

# 语义 Web 中的概念等级匹配算法

尹东斐, 梁力

(西安交通大学电子与信息工程学院, 710049, 西安)

**摘要:** 针对语义 Web 匹配方法仅适于字符串、字典近义词匹配而导致精度低、效果差的问题, 提出了一种基于概念等级的语义 Web 匹配算法. 对于一个包含大量概念的本体, 算法可根据概念之间的联系建立起概念结构, 通过计算获得本体概念之间的相互支持度, 从而使得普通概念的支持度低而特殊概念的支持度高, 进而将概念的支持度量化为概念等级. 以计算得到的概念等级为权值, 将概念间的语言学匹配度加权, 由此计算出新的概念的匹配度. 该匹配度可将各个概念之间的内在联系关联起来, 从而提高语义匹配的精度. 实验结果表明, 所提算法在本体间的语义匹配精度比字典近义词匹配法提高了 20%.

**关键词:** 语义 Web; 概念等级; 概念结构; 语义匹配

**中图分类号:** TP311 **文献标志码:** A **文章编号:** 0253-987X(2009)04-0049-03

## A Matching Algorithm in Semantic Web

YIN Dongfei, LIANG Li

(School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

**Abstract:** A conception rank based algorithm for semantic Web is presented to improve the low efficiency and poor precision of the synonymous words matching algorithm and the string matching algorithm. For an ontology with large amount of conception, the conception structure of the ontology is constructed by the algorithm based on the relationship among conceptions. Then the algorithm calculates the contribution for each conception such that the general conception has low contribution and the special conception has high contribution, and then the contributions are quantified into conception ranks. The resulting conception ranks are used as weights and the weighted linguistically matching degree is used to calculate a new conception matching degree in semantic matching. The new matching degree can improve the accuracy of semantic matching because the structure coefficient is taken into account. Experiments show that the matching degree can be increased by 20% compared with the synonymous words matching algorithm.

**Keywords:** semantic Web; conception rank; conception structure; semantic matching

语义 Web<sup>[1]</sup>就是能够根据语义进行判断的网络, 它是对当前 Web 的扩展. 语义 Web 上的信息具有良好的定义, 使得计算机之间以及计算机与人类能够更好地彼此合作<sup>[2-4]</sup>.

当前较为流行的语义 Web 匹配算法有 2 种: 基于概念名称(字典近义词)的匹配算法和基于概念结构的匹配算法<sup>[5]</sup>. 前者主要从概念名称的语言学角度进行匹配, 后者主要从概念结构相似性方面进

行匹配. 基于字典近义词的匹配算法需要一个事先建立好的领域字典, 字典中包括对本体中的所有概念的解释, 这种匹配算法的精度完全取决于字典对词解释的精度, 而且对于本体中同义词和近义词难以区分. 基于概念结构的匹配算法主要偏重于对词法、语法的分析, 依靠词法、语法分析的结果进行语义匹配, 然而词法、语法分析的结果不易量化, 难以利用<sup>[6-8]</sup>.

本文提出一种概念等级匹配算法,可将概念结构很好地量化,从而结合基于概念名称的匹配的计算结果把字典语义和概念结构结合起来,提高了语义匹配的精度。

## 1 基于概念等级的语义 Web 匹配算法

本体形式的定义为一个 5 元组  $O = \{C, R, F, A, I\}^{[9]}$ , 其中  $O$  为本体集合,  $C$  是领域概念的集合(概念是客观世界任何事物的抽象描述),  $R$  代表了关系的集合(关系代表了在领域中概念之间的交互作用),  $F$  代表一类特殊的关系称为关系函数,  $A$  是本体公理的集合, 代表永真断言,  $I$  是实例的集合(实例也称作个体, 是本体元素的一个实现)。本文的研究对象主要为  $C$  和  $R$ 。

**定义 1** 在一个领域字典  $D$  中,  $\forall c \in C$  都有  $c \in D$ , 设函数  $N(c)$  为概念  $c$  的名称, 则基于概念名称的概念相似度为

$$S_n(c_i, c_j) = |N(c_i) \oplus N(c_j)| \quad (1)$$

式中:  $c_i, c_j$  分别为  $C_i, C_j$  中的概念,  $c_i \in C_i, c_j \in C_j$ , 其中  $C_i \in O_i, C_j \in O_j, O_i, O_j$  为 2 个不同的本体,  $C_i, C_j$  分别为  $O_i, O_j$  中的概念集合;  $\oplus$  表示匹配运算。

**定义 2** 假设  $T(c_i) = \{c | c \in C \wedge \forall c \in C, \exists [(c_i, c) \in R \vee (c, c_i) \in R], R \subseteq C \times C, C \in D\}$  表示和概念  $c_i$  相关的所有概念的集合, 则基于概念结构的概念相似度为

$$S_t(c_i, c_j) = |T(c_i) \oplus T(c_j)| \quad (2)$$

设  $R_i, R_j$  分别为  $O_i, O_j$  中的关系集合,  $R_i \in O_i, R_j \in O_j$ , 则  $R_i \subseteq C_i \times C_i, C_i \in O_i, R_j \subseteq C_j \times C_j, C_j \in O_j$ , 其中  $\times$  表示叉集运算。

**定义 3** 本体  $O$  中的概念和关系可以抽象成一个概念图  $G = (V, E)$ ,  $\forall c \in V, c \in C, C \in O, V$  为概念图的节点集合,  $\forall r \in E, r \in R, R \in O, E$  为概念图中边的集合, 则称有向图  $G = (V, E)$  为本体  $O$  的概念图。

**定义 4** 设  $g_{\deg}(c)$  为概念  $c$  在概念图中对应节点的出度<sup>[10]</sup>,  $d$  为阻尼系数, 其取值范围为  $(0, 1]$ , 则概念  $c_i$  的概念等级为

$$L(c_i) = (1 - d) + d \left( \sum_{t \in T(c_i)} \frac{L(t)}{g_{\deg}(t)} \right) \quad (3)$$

当得到本体上所有概念数量  $N$  时, 可以用下面的定义计算概念等级

$$L(c_i) = \frac{1 - d}{N} + d \left( \sum_{t \in T(c_i)} \frac{L(t)}{g_{\deg}(t)} \right) \quad (4)$$

**定理 1** 一个本体中总的概念等级不变性  $(|C| =$

$N)$ , 即

$$\sum_{i=1}^N L(c_i) = N \quad (5)$$

式中:  $c_i \in C, C \in O, C$  是本体  $O$  中的领域概念集合。

在求得本体中的概念等级以后, 对于 2 个本体  $O_i, O_j$  间概念  $c_i$  和  $c_j$  的相似度为

$$S_r(c_i, c_j) = a S_n(c_i, c_j) + b \sum_{t \in T(c_i) \cap T(c_j)} \frac{L(t)}{L(c_j)} S_n(t, c_j) + c \sum_{t \in T(c_j) \cap T(c_i)} \frac{L(t)}{L(c_i)} S_n(t, c_i) \quad (6)$$

从式(2)中可以看出, 概念  $c_i$  的概念等级取决于和  $c_i$  相关的所有概念的概念等级, 所以在计算过程中, 相同的初始值经过多次迭代以后, 可以收敛到一组稳定的值, 这些值就是概念等级。概念等级只和概念之间的结构有关, 而和概念的语义没有关系, 所以用概念等级描述概念之间的关系, 较之语法分析的结果更易于在计算中使用。当计算得出本体中所有概念的概念等级以后, 就可以通过式(6)来计算本体间概念的相似度。

本文提出的基于概念等级的语义 Web 匹配算法的描述如下:

输入  $C_i, R_i \in O_i, C_j, R_j \in O_j, c_i \in C_i, c_j \in C_j$

输出  $S_r(c_i, c_j)$

(1) 设有本体  $O_i, O_j$ , 其中分别有概念集合  $C_i, C_j$  和关系集合  $R_i, R_j$ , 并初始化所有概念的概念等级为 1;

(2) 选择参数  $d \in (0, 1]$ ;

(3) 对于  $C_i, C_j$ , 重复计算所有概念的概念等级;

(4) 从  $c_0 \in C$  开始, 计算  $L(c_0) = (1 - d) + d \left( \sum_{t \in T(c_0)} \frac{L(t)}{g_{\deg}(t)} \right)$ ;

(5)  $\forall c \in C$ , 计算  $L(c) = (1 - d) +$

$d \left( \sum_{t \in T(c)} \frac{L(t)}{g_{\deg}(t)} \right)$ ;

(6) 重复计算第(4)、第(5)步, 直到所有概念的概念等级不再变化为止;

(7) 用式(6)计算概念  $c_i, c_j$  的相似度。

在第(6)步的结束条件中, 需要所有概念的概念等级不再变化, 在实际应用中, 只需要当所有概念的概念等级的变化量小于某一阈值即可。

## 2 实验结果及分析

MWSAF 是美国芝加哥大学开发的一个基于

图形界面的语义 Web 工具,该工具使用的是基于概念结构的无权重的匹配算法,并且还提供了大量的本体和网络服务器描述语言(WSDL)实例. 由于网络服务和语义 Web 尚处于研究阶段,本文采用了 MWSAF 提供的数据和工具进行实验和验证算法,并将本文算法的结果与原算法的结果进行了比较. 本文选取与天气相关的概念,对其概念进行赋权值,然后与相关的本体网络语言(OWL)中的本体做匹配,将其结果与原算法的结果进行比较. 比较的内容主要包括 2 个方面:①本文算法能否提高表达同一语义的 2 个概念间的匹配度;②本文算法能否降低表达不同语义的 2 个概念间的匹配度.

本文利用 MWSAF 开发工具实现了改进后的语义匹配算法. 采用本文算法将 WSDL 中的 GlobalWeather(本体  $O_1$ ) 与 OWL 中的 Weather-Concepts(本体  $O_2$ ) 进行匹配,取  $d=3$ ,经过 50 次迭代以后的概念等级如表 1 所示.

表 1 本体和 WSDL 中概念的概念等级

| 编号    | 概念名称                | $L(c_1)$       | $L(c_2)$       |
|-------|---------------------|----------------|----------------|
| 1     | phenomenonType      | 1. 165 344 281 | 1. 002 177 742 |
| 2     | station             | 1. 619 731 201 | 1. 479 699 457 |
| 3     | temperature         | 0. 538 501 694 | 1. 198 845 786 |
| 4     | phenomenonIntensity | 0. 719 367 821 | 1. 405 057 124 |
| 5     | wind                | 0. 619 279 415 | 0. 766 035 064 |
| 6     | weatherReport       | 1. 578 446 978 | 1. 909 628 501 |
| 7     | visibility          | 0. 943 935 623 | 0. 975 734 34  |
| 8     | layer               | 0. 958 150 162 | 0. 718 294 776 |
| 9     | precipitation       | 0. 783 334 866 | 1. 005 335 325 |
| 10    | pressure            | 0. 323 181 828 | 1. 995 832 928 |
| 11    | extreme             | 0. 964 152 52  | 0. 582 459 761 |
| 12    | direction           | 1. 917 054 402 | 0. 460 394 718 |
| 概念等级和 |                     | 21. 063 970 98 | 20. 235 096 23 |

本文算法与字典近义词算法的匹配结果比较如图 1 所示.

从图 1 中可以看出,本文算法当数值越大时,匹配度越高,本体和 WSDL 的匹配度平均提高了 20% 左右. 由图 1 还可以看出,visibility 概念在图中有较高相似度,其原因在于visibility的5个属性全部可以匹配,而 direction 的匹配度较低,这是因为 direction 虽然在本体和 WSDL 都有相同的属性,但是其属性在本体和 WSDL 中的定义是不同的,所以在本体和 WSDL 中的 direction 是不同的 2 个概念,因而其相似度较低(字符相同但语义不同).

由于匹配度不再由概念本身决定,而是由与概

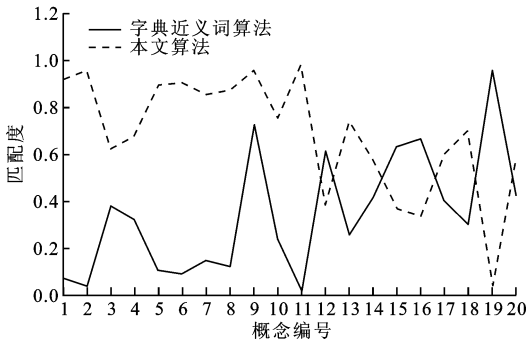


图 1 2 种算法的匹配结果比较

念相关的其他多个概念共同作用,因此本文算法提高了匹配精度.

3 结束语

本文主要讨论了语义 Web 中的概念匹配问题,研究分析了已有的语义 Web 匹配算法存在的问题,并在此基础上提出了一种基于概念等级的语义 Web 匹配算法. 本文算法可以从概念的结构关系上对概念进行分析匹配,结合基于概念名称的算法,使得匹配的精度得到了较大提高.

参考文献:

[1] BEMERS-LEE T, HENDLER J, LASSILA O. The semantic Web [J]. Scientific American, 2001, 284(5): 34-43.

[2] MADHAVAN J, BERNSTEIN P A, RAHM E. Generic schema matching with cupid [C]// Proceedings of 27th International Conference on Very Large Data Bases. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001: 49-58.

[3] 朱正东, 缪相林, 胡亚红, 等. 可用性语义 Web 服务的通用发现机制[J]. 西安交通大学学报, 2008, 42(6): 659-663.

ZHU Zhengdong, MIAO Xianglin, HU Yahong, et al. Usability based general semantic web services discovery mechanism [J]. Journal of Xi'an Jiaotong University, 2008, 42(6): 659-663.

[4] 邓志鸿, 唐世渭, 张铭, 等. Ontology 研究综述[J]. 北京大学学报(自然科学版), 2002, 38(5): 199-208.

DENG Zhihong, TANG Shiwei, ZHANG Ming, et al. Overview of Ontology [J]. Acta Scientiarum Naturalum Universitatis Pekinesis, 2002, 38(5): 199-208.

[5] PATIL A A, OUNDHAKAR S A, SHETH A P, et al. Meteor-s web service annotation framework [C]// Proceedings of the 13th International Conference on World Wide Web. New York, NY, USA: ACM, 2004: 553-562.

- [6] 许方芳,郑诚,戈明东. 语义网中的本体映射研究[J]. 计算机与现代化, 2008(2): 36-39.  
XU Fangfang, ZHENG Cheng, GE Mingdong. Research on ontology mapping on semantic Web [J]. Computer and Modernization, 2008(2): 36-39.
- [7] 张大陆,吕韬. 基于概念频率的 Web 服务语义标注[J]. 同济大学学报(自然科学版), 2008, 36(1): 103-107.  
ZHANG Dalu, LÜ Tao. Concept frequency-based Web service annotation [J]. Journal of Tongji University(Natural Science), 2008, 36(1): 103-107.
- [8] 王祥瑞,李力东. 语义网的产生与发展[J]. 吉林建筑工程学院学报, 2007, 24(3): 66-68.  
WANG Xiangrui, LI Lidong. The birth and development of semantic Web [J]. Journal of Jilin Institute Architectural and Civil, 2007, 24(3): 66-68.
- [9] 朱礼军,陶兰,黄赤. 语义万维网的概念、方法及应用[J]. 计算机工程与应用, 2004(3): 79-83.  
ZHU Lijun, TAO Lan, HUANG Chi. Semantic Web: concept, approach and application [J]. Computer Engineering and Applications, 2004(3): 79-83.
- [10] 陈健民,曾明,刘国荣. 离散的数据结构[M]. 西安:西安交通大学出版社, 2005: 170-175.
- (编辑 刘杨 苗凌)