电子信息工程、管理工程 文章编号: 1000-8608(2007)04-0598-07

领域本体信息源选取方法研究与实现

邢 军[1,2], 韩 敏*[1], 周开朋[1]

(1.大连理工大学 电子与信息工程学院, 辽宁 大连 116024;
2.大连工业大学 信息科学与工程学院, 辽宁 大连 116034)

摘要: 构造领域本体所需的信息源选取方法的研究对解决本体的构造质量、构造效率等问题,以及推广与发展领域本体有着重要意义. 传统的信息源文档选取方法只考虑概念因素,不能很好地解决该问题. 因此,首先利用抽象方法分析了领域本体所需信息源具有的概念性、关系性和预测性等特点. 然后,针对这些特点分别采用改进的VSM方法、基于本体关系距离的方法以及神经网络方法计算文档权值. 最后,通过编写的软件OnMaker产生模拟数据得到概念、关系和预测3个权值,从而计算出每个文档权值,并使用与"湿地保护"相关的真实文档验证该模型,达到了较好排序选取的效果.

关键词: 领域本体; 信息源; 本体构造
中图分类号: TP391 文献标识码: A

0 引 言

随着人们对本体研究的不断深入,本体在互操作、信息集成、知识工程、第二代WWW等方面[1-3]所起的重要作用得到越来越广泛的关注. 构造大量可实际应用的领域本体是一项紧迫而具有深远意义的工作. 本体的构造从开发模式上分为自顶向下、自底向上、中间出发等开发过程[4]. 构造方式有手工与(半)自动化构造两种方法[5]. 不论采取何种模式、方法都需要从大量的文档资料中提取相关的知识来完成本体的构造. 那么如何判定哪些文档资料具有更多的知识量需重点阅读,从而加快本体的构造速度呢?这是领域本体构造的源头问题,如果没有一个较好的解决方法,将直接影响本体的构造效率和质量. 这个问题也可归结为文档的排序选取问题.

为了解决这个问题,自20世纪80年代许多学者提出了各种方法. Lee等提出了应用向量空间模式的方法[6]; Egghe等[7]在VSM基础上提出使用模糊集技术完成文档排序的相似度方法. 这两种方法比Boolean模式有了很大的进步,但由于只考虑了术语在文档中的频率及含有术语文档在整个文档中的百分率,而没有考虑术语在文档中位置的信息,其精度不是很高. 为此, Park等[8]把术语出现的位置转换成信号信息,然后使用离散余弦转换(DCT, discrete cosine transform)方法完成文档转化,精度有了较大的提高. 另外, Danilowicz等[9]提出基于Markov链的文档排序方法, Alema-Meza等[10]提出基于关系因素的方法,以及罗三定等[11]提出基于概念的文档评价模型,都对该问题进行了深入的探讨. 上述传统方法主要考虑关键字在文档中出现频率、出现位置等因素,而构造本体所需信息源排序问题中本体结构所起的作用是不可忽视的,原因是如果把本体作为一种树型结构,由于概念、关系在树型结构中所处层次不同,其重要程度是不同的,传统方法没有区分其重要程度,这必将影响其准确度. 另外,传统方法只是考虑文档索引与查询索引的相似度,而忽略了文档特征值与所需文档信息的内在关系问题,也就是说,一篇文档重要程度的确定,与该文档是否含有构造本体所需的概念和关

收稿日期: 2005-11-05; 修回日期: 2007-06-12.
基金项目: 国家自然科学基金资助项目(60674073);国家重点基础研究发展规划("九七三")资助项目(2006CB403405);国家科技支撑计划资助项目(2006BAB14B05).
作者简介: 邢 军(1972-),男,博士生; 韩 敏*(1959-),女,教授,博士生导师, E-mail: minhan@dlut.edu.cn.

系直接相关,并且还应把含有概念和关系数量作为排序的一个判定依据. 只有在深入研究本体自身特点和文档信息源特点的基础上,才能够更好地解决上述问题. 为此,本文采用抽象方法、改进的 VSM方法、基于本体关系距离的方法以及神经网络方法对文档进行全面详细的分析.

1　领域本体所需信息源特点分析

从传统信息源的需求到构造本体所需信息源的转换,使得一些传统经典算法必须经过适当的改进才能适合现在的需要. 这就涉及两个问题: 第一个问题是构造本体所需信息源具有什么特点,以及本体知识在信息源中是如何分布的;第二个问题是如何对原有算法进行改进,从而适合构造本体的需要. 首先分析第一个问题,这个问题本质上是本体信息的表述与分布问题.

1.1　本体信息的表述

本体究竟应该包括哪些信息是和本体的定义相关的. Gomez-perez等[12]在本体定义的基础上归纳出用于描述本体的 5个基本建模元语: 概念、关系、函数、公理和实例等. 本体构造过程就是从信息源中提取上述 5个元语所表达的数据及规则信息. 由于本文只关心文档重要程度的判别,没有涉及本体知识提取问题,为了简化问题把本体信息的表述归纳为概念与关系两种形式.

1.2　信息源的概念性和关系性

本体的概念和关系信息在信息源中是如何分布的呢? 本文通过抽象化的方法来解决. 所谓抽象化就是对于一个问题以数学或形式化的方式加以说明、表示,使其具有可解性、一般性. 这里通过问题求解的领域应用本体类型来讲述抽象过程. 对于一个问题的求解,大多可以分解为输入条件、处理对象、输出结果等 3个部分,分别对应集合 X、集合 Y、集合 Z,称这种模型为输入输出需求驱动模型. 在该模型的基础上可以把文档集 D 中的一个文档分解成 3个概念子文档和 6个关系子文档,如图 1所示,也就是把单一文档中的概念、关系分布问题,转化为对其子文档的处理,大大简化了处理方法. 为了便于问题的分析、理解有如下定义:

定义 **1**　概念子文档 (CSD, concept sub-document),把一篇文档中的术语按照一个集合中的元素进行映射而得到的文档,称为概念子文档. 如 $c(X)$、$c(Y)$、$c(Z)$ 分别为含有输入集合 X 元素的文档、对象集合 Y 元素的文档以及输出集合 Z 元素的文档.

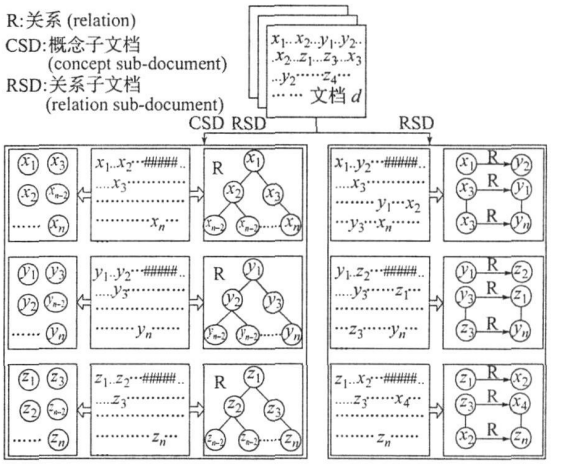

图 1　文档的分解与提取

Fig. 1　Decomposition and extraction of a document

定义 **2**　关系子文档 (RSD, relation sub-document),对于给定的一篇文档可以完成一个集合中的元素关系提取,也可以完成两个集合元素之间关系的提取而组成的文档,称为关系子文档. 如 $r(X)$、$r(Y)$、$r(Z)$ 为一个集合的关系子文档,$r(X,Y)$、$r(X,Z)$、$r(Y,Z)$ 为两个集合间的关系子文档.

定义 **3**　元素距离 d,在关系子文档中,两个有关系的元素出现在同一句子、同一个段落及同一篇文章中的不同位置具有不同相关权值,分别记为 d_1, d_2, d_3,并且 $d_1 > d_2 > d_3$.

定义 **4**　词的位置 p,元素在文档中出现在标题、句首和一般正文等不同位置,分别记为 p_1, p_2, p_3,并且 $p_1 > p_2 > p_3$.

通过上述对文档的分析可以得出,构造领域本体的信息源只有包含构造本体所需的概念和关系才是有效的,否则是无用的信息源. 也就是说这些文档必须具有概念性和关系性. 另外,需要说明的是本文在第 2章中提到的概念权值、关系权值的计算实际上已转化成对概念子文档、关系子文档权值的计算.

1.3　预测性

一般来说,信息源如果具备了概念性和关系性特点就足够了,但在实际构造本体工作中,还存在一个非常重要的特点: 本体构造过程可以描述为在专家的参与下完成初集本体的建立;然后,从给定的文档集 D 中提取相关知识,得到最终的本体. 在初集本体存在的情况下,能否通过对文档包含初集信息特征的分析,来预测文档所含新知识的多少,这是一个非常值得研究的问题,同时也是构造领域本体的信息源应该具有的另一个特点——预测性.

为了深入研究这个问题,需要弄清楚什么是新增知识;文档中包含新增知识产生的决定因素是什么;以及这些因素与新增知识是否存在函数关系,如果存在函数关系如何表示等 3个问题.

1.3.1 新增知识的定义

定义 5 假设一个初集本体包括概念有 $\{x_1, x_2, \cdots, x_n\}$,关系有 $R_1: x_1 \rightarrow x_2, R_2: x_1 \rightarrow x_3, \cdots, R_{n-1}: x_1 \rightarrow x_n$. 对于终集本体有概念 $\{x_1, x_2, \cdots, x_n, x_{n+1}, \cdots, x_{n+i}\}$,有关系 $R_1: x_1 \rightarrow x_2, R_2: x_1 \rightarrow x_3, \cdots, R_{n-1}: x_1 \rightarrow x_n, R_n: x_1 \rightarrow x_{n+1}, \cdots, R_{n+i-1}: x_1 \rightarrow x_{n+i}$,则称 $x_{n+1}, \cdots, x_{n+i}$ 为新增概念,$R_{n+1}: x_1 \rightarrow x_n, \cdots, R_{n+i-1}: x_1 \rightarrow x_{n+i}$ 为新增关系,它们统称为新增知识.

1.3.2 文档中新增知识的决定因素 一个初集本体是由专家确定的,如果把本体理解成树型结构,那么初集本体的大部分概念处于树的上层节点,而新增知识是初集本体树的扩展,处在树的下层节点. 这样,在初集本体概念和新增概念之间必然存在一定的关系,称这种关系是新增知识对初集本体的依赖关系. 正是存在这种依赖关系,在一般情况下,一篇文档中包含初集本体的概念、关系越多新增知识数量也越多. 例如,有两篇与湿地分类有关的文档,相对于同一个初集本体,一篇含有 10个初集概念,包含了 6个新增概念;而另一篇含有 5个初集概念,只有 2个新增概念. 当然,也会出现包含初集概念多,新增概念少和包含初集概念少,新增概念多等特殊情况. 但对绝大多数文档还是符合包含初集本体的概念、关系越多新增知识数量也越多这一规律的. 作者在实际的实验中,对概念、关系这两个决定因素进行了细化,按照概念的位置、关系的距离以及子文档的划分原则确定了 12个因素,分别是包含初集本体概念数:词数 X、词数 Y、词数 Z;包含初集本体概念位置的数值:位置 P_X、位置 P_Y、位置 P_Z;包含初集本体关系距离的数值:距离 D_X、距离 D_Y、距离 D_Z、距离 X_Y、距离 X_Z、距离 Y_Z 等.

1.3.3 函数关系 文档中新增概念和上述 12个因素是否存在函数关系,如果存在是什么样的函数关系,将在第 2章中具体讨论包括概念性、关系性以及预测性的函数确定问题.

综合上述几方面分析,得到构造领域本体所需信息源的特点.

(1) 概念性:文档中具有概念数量、位置的信息.

(2) 关系性:文档包含关系的种类和数量.

(3) 预测性:对于待扩充的本体,一个文档最重要的是所能够提供的新知识的多少. 而不是包含现有知识的多少.

2 文档权值的确定

一个文档的重要级别的判定是由它的概念性、关系性以及预测性决定的. 这 3个特点权值的确定是需要解决的另一个问题. 作者通过查阅大量的相关资料,根据上述特点涉及的内容分别采用不同的方法确定其权值,其中对于概念性采用改进的 VSM方法;对于关系性采用本体关系距离的方法;对于预测性采用神经网络的方法. 然后,把这 3个权值作线性相加,最终得到文档的权值.

2.1 概念权值

通过对文档的分解,可以得出文档概念权值实际上就是计算概念子文档的权值. 文献 [13]通过大量的实验在近百种 VSM方法组合方式中,得出最好的方式为 BD-ACI-BCA. 其表述为式 (1)、(2),从传统需求的角度来看选择该公式是比较好的,但对于本文提出本体所需信息源却存在不足,分析如下:

$$w_{d,t} = (1 + \ln f_{d,t}) \tag{1}$$

$$W_d = (1 - s) + s \cdot \frac{W'_d}{av_{d \in D} W'_d} \tag{2}$$

式中:$w_{d,t}$ 为文档 d 中术语 t 的权值;$f_{d,t}$ 是术语 t 在文档 d 中的频率;W_d 为文档 d 的权值;s 是斜率 (通常为 0.7);$W'_d = \sqrt{\sum_{t \in T_d} w_{d,t}^2}$,$T_d$ 是出现在文档 d 中查询术语的集合,相当于出现在初集本体中的元素集;$av_{d \in D} W'_d$ 是文档集合中所有文档 W'_d 值的平均数. 从式 (1)、(2) 中不难看出它们只考虑词频一个因素,而忽略了概念在文档中的位置以及概念在初集本体中的位置这两个重要的因素,从而造成概念权值计算的误差. 为此,对式 (1) 中的参数作相应的改进. 针对第一个因素,按照定义 4确定相关系数为 $p_1 + p_2 + p_3 = 1$,这里取 p_1、p_2、p_3 系数分别为 0.5、0.3、0.2;另一个影响因素概念在初集本体中的位置,用符号 T表示,其取值范围为 0~1,按照线性划分,1到 10层对应的系数为 1到 0.1,每层相差 0.1,大于 10层按 0.05计算. 于是得到式 (1) 的改进公式:

$$w_{d,t} = (1 + \ln(p_1 f_{1d,t} + p_2 f_{2d,t} + p_3 f_{3d,t})) \tag{3}$$

式中:$p_1 f_{1d,t}$ 表示文档 d 中概念 t 在标题出现的频率;$p_2 f_{2d,t}$、$p_3 f_{3d,t}$ 分别为在句首、一般正文中出现的频率. 把式 (3)代入式 (2)就得到文档 d 的概念权值. 类似地,把上述过程应用到本文提出的概

念子文档 $c(X)$、$c(Y)$、$c(Z)$ 中,它们的和为文档的概念权值,如式 (4) 所示.

$$W_d^c = W_{c(X)} + W_{c(Y)} + W_{c(Z)} \quad (4)$$

式中: W_d^c 为文档 d 的概念权值; $W_{c(X)}$、$W_{c(Y)}$、$W_{c(Z)}$ 分别为概念子文档 $c(X)$、$c(Y)$、$c(Z)$ 的概念权值.

为了克服传统的 VSM 方法带来的不足,本文采用经典 VSM 方法和相关本体信息相结合的方法来确定权值,包括概念的数量、概念在文档中的位置、概念在初集本体中的位置等因素,大大提高了文档权值的精度.

2.2　关系权值

关系权值确定是一个比较新的研究课题. 文献 [10]采用关系遍历的方法来确定,涉及的参数多,缺乏普遍性. 而本文通过计算关系子文档权值来确定其权值,其主要思想是假设已知初集本体中的关系,包括直接关系和经过推理得出的间接关系,通过查找、计算一个文档中关系出现的位置和频率来计算关系权值. 这里只关心文档中两个元素同时出现的情况,不考虑关系的名称,也就是说只要两个元素同时出现就认为有关系. 影响权值的因素有两个: 一个是两个相关元素同时出现的位置不同其权值也不同,按照定义 3,元素的距离分为 d_1、d_2、d_3,这里它们分别取 0.6 0.3 0.1,且满足 $d_1 + d_2 + d_3 = 1$;另一个影响因素是关系元素在本体中的位置,以及它们的层次差. 类似概念权值的计算方法,采用对数方法计算一种关系在文档 d 中的权值如式 (5),计算关系子文档的关系权值如式 (6).

$$w_{d,r} = 1 + U\ln(d_1 f_{1d,r} + d_2 f_{2d,r} + d_3 f_{3d,r}) \quad (5)$$

$$W_r = \sqrt{\sum_{r \in R} w_{d,r}^2} \quad (6)$$

式 (5) 中 $w_{d,r}$ 为关系 r 在文档 d 中的权值,U是关系元素在本体中的位置系数,它的确定是该公式的难点,实际上是相关度问题的研究. 有关这方面的研究有不同的方法,由于本文是以领域本体为研究背景,文献 [14]给出的定义比较适合,其定义如下.

定义 6　概念 A、B 之间的相关度记为 $Rel(A,B)$,其具体计算方法如下:

如果 A,B 两个概念直接相关,则 $Rel(A,B) = 1$;

如果 A,B 两个概念通过 n 个概念间接相关,则 $Rel(A,B) = \frac{1}{n+1}$;

如果 A,B 两个概念直接继承相关,则 $Rel(A,B) = 0.75$;

如果 A,B 两个概念间接继承相关,则 $Rel(A,B) = \frac{3}{4(n+1)}$.

另外,$d_1 f_{1d,r}$ 表示关系 r 在文档 d 中出现在同一句中的频率,$d_2 f_{2d,r}$、$d_3 f_{3d,r}$ 分别表示在同一段、同一篇文章中出现的频率. 式 (6) 中 W_r 表示文档 d 的关系权值,R 为初集本体中包含的直接关系和经过推论得到的间接关系.

由于一个文档可以分解为 $r(X)$、$r(Y)$、$r(Z)$、$r(X,Y)$、$r(X,Z)$、$r(Y,Z)$ 等关系子文档,可得出一个文档的关系权值如式 (7) 所示:

$$W_d^r = W_{r(X)} + W_{r(Y)} + W_{r(Z)} + W_{r(X,Y)} + W_{r(X,Z)} + W_{r(Y,Z)} \quad (7)$$

式中: W_d^r 为文档 d 的关系权值; $W_{r(X)}$、$W_{r(Y)}$、$W_{r(Z)}$、$W_{r(X,Y)}$、$W_{r(X,Z)}$、$W_{r(Y,Z)}$ 分别为关系子文档 $r(X)$、$r(Y)$、$r(Z)$、$r(X,Y)$、$r(X,Z)$、$r(Y,Z)$ 的关系权值.

2.3　预测权值

根据第 1章的分析,得出领域本体信息源的评测由 3个特点决定,其中预测性计算是本文的一个重点,同时也是一个难点. 首先给出其基本思想: 参照初集本体的概念、关系,对给定文档进行分析,提取相关特征,找到这些特征与预测权值的函数关系. 从目前的分析中,可以得出这种关系并不是精确的,很难用传统的线性函数表达. 而在软计算方法中神经网络方法对该类问题有较好的处理能力.

目前人们已经提出了许多种神经网络模型,这些网络模型都是针对某种特殊用途的,各有优缺点. BP(back propagation)网络是目前为止最有影响的一种,它是在感知器中加入隐含层并且使用广义W算法进行学习之后发展起来的,表现为多层网络结构,相邻层之间为单向完全连接. 由于 BP网络对输入输出节点的数量没有限制,很多问题可以转化为用 BP网络能够解决的问题,如模式识别、信号检测、自适应滤波、函数逼近以及逻辑映射等. 可以利用 BP网络的这一特点来逼近原始数据与期望数值之间存在的关系,即通过学习和训练来拟合文档特征与文档蕴涵知识量的内在关系. 具体步骤如下.

(1) 输入特征值确定: 包括词数 X、词数 Y、词数 Z、位置 P_X、位置 P_Y、位置 P_Z、距离 D_X、距离 D_Y、距离 D_Z、距离 X_Y、距离 X_Z、距离 Y_Z 等 12个特征.

(2) 神经网络模型: BP算法.

(3) 输出值确定: 由文档的新增概念和新增关系的数量决定.

文档预测权值 W_d^n 可以通过输入文档特征值在 BP算法中得到. 针对 BP神经网络存在的收敛速度慢问题,本文使用自适应学习和附加动量项的方法;为了避免局部最小值的影响,采用多次随机初始化网络权值的方法;对于隐层节点的确定问题,本文采用 5倍率的交叉检验方法. 上述方法的采用,避免了 BP算法存在的缺点,达到了较好的结果,其实验过程将在第 3章中给予介绍.

2.4 文档级别的判定

从上述分析可以得出一个文档的级别判定由 3个因素决定,分别是概念权值、关系权值和预测权值. 文档级别可通过两种方法得出: 一种直接根据线性关系得出具体的文档权值,得到排列顺序,如式 (8) 所示;一种根据具体的要求确定文档的级别为重要、一般、不重要等 3个级别,采用模糊集方法判定级别. 本文采用第一种方法.

$$W_d^{crn} = aW_d^c + bW_d^r + cW_d^n \tag{8}$$

式中: W_d^{crn} 为文档 d 的权值; W_d^c、W_d^r、W_d^n 分别为概念权值、关系权值和预测权值; a、b、c 为常数,并且 $a + b + c = 1$.

以上给出了适用于构造领域本体所需信息源的选取方法,当然该方法还有不完善的地方,比如: 线性关系系数确定的科学性验证,以及文档级别采用模糊集方法时隶属函数的确定,还需进行深入的研究. 和国外同类研究相比,预测权值的考虑具有一定的借鉴意义.

3 实验结果与分析

3.1 预测权值模型的产生

为了验证 BP算法是否可以反映输入特征与预测值之间存在的隐含关系,本文使用 Java语言,后台使用 Oracle数据库编写了软件 OnMaker,它包括了如下功能:

(1) 初集本体的设定;

(2) 模拟文档的产生;

(3) 文档特征的产生;

(4) 训练样本的产生;

(5) 文档级别的确定;

(6) 神经网络模型的确定,学习训练;

(7) 真实文档测试,包括初集本体确定、导入文档、文档特征产生、通过模型得到结果文档权值.

从大量相关“湿地保护”文档得出文档中集合 X、集合 Y、集合 Z 的元素分布规律呈正态分布. 按照正态分布规律由软件 OnMaker模拟产生了 110组样本数据 (如表 1所示),利用 100组作为训练样本,10组作为泛化样本. 神经网络在训练过程中,训练误差没有明显下降时停止网络参数学习,多次随机试验表明网络学习迭代 500次后就基本达到稳定,因此选择训练次数为 500次. 为保证收敛速度,采用自适应学习率的方法,在学习过程动态地调节其大小,以加快其学习速度. 另外,为克服 BP网络的局部权值,本文采用多次随机训练的方法来避免局部最小的影响. 该 BP网络输入节点 12个,对应着词数 X、词数 Y、词数 Z、位置 P_X、位置 P_Y、位置 P_Z、距离 D_X、距离 D_Y、距离 D_Z、距离 X_Y、距离 X_Z、距离 Y_Z. 输出节点 1个,对应着预测权值. 训练结果如图 2所示: 训练误差为 0.012 3,泛化误差为 0.013 9. 从结果中发现,通过该软件得到的文档初集本体特征与文档预测值之间存在着隐含关系,文档的预测权值可以经过 BP网络计算获得.

表 1 由软件 OnMaker产生的 110组模拟数据

Tab. 1 110 group of simulated data by software OnMaker

文档	X	Y	Z	P_X	P_Y	P_Z	D_X	D_Y	D_Z	X_Y	X_Z	Y_Z	输出
1	1.58	2.28	1.65	0.13	0.17	0.12	1.14	1.65	1.14	7.83	6.21	7.90	5.00
2	2.98	2.67	6.51	0.20	0.18	0.52	1.02	1.49	0.93	6.48	5.83	6.82	4.70
3	2.81	0.54	0.54	0.16	0.06	0.08	1.91	2.39	1.79	11.95	9.74	11.95	3.30
4	0.52	2.69	0.85	0.06	0.23	0.18	0.70	1.02	0.62	3.87	3.27	4.33	5.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
110	2.37	8.77	0.53	0.24	0.55	0.10	0.94	1.17	0.88	6.61	4.85	6.05	3.70

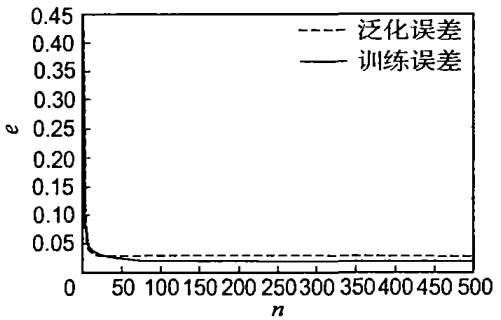

图 2　由 BP算法得到的训练与泛化误差

Fig. 2　The error of training and generalization by BP

3.2 “湿地保护”真实数据测试

OnMaker软件提供了用真实数据测试的功能,利用上述训练好的 BP网络,输入与“湿地保护”领域相关的 40个真实文档数据,产生了 40个文档的预测权值(见表 2). OnMaker同时也计算出了文档的概念权值、关系权值以及最终的文档权值,分别对应表 2的第 2 3 5列. 为了验证 OnMaker软件计算文档权值的正确性,由领域专家对这 40个文档进行判定,主要根据文档包含初级本体知识数量和专家认定增加知识的数量确定,其结果如表 2第 6列. 通过本方法与专家的判定结果相比较可以看出其对文档的排列顺序基本一致,并且它们之间存在一定的拟合关系. 由上述分析可以得出,该方法对于计算机自动评测构造本体所需信息源的重要程度是有效的.

表 2　OnMaker产生的文档值与专家判定的比较

Tab. 2　Comparison between expert weights and documents weights generated by OnMaker

文档	概念权值	关系权值	预测权值	文档权值	专家判定
1	0.230	0.801	0.018	0.377	0.327
2	0.207	0.852	0.132	0.370	0.332
3	0.626	2.724	0.084	1.103	1.089
4	0.295	1.106	0.074	0.485	0.479
5	0.268	1.240	0.086	0.558	0.498
6	0.173	1.281	0.088	0.534	0.500
7	0.460	2.298	0.125	0.924	0.864
8	0.261	0.707	0.482	0.331	0.325
9	0.275	0.398	0.169	0.258	0.215
⋮	⋮	⋮	⋮	⋮	⋮
40	0.972	3.185	0.129	1.504	1.351

图 3是使用 OnMaker软件得到的最终“湿地本体”. 其过程如下: 首先,领域专家通过 OnMaker输入初集本体,如图 3窗口左部,其中“污染”这个概念有 3个子概念. 然后,输入湿地领域文档集,并由 OnMaker软件对文档集自动完成排序,领域专家对级别较高的文档重点阅读,从中提取相关概念和关系. 最后,把提取的概念和关系输入到 OnMaker软件中得到图 3窗口右部最终“湿地本体”,其中“污染”这个概念新增了 6个概念. 通过该方法,由计算机完成信息源选择,大大提高了构造效率.

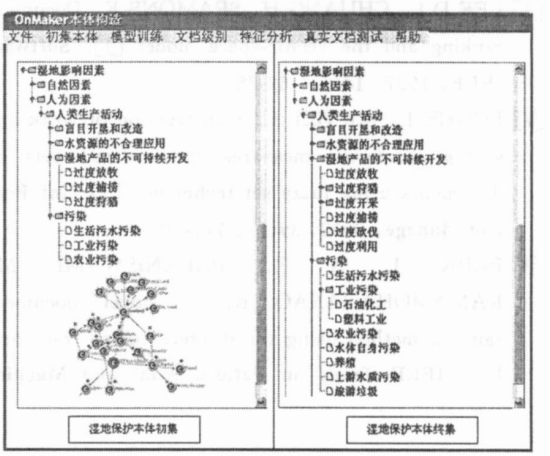

图 3　OnMaker湿地保护领域初集本体与最终本体

Fig. 3　Initial ontology and finished ontology in the OnMaker of wetland protection

4 结　语

构造领域本体所需信息源的研究是一个较新的课题. 本文通过对其特点的分析,提出了一个领域本体所需信息源判定的方法. 实验证明该方法是有效的,该项技术的研究可以应用到下一代 WWW网搜索引擎中,同时也是领域本体自动构造的关键技术. 此外,本文还为该方法专门设计了一个软件 OnMaker,下一步工作将在该方法的基础上重点研究如何把领域专家对知识的手工提取转换为机器(半)自动提取.

参考文献:

[1] ZHANG K, HU Y F, WANG Y. Multiple viewpoints based ontology integration [J]. **Lect Notes Comput Sci**, 2004, **3032**: 690-693

[2] BERNERS-LEE T, HENDLER J, LASSILA O. The semantic web [J]. **Sci Amer**, 2001, **284**(5): 34-43

[3] SHETH A, RAMAKRISHNAN C. Semantic (Web)

technology in action ontology driven information systems for search, integration and analysis [J]. **IEEE Data Eng Bulletin**, 2003, **26**(4): 40–48

[4] FERNÁNDEZ–LOPEZ M, GOMEZ–PEREZ A. Overview and analysis of methodologies for building ontologies [J]. **Knowledge Eng Rev**, 2002, **17**(2): 129–156

[5] DING Y, FOO S. Ontology research and development: part 1 — a review of ontology generation [J]. **Inf Sci**, 2002, **28**(2): 234–260

[6] LEE D L, CHUANG H, SEAMONS K. Document ranking and the vector–space model [J]. **Software IEEE**, 1997, **14**(2): 67–75

[7] EGGHE L, MICHEL C. Construction of weak and strong similarity measures for ordered sets of documents using fuzzy set techniques [J]. **Inf Proc and Manage**, 2003, **39**(5): 771–807

[8] PARK L A F, PALANISWAMI M, RAMAMOHANARAO K. A novel document ranking method using the discrete cosine transform [J]. **IEEE Trans on Pattern Anal and Machine Intell**, 2005, **27**(1): 130–135

[9] DANILOWICZ C, BALINSKI J. Document ranking based upon Markov chains [J]. **Inf Proc and Manage**, 2001, **37**(4): 623–637

[10] ALEMAN–MEZA B, HALASCHEK C, ARPINAR I B, *et al*. Ranking complex relationships on the semantic Web [J]. **Internet Comput IEEE**, 2005, **9**(3): 37–44

[11] 罗三定, 冯元勇, 沈德耀, 等. 基于概念的文档评价模型 [J]. 计算机工程, 2002, **28**(8): 79–81

[12] GOMEZ–PEREZ A, CORCHO O. A roadmap to ontology specification languages[C] // **Knowledge Engineering and Knowledge Management. Methods, Models, and Tools: 12th International Conference**. Berlin: Springer, 2000

[13] ANH N V, KRETSER D O, MOFFAT A. Vector–space ranking with effective early termination[C] // **Proceedings of ACM SIGIR Forum**. New York: ACM Press, 2001: 35–42

[14] 朱礼军. 万维网环境下基于领域知识的信息资源管理模式研究 [D]. 北京: 中国农业大学, 2004

Research on method of information sources selection for domain ontology building and its implementation

XING Jun[1,2], HAN Min*[1], ZHOU Kai-peng[1]

(1.School of Electr. and Inf. Eng., Dalian Univ. of Technol., Dalian 116024, China;
2.School of Inf. Sci. and Eng., Dalian Polytechnic Univ., Dalian 116034, China)

Abstract Method of information source selection is important to build domain ontology with regard to improving the ontology quality and efficiency, and develop the ontology. The classical methods only take concepts into account, and fall short of solving practical problems well. Therefore, an abstract method analysis is firstly used to analyze the characteristics of information sources, such as conceptuality, relativity and predictability, and then considering these properties, three methods— the improved vector space model (VSM), ontology relation distance and neural network, are introduced to calculate these characteristics weights, respectively. Finally, the simulated data is generated by implementing software OnMaker, and three weights of concept, relation and prediction are obtained, and in the following every document weight is calculated. Combined with a real document data set of "Wetland Protection", the model is tested and a good order effect on the document selection is attained.

Key words domain ontology; information sources; ontology building