一种基于概念信息量的相似度传播算法

徐德智 吴军庆 陈建二 赵于前
(中南大学信息科学与工程学院 长沙 410083)

摘 要 相似度传播在本体概念相似度计算中有着非常重要的作用。然而,目前常见的相似度传播算法大都采用了固定比例的相似度传播值,没有对相似度转播值进行合理的定量分析。针对此问题,提出了基于概念信息量的相似度传播算法,该算法根据匹配节点的概念信息量大小来判断其子父节点匹配概率大小,通过匹配概率大小调整相似度传播值,从而进行更精确的相似度传播。理论分析与实验结果证明了该算法是有效的。

关键词 本体,相似度传播,概念信息量

Algorithm of Similarity Propagation Based on Information Content of Concept

XU De-zhi WU Jun-qing CHEN Jian-er ZHAO Yu-qian
(College of Information Science and Engineering, Central South University, Changsha 410083, China)

Abstract Similarity propagation is very important for calculating similarity between two concepts. However, the existing algorithms of similarity propagation usually use a fixed proportion of spreading value, these algorithm do not take reasonable quantitative analysis for spreading value. To solve the problem, a novel algorithm of similarity propagation was proposed, which is based on information content of concept. The algorithm adopts the value of information content of matched node to determine matching probability of the matched node's children and parents, and more accurate propagated value will be obtained by adjusting spreading value according to the matching probability of node. Theoretical analysis and the results of experiment show that the algorithm is efficient.

Keywords Ontology, Similarity propagation, Information content

随着本体数量的不断增加,本体的重用和共享逐渐成为急待解决的重要问题,而本体映射则是解决该问题的关键。由于本体的异构性,本体映射是一个非常复杂的问题。目前,出现了一些基于概念相似度的本体映射系统,例如:ASMOV[1],falcon[2],Rimom[3]等。在概念相似度计算中,处于结构级的相似度传播起到了非常关键的作用。然而,在目前的相似度传播算法中,却没有对相似度传播值进行合理的定量分析,从而引起无法找到部分关键映射或者产生错误映射的问题。因此,针对此问题,本文提出基于概念信息量的相似度传播算法,以提高相似度传播值的精确度。

1 相关工作

目前,对于相似度传播算法的研究并不多。其中,具有代表性的传播算法有 Similarity Flooding[4] 和 GMO[5],它们的核心思想主要基于:如果两个概念的父类或者子类相似,那么这两个概念也可能相似,基于此特征把这两个概念的父类或子类的相似度通过相似度传播算法传播到两个待匹配概念中。它们之间主要区别在于 Similarity Flooding 相似度的传播只考虑已匹配的概念对邻居节点的传播,而 GMO 则是本体全局的相似度传播。然而,对于相似度传播值,两者都没有进行合理的定量分析,没有考虑不同概念间的相似度传播的差异,而只是给予一个固定比例的相似度传播值。

为了说明不同概念之间的相似度传播的差异性,本文给出下面的例子进行说明。例如图 1 是某源本体,本体的默认关系为 subclassof。假定通过语言级映射后,得出源本体中的 entity 和 horse 概念与目标本体中的 entity 和 horse 概念存在匹配映射关系,现进行结构级的相似度传播,相似度的传播是基于这样的认识:如果父类匹配,子类匹配的可能性越大,那么父类对子类的相似度传播越大。根据以上认识,讨论概念 entity 和 horse 对子类的相似度传播情况,对于目标本体来说,其 entity 的子类空间要比 horse 子类空间大得多,entity 的子类可以是图 1 本体中除了自身外的所有概念,而 horse 子类则受限于更小的空间内,所以对于entity的两个子类匹

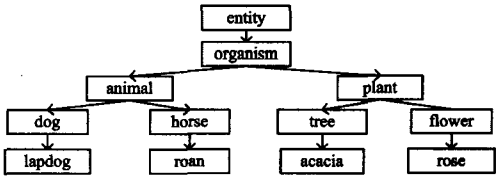

图 1 某源本体

到稿日期:2008-07-15 返修日期:2008-09-17 本文受国家自然科学基金重点项目(60433020),湖南省自然科学基金(06JJ50142),湖南省国土资源厅科技计划项目(200718)资助。

徐德智 教授,主要研究方向为 Web 计算、语义网等,E-mail:jqwu_csu@163.com;吴军庆 硕士研究生,主要研究方向为语义网;陈建二 教授,博士生导师,主要研究方向为计算机网络、计算机理论等;赵于前 副教授,主要研究方向为本体在信息工程中的应用。

配的可能性要远小于 horse 的两个子类匹配可能性，故概念 entity 对子类的相似度传播值要小于 horse 对其子类的相似度传播。根据以上分析可知，不同概念的相似度传播是有差异的。

2 基于概念信息量的相似度传播算法

针对传统方法的不足，本文提出了基于概念信息量的相似度传播算法。本节首先通过两个定理给出了匹配概念的信息量大小与其子父节点的匹配概率的关系，接着在两个定理基础之上定义了语义传播因子，然后结合语义传播因子制定了相似度传播算法，并给出了收敛性证明，最后给出了该传播算法的主要算法内容。

2.1 匹配概念的信息量大小与其子父节点的匹配概率关系

概念信息量在概念的相似度计算中有着广泛的应用，其中较典型的是文献[6]中的相似度计算方法。概念信息量是概念信息的一个量化值，概念信息越丰富，概念的信息量越大，概念信息量与概念所包含的元属性(它是用来描述概念的特征，区别于描述两个实体的关系属性)数目成正比例关系，一个概念包含的元属性越多，概念的信息越丰富，假若我们把现实中的概念组织成一棵树的形式，那么概念信息量与概念包含的元属性数目，从树根节点往下看是一个单调递增的过程，因为概念的信息与元属性都具有传递性，父类的信息与元属性都传递于子类中，因此子类的信息量与元属性数目大于父类的概念信息量与元属性数目。对于两个匹配的概念，则需要两个概念有相同的元属性。根据以上分析，本文给出以下两个定理：

定理 1 在待匹配概念对的父概念对匹配的情况下，如果该父概念对的信息量越大，那么待匹配概念对的匹配的概率越大。

证明：假设待匹配概念对 c_1，c_2 的元属性集合为 A，B，两概念匹配的概率记为 $P(x)$，可得 $P(x)=P(A=B)$，假设待匹配概念对的父概念的元属性集合为 A_p，B_p，由于父子概念的元属性继承关系，可以得出 A_p，B_p 分别被包含于 A，B 中，因此当待匹配概念对 c_1，c_2 的父概念匹配时，$P(x)=P(A=B|A_p=B_p)=P(A-A_p=B-B_p)$，由于 $A_p=B_p$，故可设 $t=A_p=B_p$，所以 $P(x)=P(A-t=B-t)$。下面证明 $P(x)$ 的值随着 t 集合大小的增加而变大，我们分别设 $m=A-t$，$n=B-t$，那么 $P(x)=P(A-t=B-t)=P(m=n)$，设 m，n 集合的大小分别为 $t1$，$t2$，且设所有元属性的个数为 s。那么 $P(m=n)$ 的概率可转化为大小为 s 的集合且集合元素互不相同，分别不放回随机取 $t1$，$t2$ 个元素，且所取的 $t1$ 个元素与 $t2$ 个元素相等的概率。保证 $t1$ 和 $t2$ 个元素相等，$t1$ 和 $t2$ 的值必须相等，其相等概率为 $(1/s)$。当 $t1=t2$ 时，设 $w=t1=t2$。最后 $P(x)$ 的概率为：

$$P(x)=1/(s)*(1/((s)*(s-1)\cdots(s-w)))$$

由此可知 $P(x)$ 的大小与 w 成反比例关系，而 w 的值代表了概念的元属性的个数，可知 $P(m=n)$ 的概率与 m 和 n 的元属性个数成反比，由此可推出，$P(x)=P(A-t=B-t)$ 的概率随着 t 集合大小的增大而增大，通过元属性与概念信息量的正比例对应关系，可得出当两个待匹配概念的父概念对匹配时，如果父概念的信息量越大，则待匹配概念对匹配的概率越大。证毕。

定理 2 在两个待匹配概念对的子概念对匹配的情况下，如果该子概念对的信息量越小，那么待匹配概念对的匹配的概率越大。

证明：设待匹配概念的元属性个数分别为 n，m。由于子概念对匹配，故可设子概念对的元属性个数同为 s。由于父子的继承关系，那么待匹配概念对的所有元属性肯定被子概念的元属性包含。设待匹配概念的匹配概率为 $P(x)$，那么概率 $P(x)$ 可转换为大小为 s 的集合且集合元素互不相同，分别不放回抽取 n，m 个元素，其中 n 和 m 个元素相等的概率。保证 n 和 m 个元素相等，n 和 m 的值必须相等，其相等概率为 $(1/s)$。当 $n=m$ 时，设 $t=n=m$。最后 $P(x)$ 的概率为：

$$P(x)=(1/s)*(1/((s)*(s-1)\cdots(s-t)))$$

因此 $P(x)$ 的值跟 s 成反比例函数，s 是子概念的元属性个数，根据概念元属性与概念信息量的正比例对应关系，最后可得出，当两个待匹配概念的子概念对匹配，如果子概念对的信息量越小，则待匹配概念对匹配的概率越大。证毕。

2.2 概念信息量获取

概念信息量的获取本文采用了文献[6]的方法，概念信息量与其子孙节点的数目成反比例函数，概念的子孙节点数目越多，概念信息量越少。其概念信息量计算公式为：

$$ic(c)=1-\frac{\log(hypo(c)+1)}{\log(hypo(root)+1)} \tag{1}$$

其中 $hypo(c)$，$hypo(root)$ 分别代表概念 c 所有子孙节点与本体树中的所有节点个数。从该公式中可以看出，如果以树根节点为上，从上往下看，概念信息量是一个递增过程，孩子节点的信息量大于父亲节点。从式(1)中我们可以看出，ic 值在 0 到 1 之间。

讨论第 1 节中的例子，从式(1)中可知，此时，图 1 中的 entity 概念信息量是一个较小的值，而 horse 概念信息量是一个较大的值，根据定理 1，可知 horse 子类的匹配可能性大于 entity 子类匹配的可能性，这与实际情况相符合。

2.3 相似度传播

根据定理 1 和定理 2 可知，待匹配概念对的父概念对匹配，那么父概念对的信息量越大，待匹配概念对匹配的可能性就越大，其相似度传播的值也应该越大；待匹配概念对的子概念对匹配，那么子概念对的信息量越小，待匹配概念对匹配的可能性越大，其相似度传播的值也越大。基于以上分析本文提出语义传播因子(SF)，我们将以 SF 值来影响相似度传播的大小，SF 值越大，相似度传播值越大。

定义 1(语义传播因子 SF)

$$SF=\begin{cases} e^{-(ic(c_1)+ic(c_2))/2} & c_1\in C(c)\cap c_2\in C(c') \\ e^{(ic(c_1)+ic(c_2))/2-1} & c_1\in P(c)\cap c_2\in P(c') \end{cases} \tag{2}$$

其中 c，c' 分别是待匹配概念，$C(c)$ 表示 c 概念的孩子集合，$P(c)$ 表示 c 概念的父亲集合。由于在不同的本体中匹配的概念的信息量会有所偏差，因此我们取它们的平均之和来代表该匹配概念对的信息量。在定义中我们可以看出，当 c_1，c_2 分别是待匹配概念对的父概念时，SF 的值随着 c_1，c_2 的增加而变大，当 c_1，c_2 分别是待匹配概念对的子概念时，SF 的值随着 c_1，c_2 的增加而变小，这与定理 1，2 相符合。

根据定义 2，我们制定了相似度传播公式：

$$r_t^{k+1}=r_t^0+(1-r_t^0)\sum_{i\in N(t)}SFi^*r_t^k/N \tag{3}$$

其中 $t=1,2,\cdots,n$，$k=1,2,\cdots$，r_t^{k+1} 表示元素 t 第 $k+1$ 次迭

代的相似度值，r_{i0}为元素 t 初始相似度值，$N(t)$表示 t 元素对应邻居节点的已匹配概念对集合，r_i^k 表示 t 的邻居 i 元素在 k 次迭代时的相似度值，N 表示 $N(t)$的集合大小。

然而根据非线性数值理论[7]，该迭代公式收敛于不动点 $R^*=\{r_1^*,r_2^*,\cdots,r_n^* \mid 0\leqslant r_i^*\leqslant 1,i=0,1,\cdots,n\}$，迭代公式才有意义，故给出定理 3，并给予证明。

定理 3(迭代序列)

$$r_i^{k+1}=r_i^0+(1-r_i^0)\sum_{i\in N(t)}SF_i * r_i^k/N \quad t=1,2\cdots,n$$

$$k=1,2\cdots$$

对任意初始

$$R_0=\{R=(r_0,r_1,\cdots,r_n) \mid 0\leqslant r_i\leqslant 1,i=0,1\cdots,n\}$$

收敛于不动点 $R*$。

证明：

第一步 单调性的证明

取任意的元素 i，在第一次迭代后，易知 $r_i^1\geqslant r_i^0$，假设任意给定的 i 都满足 $r_i^k\geqslant r_i^{k-1}$，那么

$$r_i^{k+1}=r_i{}^0+(1-r_i^0)\sum_{j\in N(i)}SF_j * r_j^k/N\geqslant$$

$$r_i^0+(1-r_i^0)\sum_{j\in N(i)}SF_j * r_j^{k-1}/N=r_i^k$$

可得 $r_i^{k+1}\geqslant r_i^k$，即迭代序列是递增序列。

第二步 迭代序列单调递增于某定值的证明

根据定义 1 可知 $SF\leqslant 1$，取任意元素 i，在第一次迭代后，易知 $r_i^1\leqslant 1$，假设任意给定的 i 都满足 $r_i^k\leqslant 1$，那么

$$r_i^{k+1}=r_i^0+(1-r_i^0)\sum_{j\in N(i)}SF_j * r_j^k/N\leqslant 1$$

可得迭代后的元素相似度值必小于等于 1，把这一结果代入迭代序列中，可得

$$r_i^{k+1}\leqslant r_i^0+(1-r_i^0)\sum_{j\in N(i)}SF_j/N$$

所以元素 i 在任意次迭代后，其值收敛于 $r_i{}^*$，其中

$$r_i^*=r_i^0+(1-r_i^0)\sum_{j\in N(i)}SF_j/N$$

对于任意的 r_i，都能收敛于 r_i^*，可知 R_0 收敛于 R^*。证毕。

2.4 传播算法

针对以上分析，本文给出基于概念信息量的相似度传播的算法主要过程，算法参见图 2。其中，caculateInformation-Content()是概念信息量获取函数，其实现方法在 2.2 节中介绍。similarityPropagate()是相似度传播函数，其实现内容采用了 2.3 节中式(8)。相似度传播是一个迭代过程，迭代终止于无新映射对发现。

```
输入：similarity table after language_based ontology matching
输出：final mappings between Ont1 and Ont2
//对本体 Ont1，Ont2 的每个概念计算概念信息量
foreach(concept_i in Ont1)
caculateInformationContent(concept_i);
foreach(concept_j in Ont2)
  caculateInformationContent(concept_j);
  //迭代相似度传播，直到无新映射对发现
do
foreach(concept_i in Ont1)
  foreach(concept_j in Ont2)
    similarityPropagate(concept_i,concept_j);
repeat-do(while new mappings is found)
```

图 2 基于概念信息量的相似度传播算法

3 实验结果及分析

3.1 实验系统

在实验系统设计中，为了能够进行有效的分析，本文把 SNAX_Map[8] 系统进行调整，去除了系统的结构级策略，并以该系统的输出作为基于概念信息量相似度传播算法的输入，实现 SNAX_Map+ 系统。

3.2 实验数据

本文在以下两个数据集上做了实验：

EON. 此数据集包含了 51 个本体，这些本体都描述了书籍参考目录的信息，映射的任务是建立参考本体与 51 个本体(包括参考本体自身)之间的映射关系。在此测试数据集上 OAEI 提供了参考映射，实验根据参考映射，计算出系统在此数据集上的映射效果；

Russia. 此数据集包含了 2 个本体，分别为 russia1 和 russia2，两个本体分别描述了同一个关于 Russia 的旅游网站信息。映射任务是完成 russia1 与 russia2 映射，并根据提供的参考映射，进行实验效果分析。

3.3 评价标准

查全率(Recall)和查准率(Precision)是信息检索领域的重要指标，前者是衡量检索系统和检索者检出相关信息的能力，后者是衡量检索系统和检索者拒绝非相关信息的能力，两者结合成 F-Measure，即表示检索效果。其具体的计算方法如下所示：

$$\text{Precision}=\frac{\#\text{correct\_found\_alignments}}{\#\text{found\_alignments}}$$

$$\text{Recall}=\frac{\#\text{correct\_found\_alignments}}{\#\text{exsiting\_alignmens}}$$

$$\text{F-Measure}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

3.4 实验结果及分析

为了简便，本文只给出了 EON 数据集的平均结果。实验结果见表 2 和图 3。从图 3 中可以看出，SNAX_Map+ 系统在 EON数据集上，其 F 值比 SNAX_Map 系统提高了 7%，在 Russia 数据集上提高了 9%，说明了该传播算法能够有效地提高系统的映射性能。在查全率上，SNAX_Map+ 系统在 EON 和 Russia 数据集上分别提高了 8%，11%，在查准率上，SNAX_Map+ 系统在 EON 和 Russia 数据集上分别提高了 4%，5%，从中也可以看出 SNAX_Map+ 系统在 Russia 数据集上对查全率与查准率的提高较明显，这跟数据集的本体结构存在差异有关系，在 Russia 数据集中，本体概念数目较多，结构信息较为丰富，这一方面有利于扩大相似度传播的覆盖范围，使相似度传播的可利用资源更多，而另一方面这也有利于概念的信息量的计算，在结构信息越丰富的情况下，概念信息量的计算越精确，因此基于概念信息量的相似度传播算法在这种情况下能够更加有效，同时，在 EON 数据集中也可以发现，对于一些没有结构的或者一些扁本体该传播算法的效果比较差。根据以上分析我们可以得出，该传播算法比较适用于结构信息较为丰富的本体。

表 1 SNAX_Map 与 SNAX_Map+ 查全率查准率对比

	SNAX_Map Rec.	SNAX_MAP+ Rec.	SNAX_Map Pre.	SNAX_Map+ Pre.

EON	76%	84%	91%	95%
Russia	72%	83%	80%	85%

表 2　SNAX_Map 与 SNAX_Map+的 F 值对比

	SNAX_Map	SNAX_Map+
EON	0.82	0.89
Russia	0.75	0.84

相似度传播公式是一个迭代的公式，迭代收敛速度也是判断迭代公式好坏的一个因数，一般来说，迭代收敛速度越快，迭代公式相对也较好。本文以 SNAX_Map+系统不再产生新映射对为迭代终止记号。从表 3 中看出 SNAX_Map+系统的相似度传播迭代算法在收敛速度上还是比较好的，基本在迭代 3 次以内实现稳定，从表中也可以看出，迭代的次数跟本体的规模也有关系，规模越大，迭代次数也相应增加。

表 3　SNAX_Map+在 EON 与 Russia 数据集上的平均迭代次数

	EON	Russia
Iteration times	2.78	3

结束语　相似度传播在本体概念相似度计算中有着广泛的运用，然而目前的相似度传播算法却并未对相似度传播值进行合理定量分析。本文分析了对于已匹配的节点，其概念信息量的大小对于相似度传播的影响关系，提出了基于概念信息量的相似度传播算法，从而能够进行更精确的相似度传播。实验验证，该算法对系统的映射性能有一定的提高。

参 考 文 献

[1] Jean-Mary YR, Kabuka MR. ASMOV Results for OAEI 2007 [C]// International Semantic Web Conference (ISWC). Busan, Korean, 2007: 141-151

[2] Introduction of Falcon-AO[EB/OL]. http://xobjects.seu.edu.cn/project/falcon/matching/index.html. 2006

[3] Li Y, Zhong Q, Li J, et al. Result of Ontology Alignment with Rimom at OAEI 2007 [C]// OAEI. 2007: 227-235

[4] Melnik S, Garcia-Molina H, Rahm E. Similarity Flooding: A Versatile Graph Matching Algorithm[C]// The 18th International Conference on Data Engineering. San Jose, California, USA, February 26th-March 1st, 2002: 112-126

[5] Hu W, Jian NS, Qum YZ, et al. GMO: A Graph Matching for Ontologies[C]// K-CAP Workshop on Integrating Ontologies. Banff, Alberta, Canada, October 2005: 1-8

[6] Jiang J J, Conrath D W. Semantic Similarity Based on Corpus Statitics and Lexical Taxonomy[C]// Proceedings of International Conference Reaserch on Computational Linguistic. Tai-Wan, 1997: 1-15

[7] Ortega JM, Rheinboldt WC. Iterative Solution of Nonliner Equations in Several Variables[M]. New York: Academic Press, 1970

[8] Zhang ZW, Xu DZ, Zhang T. Ontology Mapping Based on Conditional Information Quantity[C]// Proceedings of ICNSC 2008. Sanya, 2008: 587-591

(上接第 173 页)

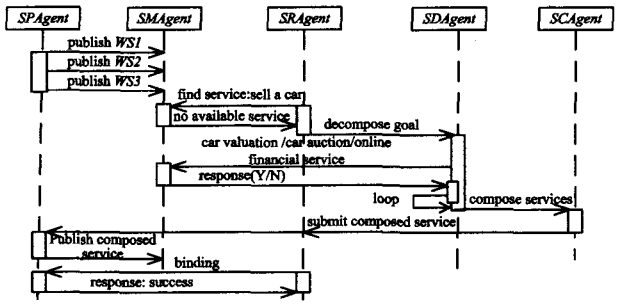

图 4　卖车服务协作顺序图

务、在线支付服务组装起来形成组合服务(3 个服务组装是个简单的顺序结构)，即汽车估价服务给卖主估计汽车可能卖出的价格；这一信息提供给拍卖行后，拍卖行就在此基础上给定拍卖价格；拍卖完成后，买卖双方在线完成支付，最后完成了这个交易。

结束语　本文通过研究 Agent 技术与 Web 技术相互间的关系和相似性，将 Agent 引入到 Web 服务的经典模型 SOA 中，提出了一种基于多 Agent 的语义 Web 服务组合框架。在该框架中，使用 Agent 来封装 SOA 中 3 个不同的角色，使得角色之间的交互表现为 Agent 的协商和协作，同时可以使用多 Agent 协作来搜索发现服务资源，发现用户意图，分解用户目标及动态规划组合服务。有了以多 Agent 为内核的 Web 服务组合管理模块，语义 Web 服务能够很好地适应当今极其开放动态的分布式计算环境，给用户提供高质量稳健的服务，同时我们认为有 Agent 封装用户意图，就可以基于用户偏好，更有针对性地提供相应的服务。下一步需要解决的是基于 Agent 的目标分解及目标-服务匹配等问题。

参 考 文 献

[1] McIlraith S, Son T C. Semantic Web Services[J]. IEEE Intelligent Systems, 2001(Special Issue on the Semantic Web)

[2] Handler J. Agents and the Semantic Web[J]. IEEE Intelligent System, 2001, Mar/Apr: 30-37

[3] 姚莉，张维明. 智能协作信息技术[M]. 北京：电子工业出版社，2002

[4] 叶云，李舟军，李梦君，等. 整合 Agent 与语义 Web 服务[J]. 计算机科学，2007，34(5)：144-146

[5] Chen H, Finin T, Joshi A, et al. Intelligent Agent meets the Semantic Web in Smart Space [J]. IEEE Internet Computing, 2004, Nov/Dec: 69-79

[6] Buhler P, Vidal J M. Semantic Web Services as Agent Behaviors[M]. Agentcities: Challenges in Open Agent Environments, LNCS/LNAI, Berlin: Springer-Verlag, 2003

[7] 任磊，李玉忱，李璟. 基于多 Agent 的 Web 服务动态合成的研究[J]. 计算机应用，2005(4)：802-804

[8] 姚世军. 基于 Agent 和 QoS 动态选择服务的方法[J]. 计算机应用，2005(12)：345-346

[9] 张尧学，方存好. 主动服务—概念、结构与实现[M]. 北京：科学出版社，2005

[10] Payne T R. Web Services from an Agent Perspective[J]. IEEE Intelligent System, 2008, Mar/Apr: 12-14