第35卷 第11期 计 算 机 工 程 2009年6月
Vol.35 No.11 Computer Engineering June 2009

·人工智能及识别技术· 文章编号：1000—3428(2009)11—0184—03 文献标识码：A 中图分类号：TP242

面向垂直搜索引擎的基于知识的语义关联算法

高一波[1]，赵先章[1,2]，孙 硕[1]，黄 河[3]

(1. 中国科学院自动化研究所，北京 100190；
2. 中国农业大学信息与电气工程学院，北京 100083；
3. 哈尔滨工业大学计算机科学与技术学院，哈尔滨 150001)

摘 要：提出一种面向垂直搜索引擎的基于知识的语义关联算法，以知识表达技术及语义关联度计算为核心，建立语义关联规则，在语义扩展基础上提高查询招回率和语义相关度计算高检索的准确性，同时根据农产品物流领域的特点，设计并实现了用于农产品物流ASP平台的垂直搜索引擎。

关键词：概念知识树；知识表达；语义计算；垂直搜索

Vertical Search Engine-aimed Semantic Correlation Algorithm Based on Knowledge

GAO Yi-bo[1], ZHAO Xian-zhang[1,2], SUN Shuo[1], HUANG He[3]

(1. Institute of Automation, Chinese Academy of Sciences, Beijing 100190;
2. College of Information & Electrical Engineering, China Agricultural University, Beijing 100083;
3. College of Computer Science & Technology, Harbin Institute of Technology, Harbin 150001)

【Abstract】A new vertical search engine-aimed semantic correlation algorithm based on knowledge is presented, which emphasizes knowledge representation technique and semantic association calculation to build up semantic rules. The query recall rate and the retrieval accuracy of both semantic extension and semantic association calculation are promoted. According to the characteristic in fields of agricultural products logistics, the vertical search engine in the Application Service Probider(ASP) platform is designed and implemented.

【Key words】concept knowledge tree; knowledge representation; semantic calculation; vertical search

1 概述

垂直搜索引擎是种针对某特定领域业务而设计的专业搜索引擎。本文依托的农产品物流ASP服务平台，以数据中心及互联网为信息资源，包含自主建站、应用服务管理和信息交互引擎等多项关键技术，为农产品流通领域的各类用户在多个环节提供信息支撑技术方面的集成服务。对农产品物流相关信息进行有效检索是农产品物流ASP平台的基础功能，有助于农业企业间实现信息的资源共享，跨越企业间地域差别使农业企业更好地承担起农业产业化的重任[1]。垂直搜索引擎较综合搜索引擎获取信息更为有效[2]。

针对农产品物流这一特定领域，垂直搜索引擎更符合农产品物流ASP平台功能的要求。传统的搜索引擎通过目录、索引和关键词匹配来实现信息检索，缺少对搜索意向的语义理解，没有对信息进行基于领域知识的关联性分析，难于为用户提供全面准确的信息资源。以领域知识的关联性为支撑，从语义层面上对信息资源进行分析和整合，为信息用户提供包含语义关联和分类的检索服务是克服传统搜索引擎不足的重要手段。

知识表达是对人类认知在计算机系统中的表达与刻画方法，知识表达体系是实现语义分析计算的基础，可以改善搜索引擎获取信息的效率，长期为学界所关注：文献[3]提出的WordNet是种通用型的知识表达体系，在WordNet知识表达体系基础上，提出基于知识的信息搜索方法，其基本思想是计算搜索关键词与页面内容的相似度，这种相似度通过概念之间的语义距离来表征。“知网”是面向汉语及英语概念的知识表达体系，其基本思想是通过概念之间的关系以及概念属性之间的关系将知识组织成网。“知网”认为世界上一切事物都在特定的时间和空间内运动和变化，通常从一种状态变化到另一种状态，并由其属性的变化来体现。

本文在农产品物流领域进行垂直搜索，依托概念知识树表达模型，描述专业知识、组织知识内容，在领域知识和语言知识的基础上对信息内容做语义分析，获得具备语义关联及知识关联的搜索结果。在概念知识树的基础上，提出语义关联度计算模型，针对农产品物流这一特定的应用领域，建立起知识库，并开发领域垂直搜索引擎，以支持农产品物流领域的信息资源在语义层面上的融合。

2 知识表达体系

概念知识树是本文的基础模型，由中国科学院自动化研究所的研究团队提出。该模型针对人工智能领域的知识表达

基金项目：国家科技支撑计划基金资助项目“农产品物流企业信息门户ASP服务平台”(2006BAD10A0401)；“基于语义跨媒体检索与智能搜索平台”(2006BAH02A01)

作者简介：高一波(1973－)，男，助理研究员，主研方向：自然语言理解，语义信息处理；赵先章，博士后；孙 硕，工程师；黄 河，博士研究生

收稿日期：2008-11-20 **E-mail：**yibo.gao@ia.ac.cn

问题，从认知心理学的角度出发，借鉴传统知识表达方法，在提出思维活动的心理模型假说[4]的基础上，建立知识内容的语义表达方式和知识结构的语义计算方法。概念知识树模型在不断探索、完善表达机制与计算模式的同时，已经在以语义信息处理为技术核心的多个领域中得到了较好的应用。

概念知识树表达体系从本体论思想出发，作为表达知识的语义内容的核心，由 3 个层面的子模型组成，即知识本体模型、知识树模型和语义复合模型。

知识本体模型将基本概念作为表达的实体，基本概念以字、词为名称的语言载体，是表达知识的最小语义单元。每个基本概念使用属性、关系和行为 3 种要素表达概念的语义内涵，即概念={属性，关系，行为}。基本概念依靠名称、属性、关系和行为描述彼此间的语义关联，形成网状结构的基础知识层。

知识树模型是在知识本体的基础上构造的层次化分类表达结构，它从不同侧面描述知识块的语义组成成分，并反映知识节点间的语义分类关系。知识树是由知识节点组成的一个树状知识分布，每个知识节点使用基本概念或复合概念描述其语义内容，树枝表示知识节点间的语义关系(上下位关系或成员关系)。

知识树建立在基本概念的基础之上，是种高层语义分类知识的描述模型，如图 1 所示，概念知识树并非独立存在，也不是简单地对基本概念进行复合，而是借助知识树的内部结构，突出了概念的某些属性特征。通过基本概念间的关联，知识树之间形成节点间的横向语义关系，构成对领域知识的描述。知识树的内容由具体应用导向形成体系，是对具体领域、具体问题的分类知识的表达，形成了树状结构的应用知识层。

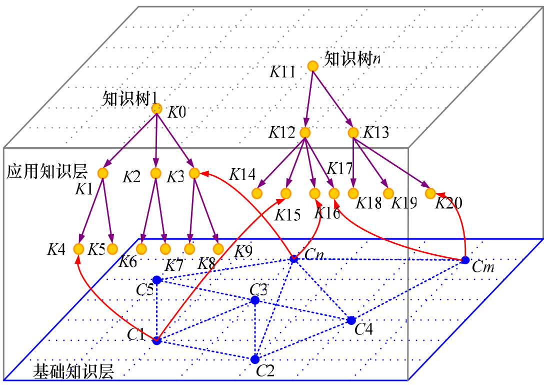

图 1　概念知识树模型

语义复合模型定义了由基本概念生成复合概念的组织规则，包括语义约束、语义状态和语义逻辑 3 种结构[4]。在知识体系的构造过程中，更多地使用语义约束规则，对名词性知识点给予描述，如“诺基亚手机”，“诺基亚”是对核心概念“手机”品牌属性的约束。该模型支撑知识内容由简单到复杂、由点到块的组织过程。

3　语义关联计算模型

基于上述知识表达模型，在领域知识内容的支撑下，提出一套量化计算知识点间语义相关性的算法模型，由同义概念扩展计算、属性关联计算和信息熵语义相关度计算 3 个部分组成。其中，前 2 种计算在检索系统中用于对检索词的语义关联分析，第 3 种计算用于数据预处理时的文本分类。此模型支撑语义关联、语义扩展和语义归并的量化计算，作为垂直搜索引擎中语义分析的技术基础，提供对信息内容进行分析的计算模型。

3.1　同义概念扩展计算

传统搜索引擎中对于检索词进行机械匹配，往往遗漏大量语义相关内容，特别是同义概念表达的内容，造成查全率较低。依靠概念知识树表达模型，本文建立 2 种“同义概念”计算方法：指称法和内涵法，以判断概念之间的语义等价性。

指称法是在给基本概念下定义时，建立“同义”和“反义”关系，当碰到相关概念时，再辅以“非”、“不是”等否定词组合即可完成同义概念的确定性推理计算，如用户输入“泰山”，则扩展得到“东岳”也是相关检索对象，于是搜索结果在同义概念层面得到扩展。

内涵法是通过属性判断概念的同义性，即计算 2 个概念的同名且同值属性的比例，是针对概念内涵重合程度的同义计算。概念内涵重合程度通过概念相似度来表征，计算过程如式(1)所示：

$$S(c_1,c_2)=\frac{A(c_1,c_2)}{a(c_1)+a(c_2)-a(c_1,c_2)} \tag{1}$$

其中，c_1, c_2 表示 2 个概念；S 表示概念相似度；$A(c_1, c_2)$表示 c_1，c_2 所具有同名且同值属性的个数；$a(c)$表示概念 c 的属性个数；$a(c_1, c_2)$表示概念 c_1, c_2 具有同名属性的个数，显然，$S(c_1, c_2)\in[0,1]$。

3.2　属性关联计算

用户的查询输入经常是由查询对象的各种属性组成，这些属性限定了查询的范围，则由属性推测查询的主题就显得十分重要。

针对这个问题，在响应用户的查询过程中使用属性关联计算，通过对属性的属性名称和属性值匹配分析，得到关联对象集合，在多属性匹配时对结果对象求与。

经统计，在查询过程中多以属性值作为输入，有时会连带属性名称。针对只出现属性值的情况，依靠先验领域知识辅助处理，即在知识库中建立描述属性值的类型与属性名称对应关系的知识树。对于属性值与属性名称同时出现的情况，使用语义约束模型的自动分析计算，得到“约束”成分为“属性值”，“核心”成分为“属性名”。

属性关联计算利用基本概念或知识点的属性内容，得到对应知识点，达到由明确内涵以缩小外延的语义分析的目的，进一步提高系统的查全率和查准率，如用户查询“500 万像素”，通过属性关联计算可查询到具有“500 万像素”属性的手机或照相机的相关结果。

3.3　基于信息熵的语义相关度计算

通过计算文本内容与知识节点之间的语义相关度，可以实现基于语义的信息内容分类。信息内容除与知识点直接匹配外，其语义内容还会沿知识树结构向上归纳，由下向上的语义的贡献值取决于下层语义点的实例分布情况，研究中借用信息熵来计算。基于信息熵的语义相关度计算过程如下：

$$S^{ij}=S^{ij}{}_o+S^{ij}{}_e \tag{2}$$

其中，S^{ij} 表示文本与知识树中第 i 层第 j 个节点的语义相关度，由自语义相关度 $S^{ij}{}_o$ 和扩展语义相关度 $S^{ij}{}_e$ 构成。自语义相关度由式(3)获得：

$$S^{ij}{}_o=m^{ij} \tag{3}$$

其中，m^{ij} 为第 i 层第 j 个知识点在给定文本中出现的次数。扩展语义相关度是子层节点对当前知识点的语义贡献值，由

式(4)给出：

$$S^{ij}_{e} = Cov_{i+1}H(S_{i+1}).\sum_{k=1}^{N_{i+1}} m_k \tag{4}$$

其中，Cov_{i+1} 为分析文本对知识树中第 i 层第 j 个节点的 $i+1$ 子层的语义覆盖度，由式(5)给出：

$$Cov_{i+1} = {M_{i+1}}/{N_{i+1}} \tag{5}$$

而 $H(S_{i+1})$ 为子层语义分布的信息熵，由式(6)给出：

$$H(S_{i+1}) = -\sum_{l=1}^{N_{i+1}} P_l \log P_l \tag{6}$$

在式(4)、式(5)中，N_{i+1} 为第 i 层第 j 个节点的 $i+1$ 层子节点的总数；M_{i+1} 为子节点在文本中出现的个数。式(6)是个信息熵模型，$H(.)$表示分析对象在子知识层的语义分布的均匀度，该值越大，表示在 $i+1$ 子层的语义分布越均匀，语义分散，而对上层的 S^{ij}_{e}“贡献”越大；H(.)值越小，则子层节点语义分布越不均匀，语义汇聚度越高，相对语义越明确，对上层语义相关度“贡献”相对越小；P_l 为子层第 l 个节点出现的概率，由式(7)给出：

$$P_l = {m_l}/{\sum_{k=1}^{N_{i+1}} m_k} \tag{7}$$

其中，分母同式(4)定义；m_l 为子层第 l 个节点出现的总次数。

4 搜索引擎架构

以概念知识树知识表达体系为核心的语义垂直搜索引擎架构如图 2 所示。

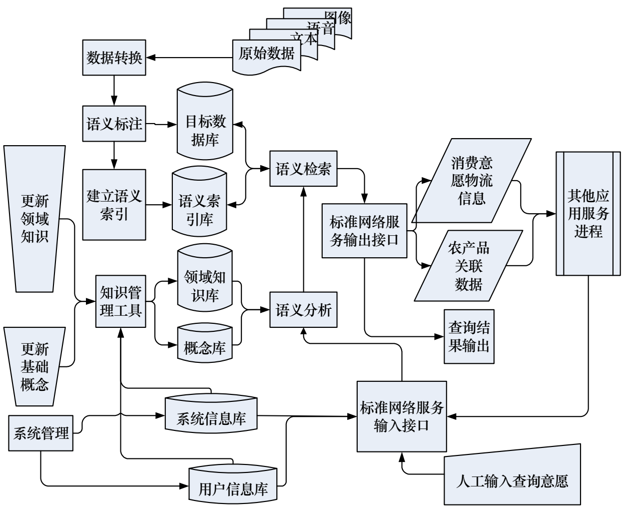

图 2　垂直搜索引擎构架

系统通过信息采集程序对农产品相关的站点进行原始数据采集与概念知识树表达体系进行语义匹配，建立索引数据库；当系统使用者通过查询界面输入查询意愿，同样地，经过概念知识树进行语义匹配获取使用者的查询语义，通过索引库将目标库中的相关内容呈现给使用者。

在以知识树作为表达模型构建的知识体系中包含知识树体系和概念库 2 个部分。通过以上知识表达模型，本文构建针对农产品物流领域知识库，包括基本概念近 50 000 个，为满足领域应用需要，填充内容定义的基本概念 200 个、知识树 50 棵、知识节点 7 000 个。

5 实验

本文收集了 1 500 篇各类文档作为测试集，验证以上语义关联计算模型的有效性。这些文档中涉及农产品物流、商务、计算机等多个学科，通过式(8)、式(9)计算出查询的准确率与招回率，查询 50 次获得相关试验结果列于图 3、图 4 中。

$$Pre = \frac{|Retrieved \cap Relevant|}{|Retrieved|} \tag{8}$$

$$Rec = \frac{|Retrieved \cap Relevant|}{|Relevant|} \tag{9}$$

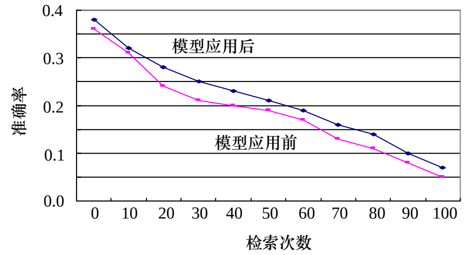

图 3　查询准确率曲线

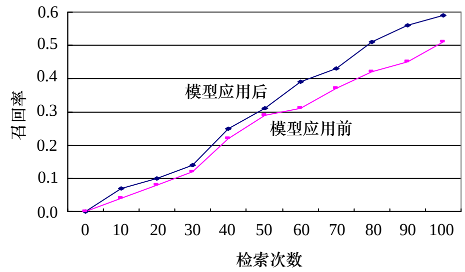

图 4　查询招回率曲线

从图 3、图 4 可以看出，在领域知识支撑下，应用语义关联算法对目标文本和查询短语进行语义分析后，查询的准确率与招回率均有所提升。

6 结束语

本文在概念知识树表达体系的基础上，建立一套语义关联计算模型，实现基于语义计算的信息分类算法，在以上核心思想基础上设计并实现为农产品物流ASP平台提供服务的垂直搜索引擎系统，通过实验证明了算法的合理性和有效性。本文的语义关联计算模型有助于提高搜索引擎的召回率和准确率。垂直搜索引擎被应用于农业物流 ASP 服务平台，有效增强了该系统的功能性，推动我国农业信息化建设的发展。同时，领域知识内容的构建方法具有通用性，可扩展到其他应用领域，支持架构相关的垂直搜索引擎，即本系统具备较好的可移植性。

参考文献

[1] 王红蕾，何东键．农业信息化 ASP 服务平台研究与建模[J]．中国管理信息化, 2007, 10(2): 14-16.

[2] 刘　畅．综合搜索引擎与垂直搜索引擎的比较研究[J]．情报科学, 2005, 25(1): 97-102.

[3] Miller G. A Lexical Inheritance System[J]. International Journal of Lexicography, 1990, 3(4): 245-264.

[4] 高一波．一种基于概念的知识表达体系[J]．微电子学与计算机, 2004, 21(9): 71-74.

编辑　陈　文