相似度. 词汇-知识相似度函数 $\text{SIM} : w \rightarrow e$, 计算词汇集合 Γ 中一个词汇 $w \in \Gamma$ 与本体实体 e 相似度,有

$$\text{SIM}(w,e) = \begin{cases} 0, & \neg \exists w' \in T_e, s. t. \text{SrtSim}(w,w') > \text{threshold}, \\ (1-\lambda) \times \text{MAXSTRSIM}(w,e) + \lambda, & \text{其他}. \end{cases} \tag{2}$$

其中: $\text{MAXSTRSIM}(w,e) = \text{Max}(\text{SrtSim}(w,w'_1), \text{SrtSim}(w,w'_2), \dots, \text{SrtSim}(w,w'_{|T_e|}))$; $w', \dots, w'_{|T_e|} \in T_e$; $\lambda = |\Phi|/|\Gamma|$, 而 $\Phi = \{w' | w' \angle \text{Context}_e, w' \neq w\}$.

依据以上工作,得出算法 2 的词汇-本体知识相似度计算伪代码,如表 1 所示.

表 1 算法的词汇-文档-本体知识相关度计算伪代码

Tab. 1 Pseudocode of algorithm of calculating word-document-ontoloy relation

	Rel RS(w)	Rel ES(w)	词汇集合-文档-本体知识相关度
算法 1	输入数据 1: 词汇 w, 本体 o 输入数据 2: 文档集合 DS, 输出数据: 与 w 相关的文档集合 result={}; For each document d in DS do freq+=count(w,d)/len(d) If (freq>threshold) 把文档 d 加到集合 result 中 End 返回 result	输入数据 1: 词汇 w, 本体 O 输入数据 2: Rel RS(w) 输出数据: 与 w 相关的文档集合 result={}; For each document d in Rel RS(w) For each entity e in O do If NPRO(e,d)>threshold 把本体知识 e 加到集合 result 中 取下一个文档 End 返回 result	输入数据 1: 词汇 w, 本体实体 e 输入数据 2: Rel RS(w), Rel ES(w), Ann RS(e) 输出数据: w 与 e 的相关度 result=0;1 sumFreq=0 For each document d in Ann RS(e) do freq +=count(w,d)/len(d) sumFreq+=freq End 计算 $\chi = \text{Rel RS}(w) \cap \text{Ann RS}(e)$ 计算 $\lambda = \text{Rel RS}(w) \cup \text{Ann RS}(e)$ 在 χ 中计算 $\sum_{d_i \in \chi} \text{NPOR}(e, d_i)$ 按式 1 计算 result=W2E(w,e) 返回 result
算法 2	词汇集合-文档-本体知识相关度		
	输入数据 1: 本体知识库 O, 本体实体 e 输入数据 2: 自然语言词汇集合 ts 和词汇 w 输出数据: w 映射到 e 的相似度 获取实体 e 的相关词汇集合 T_e 和 Context_e . 初始化最大似度 max_r=0 for each w' in T_e do 计算 w 与 w' 相似度 r if r 大于 max_r do max_r:=r end.	if max_r< 阈值 do return 0; for each triple p in Context_e do for each word w' in ts If($w' \neq w$ 且 w' 出现在 p) 中 do $\Phi = \Phi + \{w'\}$ end end 按式(2)计算最后结果 result 返回 result	

2.3 词汇-本体知识映射相关度

词汇-标注文档-本体知识相关关系体现的是词汇与本体知识之间的联系紧密程度,而基于知识关系词典的词汇-本体知识相似度体现的是两者之间的相似程度. 二者分别从不同的角度考查两者之间的关系. 在这里,可以认为两者对词语-本体知识映射相关度具有相同的影响力,两种因素相互作用的结果更能有效说明词汇映射相关度的强弱. 因此,采用两种因素直接相乘的方式来定义词语-本体知识映射相关度,即

$$\text{Map}(w,e) = \text{W2E}(w,e) \times \text{SIM}(w,e), \tag{3}$$

3 基于领域本体的语义扩展

根据疑问对象和疑问焦点在本体知识库中映射的距离,可分为直接关系检索和间接关系检索. 在检索处理中,两种关系检索方式的难度和实现策略不同. 直接检索的实现较为简单,可通过相似度计算将疑问对象和疑问焦点映射为三元组的主体和谓词结构进而生成检索表达式. 有时直接检索结果不能满

足用户需求,就需要根据领域本体中知识关系进行语义扩展,然后提交给检索表达式生成模块,组成 SPARQL 表达式进行检索,即扩展检索. 文中的工作是把两者合并在一块,通过相关度体现出不同.

本体实体分为类、属性、实例 3 种,其扩展方式和扩展假设有所不同.

1) 类(概念)扩展. 对一个概念实体 e ,可以从知识库中选取与之相关的概念作为 e 的扩展,如上下位概念、等同概念(包括 owl:sameAs 和 owl:equivalentTo)、参照概念(owl:seeAlso)等.

2) 属性扩展. 与类的扩展类似,对一个属性实体 e ,可以从知识库中选取与之相关的属性作为 e 的扩展,如上下位属性、等同属性(包括 owl:sameAs 和 owl:equivalentTo)、参照属性(owl:seeAlso)等.

3) 实例扩展. 除了可以通过等同关系、参照关系实现扩展外,还可以通过一些特殊属性关系推导出与其有等同关系的实体.

在农业领域本体中常用到这些特殊属性,主要有 FunctionalObjectProperty, SymmetricObjectProperty 和 Transtivity Property.

1) FunctionalObjectProperty(功能性属性):对于一个 owl:FunctionalObjectProperty P 而言,如果 $P(X, Y)$ 与 $P(X, Z)$ 都成立, $Y=Z$. 那么,根据这种逻辑关系,可以把使用 owl:FunctionalObjectProperty 关系可以推导出与实体 e 等同的其他实体作为 e 的扩展集.

2) SymmetricObjectProperty(对称性属性):与 FunctionalObjectProperty 类似,可以把使用 owl:SymmetricObjectProperty 关系可以推导出与实体 e 等同的其他实体作为 e 的扩展集.

3) Transtivity Property(传播性属性):对于一个 owl:TranstivityProperty ObjectProperty P 而言,如果 $P(X, Y)$ 与 $P(Y, Z)$ 都成立,则 $P(X, Z)$ 成立. 根据这种逻辑关系,可以把使用 owl:TranstivityProperty 关系可以推导出与实体 e 相关的其他实体作为 e 的扩展集.

4 生成 SPARQL 形式化查询语句

SPARQL 现在已成为 W3C 的推荐标准,文中方法的查询最终都转化成 SPARQL. 因为用户的提问方式多种多样,不同的问题也会有不同的检索策略. 问句中的关键词可能是本体中类、属性、实例中任何一种. 问句(关键词)提交后,将启动问句解析模块分析问句,过滤无意义词汇,分解成词汇集合,进行词汇-本体知识映射,确定知识关系;然后,根据问句的疑问词及句法和语法特征对问题进行分类,确定问题检索策略和答案的组织方式. 最后,系统访问领域本体库,判定哪些关键词是本体库中包含的类、属性、个体,进而对用户提问概念进行规范化.

标准的中文问句结构有一定的规则,而本体中三元组<Subject, Predicate, Object>的形式化表示正好符合问句的表达顺序. 这就为问句成分映射成本体知识库的词汇提供了条件. 在 SPARQL 的检索表达式中,三元组处在后方顺次排列,其中的未知变量就是问题所在的位置.

疑问对象是问题的主体,一般问句的提问针对的是主体的某个属性. 尽管用户提问的方式多种多样,但归纳起来主要有两种排列句型:一种是疑问对象位于句子的前部,疑问焦点排列在疑问对象的后面,句子的尾部是疑问词,问题的答案就是疑问词所指代的信息,如白斑病怎么治;另外一种问句的提问方式与此相反,疑问词位于句子的前部充当疑问对象,疑问焦点位于疑问词的后方,句子的尾部为疑问对象的相关信息,如什么药能治白斑病. 虽然两种句型的排列顺序不同,但其疑问的意向是相同的,句中的疑问成分也一致,经过问句解析后形成的结果,如表 2 所示.

表 2 解析示例
Tab. 2 Examples

什么药能治白斑病?	疑问词(inquiryword): 什么 疑问对象(topic): 药 疑问焦点: 治 答案类型: 简单查找(simple) 三元组: { ? medicine 治 白斑病} 疑问限制: 无	SPA-RQL	Prefix pf: <http://www. owl-ontologies. com/agr. owl#> SELECT ? medicine FROM <http://www. owl-ontologies. com/agr. owl#> WHERE { ? medicine pf: 治 pf: 白斑病}
-----------	--	---------	---

在进行三元组元素的映射时,疑问对象放置在三元组主体的位置,中间的谓词由疑问焦点充当,客体的位置是未知变量,也就是问题的所在之处. 当用户从接口输入检索关键词时,可以直接将输入的关

关键词与本体中的词汇进行相似度计算.

如果依据用户输入的内容找不到结果,则按前述的方式进行语义扩展,对每一个实体扩展的集合取笛卡尔乘积,即 $M_1 \times M_2 \times \cdots \times M_n$. 其中 M_i 为第 i 个实体的扩展集合. 对乘积结果中的每组元素,重新按上述方法创建查询语句,并根据扩展实体与原实体的相似度计算新生成的查询语句与原始查询语句相似度.

5 扩展查询结果

如前所述,过去的一段时期内语义检索方法的检索结果缺乏合适的排序,不存在那种可以用百分比来表示检索结果可信度的答案. 文中的检索模型与布尔语义检索系统不同之处在于,返回排序的结果不仅仅是知识库中用本体所描述的相关知识,还有与知识相关信息资源,排序过程以每条结果项的综合相关度为基准. 每个结果项的综合相关度包含两部分,即

$$\text{SIM}(\text{result}_{\text{ind},d}) = W \times \text{SIMSPARQL}_{\text{ind}} + (1 - W) \times \text{NPROR}(\text{ind},d). \tag{4}$$

式(4)中: $\text{result}_{\text{ind},d}$ 表示结果项,包含知识 ind 与文档 d ; $\text{SIMSPARQL}_{\text{ind}}$ 为生成的 SPARQL 的查询语句与原始查询的相似度,该查询语句的执行结果中包含实例 ind; $\text{NPROR}(\text{ind},d)$ 为文档 d 与实例 ind 之间的相关度; W 为权重.

6 实测评价

根据前面的工作,实现一个现向农业领域的语义检索系统. 开发工具是 MyEclipse 6.0,本体知识用 OWL 表达,资源标结果存储于 MySQL 数据库中. 目前,在语义搜索领域还没有一个公认的测设数据集和评价方法,而文中的工作也只是针对特定的农业领域. 因而,测试时使用的数据都是自己建立的农业相关领域本体知识及农业相关的资源,而这些资源都经过语义标注工具或手工方式标注过.

本体知识使用的农业病虫害本体, # Concept 和 # Individual 的数量分别为 274, 3 730;而资源是相对应的农作物病虫害知识文档,其数量为 1 119. 使用 $\text{Precision}@ (n,k)$ 和 $\text{Recall}@ (n,t)$ 作为主要的评价方法和指标,其计算式为

$$\text{Precision}@ (n,k) = (\text{前 } n \text{ 个正确的检索结果})/n, \tag{5}$$

$$\text{Recall}@ \{n,t\} = \frac{\alpha \cap \beta}{\beta}. \tag{6}$$

其中: $\text{Precision}@ (n,k)$ 表示前 n 个结果中相似度大于 k 的查准率, $\text{Recall}@ (n,t)$ 表示前 n 个结果中相似度大于 k 的查全率; α 为语义检索的前 n 个结果中相似度大于 k 的集合; β 为人工判断的检索结果中前 n 个结果.

表 3 为查询实验数据. 由表 3 可以看出,检索方法取得了一定的效果,当查询语句比较简单且能按查询示例结构输入时,查询结果比较准确.

表 3 查询实验数据
Tab. 3 Query data

查询语句	结果	Precision@ (n,k)			Recall@ (n,k)		
		n	k	$P/\%$	n	k	$P/\%$
黄瓜白斑病怎么治	5	5	0.40	95	5	0.35	90
水稻要注意哪些病	42	10	1.00	100	9	1.00	100
白菜叶子发霉	8	8	0.40	85	7	0.50	75
白菜生长有哪些特点	10	6	0.60	80	8	0.40	77

7 结论

目前,大部分的基于本体的语检索方法很难把存在的海量非结构化知识完全用本体形式化地表示. 同时,在自然语言词汇到本体知识映射方面,现有的方法多是以字符串匹配方式简单而直接地完成映射,使得映射成功率相对较低. 此外,对于检索结果也都没有提供一个有效的排序,使得最终用户很难

区分结果好坏. 针对这些问题,文中给出一个基于农业领域本体和词汇的改进检索模型.

该模型接受用户自然语言输入,通过计算词语与本体知识之间的相关度和相似度,来实现词汇到知识的映射;针对农业领域本体特点,给出相应的语义扩展规则;把检索结果从本体知识扩展到与之相关的资源文档,并对检索结果进行了排序. 基于这个模型,实现一个针对农业领域的检索系统,并进行一些小规模的实验,取得了良好的效果.

参考文献:

[1] HEFLIN J, HENDLER J. Searching the web with SHOE[C] // Proc of AAAI 2000 Workshop on AI for Web Search. Austin: AAAI Press, 2000; 35-40.

[2] SHAH U, FININ T, JOSHI A, et al. Information retrieval on the semantic web[C] // Proc of the 11th International Conference on Information and Knowledge Management. New York: ACM Press, 2000; 461-468.

[3] GUHA R, Mc COOL R, MILLER E. Semantic search[C] // Proc of the 12th international conference on World Wide Web. New York: ACM Press, 2003; 700-709.

[4] PICARD J, SAVOY J. Enhancing retrieval with hyperlinks: A general model based on propositional argumentation systems[J]. Journal of the American Society for Information Science and Technology, 2003, 54(4): 347-355.

[5] LOSADA D E, BARREIRO A. A logical model for information retrieval based on propositional logic and belief revision[J]. The Computer Journal, 2001, 44(5): 410-424.

[6] POPOV B, KIRYAKOV A, OGNYANOFF D, et al. KIM: A semantic platform for information extraction and retrieval[J]. Journal of Natural Language Engineering, 2004, 10(3/4): 375-392.

[7] BERNERS-LEE T, HENDLER J, LASSILA O. The semantic web[J]. Scientific American, 2001, 284(5): 34-43.

[8] VALLET D, FERNÁNDEZ M, CASTELLS P. An ontology-based information retrieval model[J]. ESWC, 2005, 3532: 455-470.

[9] DAVIES J, FENSEL D, BUSSLER C, et al. The semantic web: Research and applications[M]. Berlin: Springer-Verlag, 2004; 473-487.

[10] CHRISTOPHIDES V, KARVOUNARAKIS G, PLEXOUSAKIS D, et al. Optimizing taxonomic semantic web queries using labeling schemes[J]. Journal of Web Semantics, 2004, 1(2): 207-228.

A Semantic Retrieval Model Based on
Agricultural Field Ontology

CHEN Ye-wang, LI Hai-bo, YU Jin-shan

(College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China)

Abstract: A semantic retrieval model is proposed based on the ontology and vocabulary of agriculture domain. This model provides an interface for user to input natural language, then it maps identified keywords to ontology entities by calculating the correlation and similarity between them; furthermore, we give a set of rules for semantic extension based on the features of agricultural ontology; and we extend and order the result got from the ontology to the annotated documents.

Keywords: ontology; semantic extension; semantic retrieval model; agricultural domain

(责任编辑: 陈志贤 英文审校: 吴逢铁)