一种基于义素的网页信息项语义匹配方法研究

卢正鼎　张茂元

（华中科技大学计算机科学与技术学院　武汉430074）

摘　要　本文提出了一种改进的基于语义的义素相似度，并从理论上分析参数β值的影响效果。在这个基础上，提出一种基于义素的词相似度，从语义上去匹配新名词和旧名词。在基于义素的词相似度基础上，提出一种网页信息项的语义匹配方法，来识别网页信息项的类别。实验结果表明，基于义素相似度的网页信息项语义匹配方法具有较好的匹配效果。

关键词　义素，相似，语义，匹配

A Sememe Based Semantic Matching of Web Information Item

LU Zheng-Ding　ZHANG Mao-Yuan

(Department of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074)

Abstract　Because some new words will be created in the development of human knowledge and the expression forms of words may be various, the word matching method is needed to research in order to adapt the changes from the point of the semantic features of words. This paper proposes an improved sememe similarity, which is based on semantic features of words. To analyze the effect of the coefficient β, this paper deducts three theorems in theory and then educes that the value of the coefficient β must be set in a range, which is to say that the coefficient β cannot be set much little value or much large value. Based on the sememe similarity, a word similarity is put forward to match new words with old words from the view of the semantic features. To adapt the Web page information, which includes the human knowledge and has various expression forms, a novel semantic matching method is proposed to identify the class of the Web information items. The semantic matching method is based on both the word similarity and the sememe similarity. As the experiment results show, the sememe based semantic matching method gains higher accuracy to identify the class of the Web information items.

Keywords　Sememe, Similarity, Semantic, Matching

1　引言

Internet 拥有海量的网页信息，为准确地获取有用信息，网页信息获取方法得到广泛的研究。词匹配是自然语言处理和信息获取中重要的一个难题，已经被应用到人工智能等许多领域，如语音识别[1]。

词匹配的方法可分为基于词频的方法和基于词分类关系的方法，前者用词在文档中出现的共同程度来体现词间的相似度；后者是建立在词汇语义网络中的分类关系层次上，一种基于直接相关的词匹配方法[2]，用词在各个文档中出现的频度来组成词向量。这种方法体现词在出现分布上的语义关系，但未考虑词分类学中的结构关系。一种基于本体论的词相似法[3]和一种基于语义的模糊匹配方法[4]，建立在词汇语义网络中层次关系间的距离因素，但未考虑层次关系中的深度因素。基于不同本体论的词相似法[5]考虑了层次关系中的深度因素，并获得了较好的效果，但面对新出现的词，词汇语义网络就需要扩充。

网页信息内容经常发生变化，且新名词会随着人类知识的发展而出现，因此词匹配方法就得具有较好的自适应性来匹配新名词。本文提出了一种改进的义素相似度，并从理论上分析参数β值的影响效果。在这个基础上，提出一种基于义素的词相似度来适应新名词的匹配。同时，本文在基于义素的词相似度基础上，提出一种网页信息项的语义匹配方法，来识别网页信息项的类别。实验结果表明，这种方法具有较好的匹配结果。

2　基于语义的义素相似度

2.1　义素网络

《知网》(HowNet)是一个网状的有机知识系统，以汉语和英语的词语所代表的概念为描述对象，来表示概念与概念以及概念属性之间的关系。在知网中，“义素”是从所有汉语词汇中提炼出的可以用来描述其它词汇的不可再分的基本语义元素，每一个概念是通过一组义素来表示的。

义素关系有上下位关系、同义关系、反义关系等复杂的关系，但上下位关系是最主要的义素关系。如图1，根据义素的上下位关系，所有“义素”组成了一个树状层次体系结构。

从药品信息的主要关键词（如药品名称、药名、药品商用名、功能主治），提取出每个义素，组成药品信息义素集合。然后按照 HowNet 的构建原理，对义素集合构建药品信息的语义网络 HowNet-medicine。

卢正鼎　教授，博士生导师，主要从事信息系统、软件工程和数据库等方面研究；张茂元　博士生，主要从事人工智能和网络信息系统等方面研究。

```
- entity|实 体
  ├ thing|万 物
  …  ├ physical|物 质
     …  ├ animate|生 物
        …  ├ AnimalHuman|动 物
           …  ├ human|人
              |  └ humanized|拟 人
              └ animal|兽
                 ├ beast|走 兽
           …
```

图1 HowNet 义素的树状层次结构

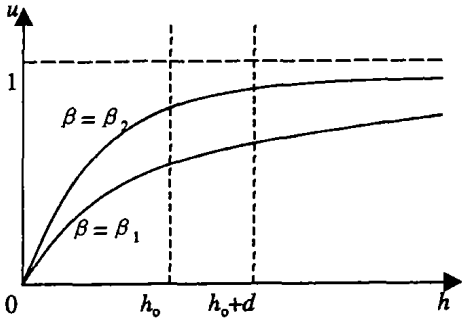

图2 函数 $u(h)$的曲线图

2.2 义素相似度函数

2.2.1 *基于语义路径的相似度* 语义网络的构建应用了词汇分类法，义素间的语义距离可以用义素间连接边的数量来表示，所以义素相似度可以用语义路径长度来计算。

定义1 设两个义素 $seme_1$和 $seme_2$之间的路径长度为 L，基于语义路径的相似度为：

$$Sim_1(seme_1,seme_2)=f_1(l)=e^{-\alpha l} \tag{1}$$

其中 $\alpha>0$是常数，$l\in[0+\infty)$。

2.2.2 *改进的语义相似度* 尽管基于语义路径的相似度在一些问题上取得了较好的结果，但在大型或通用的语义网络应用中，这种计算方法在准确度上存在一定的误差。为了改进这个不足，相似度计算还需引入更多的语义网络结构信息。从直观角度上，位于语义网络中较高层次的义素含有较通用的语义和较弱的相似度，而位于语义网络中较低层次的义素含有较具体的语义和较强的相似度，所以，在计算相似度时，义素的层次深度应当得到考虑。

定义2 $f_2(h_1,h_2)=\dfrac{e^{\beta(h_1+h_2)/2}-e^{-\beta(h_1+h_2)/2}}{e^{\beta(h_1+h_2)/2}+e^{-\beta(h_1+h_2)/2}}$，其中 $\beta>0$是常数，$h_1,h_2\in[0,+\infty)$。

定义3 设两个义素 $seme_1$和 $seme_2$之间的路径长度为 L，且它们的层次深度分别为 h_1和 h_2，则改进的义素相似度为：

$$\begin{aligned}Sim_2(seme_1-seme_2)&=f(f_1(l),f_2(h_1,h_2))\\&=f_1(l)\times f_2(h_1,h_2)\end{aligned} \tag{2}$$

2.3 相似度函数的相关定理

定理1 设函数 $u(h)=\dfrac{e^{\beta h}-e^{-\beta h}}{e^{\beta h}+e^{-\beta h}}$，其中 $\beta>0$是常数，$h\in[0,+\infty)$，则函数 $u(h)$的值域区间是[0,1]。设 a 为常数，若 $u(h_0)\geqslant a$，则 $\beta\geqslant\dfrac{1}{2h_0}\ln\dfrac{1+a}{1-a}$。

证明：

$$\begin{aligned}du/dh&=(1/(e^{\beta h}+e^{-\beta h})^2)][\beta(e^{\beta h}+e^{-\beta h})^2-\beta(e^{\beta h}-\\&\quad e^{-\beta h})^2]=4\beta/(e^{\beta h}+e^{-\beta h})^2\end{aligned} \tag{3}$$

由 $\beta>0$，可得 $du/dh>0$，因此函数 $u(h)$是严格单调递增的。当 $h=0$时，$u(h)$有最小值 $u(0)=0$；当 $h\to+\infty$时，$u(h)$有最大值1。所以函数 $u(h)$的值域区间是[0,1]。

$$\begin{aligned}d^2u/dh^2&=(-8\beta/(e^{\beta h}+e^{-\beta h})^3)(\beta e^{\beta h}-\beta e^{-\beta h})\\&=-8\beta^2(e^{\beta h}-e^{-\beta h})/(e^{\beta h}+e^{-\beta h})^3\end{aligned} \tag{4}$$

由 $\beta h\geqslant0$，可得 $e^{\beta h}-e^{-\beta h}\leqslant0$。因此 $d^2u/dh^2\leqslant0$，函数 $u(h)$的导数是单调递减的(如图2所示)。

由 $u(h_0)\geqslant a$，可得$(e^{\beta h}-e^{-\beta h})/(e^{\beta h}+e^{-\beta h})\geqslant a$，即$(e^{2\beta h}-1)/(e^{2\beta h}+1)\geqslant a$。并由此可推得 $e^{2\beta h}\geqslant(1+a)/(1-a)$，即 $\beta\geqslant\dfrac{1}{2h_0}\ln\dfrac{1+a}{1-a}$，证明完毕。

定理2 设 $w(h)=4h/(e^{\beta h}+e^{-\beta h})^2$，存在满足 $b=c/(2\beta)$和1.54<c<1.55的自然数 b、c，使得 $w(h)$在区间[0,b]单调递增，在区间(b,+∞)单调递减。

证明：

$$\begin{aligned}dw/dh&=(4/(e^{\beta h}+e^{-\beta h})^4)[(e^{\beta h}+e^{-\beta h})^2-2\beta h(e^{\beta h}+e^{-\beta h})\\&\quad(e^{\beta h}-e^{-\beta h})]\\&=(4/(e^{\beta h}+e^{-\beta h})^3)[(e^{\beta h}+e^{-\beta h})-2\beta h(e^{\beta h}-e^{-\beta h})]\\&=(4e^{\beta h}/(e^{\beta h}+e^{-\beta h})^3)[(1+e^{-2\beta h})-2\beta h(1-\\&\quad e^{-2\beta h})]\\&=(4e^{\beta h}/(e^{\beta h}+e^{-\beta h})^3)[(1+2\beta h)e^{-2\beta h}+1-2\beta h]\end{aligned} \tag{5}$$

令 $g(x)=(1+x)e^{-x}+1-x$，其中 $x\geqslant0$。

$dg/dx=e^{-x}-(1+x)e^{-x}-1=-1-xe^{-x}<0$

所以 $g(x)$严格单调递减。当 $x=0$时，$g(x)$的最大值是2；当 x 趋近于正无穷大时，$g(x)$的最小值是负无穷大。所以 $g(x)=0$有唯一解，设该零值解是 c。因此当$0\leqslant x<c$ 时，$g(x)>0$；当 $x>c$，$g(x)<0$。由于 $g(1.54)=0.5445-0.54=0.0045$，$g(1.55)=0.5412-0.55=-0.0088$，因此$1.54<c<1.55$。

因为 $dw/dh=(4e^{\beta h}(e^{\beta h}+e^{-\beta h})^3)\times g(2\beta h)$，所以当$0\leqslant 2\beta h<c$ 时，$g(2\beta h)>0$，可得 $dw/dh>0$；当$2\beta h>c$ 时，$g(2\beta h)<0$，可得 $dw/dh<0$；

依题意，$b=c/(2\beta)$，可得当$0\leqslant h<b$ 时，$w(h)$单调递增；当 $h>b$ 时，$w(h)$单调递减。命题得证。

定理3 设二元函数 $v(\beta,h)=u(h+d)-u(h)=\dfrac{e^{\beta(h+d)}-e^{-\beta(h+d)}}{e^{\beta(h+d)}+e^{-\beta(h+d)}}-\dfrac{e^{\beta h}-e^{-\beta h}}{e^{\beta h}+e^{-\beta h}}$，其中 $d\geqslant0$为常数。当 $h=h_0$且 $2\beta h_0>c$，则函数 $v(\beta,h_0)$对 β 是单调递减的，其中常数 $c>1.55$。

证明：令 $w(h)=4h/(e^{\beta h}+e^{-\beta h})^2$，则

$$\begin{aligned}\partial v/\partial\beta&=[4(h+d)/(e^{\beta(h+d)}+e^{-\beta(h+d)})^2]-[4h/(e^{\beta h}+\\&\quad e^{-\beta h})^2]=w(h+d)-w(h)\end{aligned}$$

依题意，$2\beta h_0>c$，则 $h_0>b=c/2\beta$。

由定理2可得，当 $h\geqslant h_0$ 时，$w(h+d)\leqslant w(h)$。

所以 $\partial v/\partial\beta=w(h+d)-w(h)\leqslant0$，当 $h=h_0$ 时，$v(\beta,h_0)$对 β 单调递减，命题得证。

推论1 设 $seme_1$和 $seme_2$是两个义素，相似函数 $Sim_2(seme_1,seme_2)=f(f_1(L),f_2(h_1,h_2))$的值区间为[0,1]，并且相似函数 Sim_2是非线性的。

证明：令 $h=(h_1+h_2)/2$，可得

$$f_2(h1,h2)=\frac{e^{\beta(h_1+h_2)/2}-e^{-\beta(h_1+h_2)/2}}{e^{\beta(h_1+h_2)/2}+e^{-\beta(h_1+h_2)/2}}=u(h)$$

由定理1可知，$0\leqslant u(h)\leqslant1$，所以$0\leqslant f_2(h_1,h_2)\leqslant1$。

又$0 \leqslant f_1(l)=e^{-\alpha l} \leqslant 1$，

所以$0 \leqslant f(f_1(L), f_2(h_1, h_2))= f_1(L) * f_2(h_1, h_2) \leqslant 1$，即相似函数$Sim_2(seme_1, seme_2)$的值域为[0,1]。

相似函数$Sim_2(seme_1, seme_2)=f(f_1(L), f_2(h_1, h_2))$的参数$L$、$h_1$和$h_2$的定义域为$[0,+\infty]$。假设相似函数是线性的，当$L \to +\infty$时，$Sim_2(seme_1, seme_2) \to +\infty$，这与$Sim_2(sem_1, seme_2) \in [0,1]$相矛盾。所以相似函数是非线性的，证明完毕。

2.4 参数β的影响效果分析

设语义网络的最大层次深度是h_{max}，因此，当$h > h_{max}$时函数$u(h)$的值对相似度计算结果没有影响，但函数$u(h)$在$h \in [0, h_{max}]$的值域对其有影响。所以为获得较好的影响效果，函数$u(h)$在$h \in [0, h_{max}]$的值域区间应尽量大，尽量接近[0,1]，也就是指$u(h_{max})$值尽量大，尽量接近1。

若要使$u(h_{max}) > a$，则由定理1可知，参数β必须$\beta \geqslant \frac{1}{2h_{max}} \ln \frac{1+a}{1-a}$。例如$a=0.95$，$h_{max}=10$时，$\beta \geqslant 0.183$。

但大的β值不一定会获得较好的影响效果。如图2所示，设$\beta_1 \leqslant \beta_2$，$2h_0\beta_1 > c$，$c > 1.55$，$d=1$。当$\beta=\beta_1$时，函数$u(h)$从$h_0$到$h_0+1$的增量值等于$v(\beta_1, h_0)$；当$\beta=\beta_2$时，函数$u(h)$从$h_0$到$h_0+1$的增量值等于$v(\beta_2, h_0)$。因为$\beta_1 \leqslant \beta_2$，所以由定理3可知，$v(\beta_2, h_0) \leqslant v(\beta_1, h_0)$。这样，大的$\beta$值会使差值$u(h_0+1)-u(h_0)$缩小，在一定程度上降低层次深度$h_0+1$对相似度的影响和层次深度$h_0$对相似度的影响的差别。

综上，为了获得较好的层次深度对相似度的影响效果，参数β不仅要使$\beta \geqslant \frac{1}{2h_{max}} \ln \frac{1+a}{1-a}$，而且不能设置成太大的值。

3 基于语义的信息项匹配系统

3.1 基于义素的词相似度

定义4 设词w含有n个义素$seme_1, seme_2, \cdots seme_n$，则该词可以用义素向量$semeV=(seme_1, seme_2, \cdots, seme_n)$表示。

定义5 设义素$seme$和义素向量$semeV=(seme_1, seme_2, \cdots, seme_n)$，则义素与义素向量之间相似度是$Sim3(seme, semeV)=\max_{j=1}^{n} Sim2(seme, seme_j)$。

定义6 设词w_1的义素向量是$semeV_1=(seme_{11}, seme_{12}, \cdots, seme_{1m})$，词$w_2$的义素向量是$semeV_2=(seme_{21}, seme_{22}, \cdots, seme_{2n})$，则基于义素的词$w_1$与词$w_2$的相似度是：

$$Sim4(w1, w2)=\frac{1}{|semeV_1|}\sum_{i=1}^{m} Sim3(seme_i, semeV_2) \quad (6)$$

其中，义素向量$semeV_1$的模$|semeV_1|=m$。

人类知识发展中出现的新词，像旧词一样含有义素，因此新词可用义素向量来表示。这样，新词和旧词之间的相似度可用式(6)来计算，所以基于义素的词相似度不仅用于旧词，同样能适用于新词，因而具有较强的自适应性。

3.2 相关工作

如图3，网页信息挖掘系统包含网页搜索、网页分类、信息项提取和信息项匹配四个子系统。网页搜索子系统用搜索引擎搜索网页；网页分类子系统采用一种基于特征选取及模糊学习的网页分类方法[6]，过滤掉非相关的网页；信息提取子系统用HTML解析器，根据网页标记从网页中提取出信息项及其相应关键词w；分词子系统采用一种基于语境的分词方法[7]，提取出信息项的关键词w；信息项匹配子系统将信息项的关键词w与各个项类特征词T进行语义匹配，然后按匹配结果把信息项和匹配到的项类特征词一起存到数据库中。

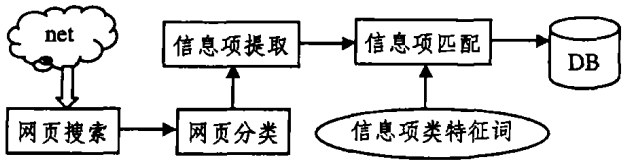

图3 网页信息挖掘系统

药品广告信息，有药品名称、功能主治、生厂企业等项目类。例如，一药品广告被信息项提取子系统，提取了三个信息项及相应关键词“商用药名”、“用法与用途”、“厂商”。信息项匹配子系统把对应“商用药名”的信息项匹配到“药品名称”项目类，把对应“用法与用途”的信息项匹配到“功能主治”项目类，把对应“厂商”的信息项匹配到“生厂企业”项目类。

3.3 信息项匹配系统结构

如图4所示，基于语义的信息匹配模块分为4层。第一层是义素转换层，用义素向量$SemeV_i$表示从各个信息项$Item_i$中提取的信息关键词w_i；第二层用式(6)计算义素向量$semeV_i$与信息项类特征词T_j的相似度m_{ij}；第三层采用竞争机制，比较第二层的结果，如果m_{ij}最大，则使$y_j=i$，但如果所有值都很小，则使$y_j=0$；第四层合并各分量，输出信息项匹配矢量。

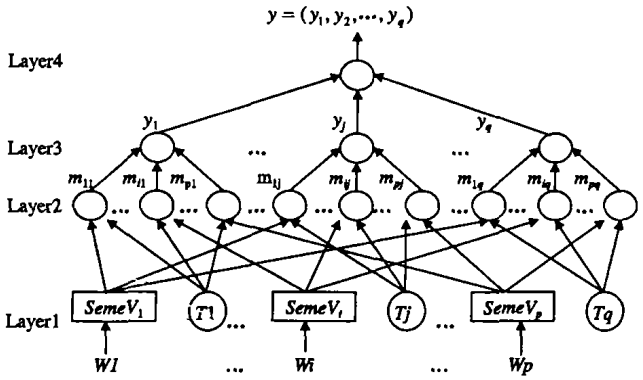

图4 基于语义的信息项匹配系统结构图

4 实验

从网上选取930张药品广告信息网页，网页解析器依据网页标记提取出各类信息项及其对应关键词。用基于词语义网络距离的模糊匹配方法和基于义素的信息项匹配方法($\alpha=0.2$，$\beta=0.6$)对这些信息项进行匹配，分别定义为方法A和B。得到的匹配结果如表1。

表1 两种方法的信息项匹配结果

项类	数量	方法A		方法B	
		正确数	准确率(%)	正确数	准确率(%)
药品名称	930	765	82.3	805	86.6
功能主治	930	757	81.4	798	85.8
生厂企业	620	513	82.7	538	86.8
用法用量	760	614	80.8	650	85.5
总计	3240	2649	81.8	2791	86.1

从平均准确率角度，方法B的平均准确率是86.1%，比方法A高4%。这表明作为语义基本单元的义素，在一定程度上可以提高语义匹配的准确率。

从最大准确率与最小准确率之差较小角度，方法A的最大准确率与最小准确率之差是1.9%，而方法B的差值是

(下转第54页)

的 GoTo 语句——用它可以做很多事情，但却使程序变得难于理解。

5 动态截断——动态地阻断回溯

传统的 Prolog 截断机制是静态的。静态截断机制存在两个问题：1) 当执行过程通过截断符号"!"时，截断即发生作用，但是它只影响那些已经把它放入其中的子句(在源文本中)。无法把一个截断的作用通过参数传递给另外一个谓词，如果条件满足，截断就只能被评估。2) 如果还没有截断谓词中下一个子句的回溯点，则先要截断子句中子目标更多的解是不可能的。

Visual Prolog 有一个动态截断机制，通过两个标准谓词 getbacktrack 和 cutbacktrack 来实现。这种机制使我们能够处理上述两个问题。谓词 getbacktrack 返回一个指针，指向当前回溯点堆栈的栈顶。在以后的某个时候，通过给出谓词 cutbacktrack 所找到的指针，就可以删除这个位置之上的所有回溯点，

如果出现回溯，尽管谓词可能返回许多答案，但是调用 cutbacktrack 来截断是可能的。随后的一个失败将回溯到上一个谓词。

动态截断更重要的应用是将回溯指针传递给另一个谓词并有条件地执行截断。指针是 unsigned 类型的，并且能够以 unsigned 类型的参数形式进行传递。

使用动态截断需要十分理智。利用动态截断非常容易破坏程序的结构，粗心的使用会产生许多难以解决的问题。

6 否定谓词 not

谓词 not 称为否定谓词，常用来对一个子目标的结果取反。考察下面的程序，它说明如何使用谓词 not 来识别一个好学生：积分点(GPA)不低于3.5且不是受处罚的试读期。

```
DOMAINS
    name = symbol
    gpa = real
PREDICATES
    honor_student(name)
    student(name, gpa)
    probation(name)
CLAUSES
    honor_student(Name):-
        student(Name, GPA),
        GPA>=3.5,
        not(probation(Name)). /* 使用谓词 not */
    student("Betty Blue", 3.5).
    student("David Smith", 2.0).
    student("John Johnson", 3.7).
    probation("Betty Blue").
    probation("David Smith").
GOAL
    honor_student(X).
```

使用谓词 not 时需要注意的一点是：当子目标不能被证明为真时，谓词 not 运行成功。这导致发生这样一种情形：未绑定变量在谓词 not 中被绑定。当带有一个自由变量的子目标被谓词 not 调用时，Visual Prolog 将返回错误信息：在 not 或 retrackall 中不允许有自由变量。产生这个信息的原因是，Visual Prolog 在一个子目标中绑定了自由变量，而这个子目标必须与其它一些子句合一且该子句必须成功。在含有谓词 not 的子目标中处理未绑定变量的正确方法是使用匿名变量。

结束语 回溯机制是逻辑程序设计的重要设施。回溯本身是一种获得目标所有可能解的良好方法。然而，回溯也有副作用，一是它可能导致 Visual Prolog 给出多余的答案，而 Visual Prolog 自己不能区分实质上相同的两个解，所以当寻找给定问题的唯一解时可能要搜索好几遍，因此会降低效率。二是尽管一个特殊的目标已被满足，但是回溯机制可能还会强迫 Visual Prolog 继续寻找另外的解。在这些情况下，必须控制回溯过程。Visual Prolog 的静态截断机制、失败谓词 fail 与否定谓词 not 等控制谓词，以及动态截断机制，构成了完整的目标搜索求解控制机制，可以实现对搜索过程的仔细控制。

本文在考察 Visual Prolog 回溯机制基本问题的基础上，通过实例，对搜索求解控制机制进行了详细分析，从而揭示回溯机制和搜索求解控制机制的本质特性和应用机理。

参 考 文 献

1 雷英杰，邢清华，孙金萍，张雷. Visual Prolog 智能集成开发环境评述[J]. 空军工程大学学报(自然科学版)，2002，3 (5)：39～43

2 雷英杰，张雷，邢清华，孙金萍. Visual Prolog 语言教程[M]. 西安：陕西科学技术出版社，2002

3 雷英杰，邢清华，孙金萍，张雷. Visual Prolog 编程、环境及接口[M]. 北京：国防工业出版社，2004

4 雷英杰，王涛，赵晔. Visual Prolog 的回溯机制分析[J]. 空军工程大学学报(自然科学版)，2004，5 (5)：80～84

(上接第51页)

1.3%。这表明语义网络的深度因数，在一定程度上影响匹配的效果。

综上，基于义素的信息项匹配方法不仅有较高的准确率，而且最大准确率与最小准确率之差较小，是一种较好的语义匹配方法。

结束语 基于义素的网页信息项的语义匹配方法，从义素这个基本语义单元角度，给出适用于变化网页信息的语义匹配方法。这个方法是目前较好的语义匹配方法中的一种，已经用于网上非法药品广告的检测系统。

参 考 文 献

1 Da L G, Facon J, Borges D L. Visual speech recognition: a solution from feature extraction to words classification. In: Proc. of Symposium on Computer Graphics and Image Processing, XVI Brazilian, 2003.399～405

2 Nakashima T. Classification of characteristic words of electronic newspaper based on the directed relation. In: 2001 IEEE Pacific Rim Conf. on Communications, Computers and signal Processing, 2001(2):591～594

3 Vladimir A O. Ontology based semantic similarity comparison of documents. In: 14th Intl. Workshop on Database and Expert Systems Applications (DEXA'03), 2003.735～738

4 程莉，卢正鼎，文坤梅. 基于语义的模糊匹配探索与应用. 华中科技大学学报(自然科学版)，2003，31(2)：23～25

5 Rodriguez M A, Egenhofer M J. Determining semantic similarity among entity classes from different ontologies, Knowledge and Data Engineering. IEEE Transa.on, 2003, 15(2): 442～456

6 张茂元，卢正鼎. 基于特征选取及模糊学习的网页分类方法研究. 小型微型计算机系统，已录用。