基于AHP-熵权决策的本体映射优化算法

徐德智，易晓媛，汤 哲
（中南大学 信息科学与工程学院，湖南 长沙 410083）

摘 要：针对现阶段语义网本体映射结果缺乏针对性，映射效率低下等问题，现提出一种基于AHP(Analytic hierarchy process)层次分析法及熵权决策的本体映射优化算法. 该方法首先解析目标本体，列出本体的各个特征向量，依据相应的公式计算出特征向量的置信度. 其次，给出各映射策略，依据所提出的层次分析-熵权决策组合优化算法，进行映射多策略权值计算. 最终，计算出映射结果. 该方法有效的减少了映射冗余，更合理的分配了多策略权值，实验结果证明了其择优性.

关键词：本体映射；置信度；层次分析法；熵权决策；配准策略
中图分类号：TP301.6 **文献标识码**：A **文章编号**：1000－7180(2017)11－0048－05
DOI:10.19304/j.cnki.issn1000-7180.2017.11.010

An Ontology Mapping Optimization Algorithm Based on AHP and Entropy Weight Decision

XU De-zhi, YI Xiao-yuan, TANG Zhe
(Central South University, School of Information Science and Engineering, Changsha 410083, China)

Abstract: Currently, there are many problems in mapping Semantic Web ontology. For example, Ontology mapping results are inaccurate or time complexity is too high. This paper presents an ontology mapping algorithm based on Firstly, we should analysis the ontology and list the feature vector. Secondly, according to the proposed AHP and entropy weight decision, we map the multi-strategy weights. Finally, the result of the mapping is calculated. The method effectively reduces redundant and improve the efficiency. Experiment result has proved the effectiveness of these proposed methods.

Key words: ontology mapping; confidence; analytic hierarchy process; entropy decision; registration strategy

1 引言

在当前信息化与大数据时代，语义网的作用将更加突显. 其主要目标是使计算机更能解读万维网[1]. 其存在给予了用户参与日常在线活动的权利. 语义网中，本体及本体映射是两个核心问题.

层出不穷的多策略系统带来了多种映射的选择方案. 但是，当两个或多个本体出现时，如何正确的根据其特征选择合适的映射方法，是目前比较突出的问题. 单一的系统能够得到对应单一的特征较精确的映射，而这就忽略了其他特征. 所以单一的系统并不能全面的表示出本体映射的结果.

针对这个问题，现阶段已经提出很多策略组合选择方法. 例如取平均值法，取最大值法. 目前研究关于权重分配的方法有：基于冲突集的方法[1]、基于层次分析方法[2]，基于三角模糊数的方法[3]，基于熵权决策的方法[4]等，但是这些方法也有其自身的缺陷. 因此，本文提出一种基于层次分析和熵权决策的优化方法.

2 分析样本及解析映射策略

2.1 分析本体构造及计算对应置信度

本体的定义如下：本体为一个七元组，$O(C, A^C, R, A^R, H, I, X)$. 其中 C 指概念；A^C 指概念属

收稿日期：2017－03－15；修回日期：2017－04－12
基金项目：国家自然科学基金项目(31470028)

性；R 指关系；A^R 指关系属性；H 是层次；I 是实例；X 是公理[5]. 在定义的七元组中，包含如下特征：概念特征、概念属性特征、关系特征、关系属性特征及实例特征.

2.1.1　概念特征置信度计算

参与映射的本体从概念构造来说就不尽相同. 本体一般由多个概念组合而成. 对于这些使用的概念，运用概念置信度来进行计算. 本文对两个本体的概念置信度定义如下：

$$L_{\text{value}}(O_s, O_t) = \frac{\text{Num}(\text{common}(\text{normalised}-\text{entities}))}{\min(|O_s|, |O_t|)} \tag{1}$$

normalised－entities 指的是本体中，将其格式化后的实体表示. 格式化操作，是指将本体统一成一定格式.

Num(common(normalised－entities)) 指的是两个本体中实体相同的数量. $|O_s|$ 表示本体样本一的实体数，$|O_t|$ 表示本体样本二的实体数.

2.1.2　概念属性特征置信度计算

本体概念用标签和注释进行描述. 即当两个本体的概念属性相似时，这两个本体的概念也会相似. 利用概念属性计算置信度时，需要借助于向量空间模型. 给出向量空间模型：

$$\overrightarrow{V_{ek}} = (w_{1k}, w_{2k}, \cdots, w_{Wk}) \tag{2}$$

式中，W 为向量维度. W_{ik} 表示概念属性中的关键词 t_i 出现的次数. 将两个本体中的概念属性表示成 $m \times W$ 和 $n \times W$ 的矩阵模型.

本文定义概念属性特征置信度为：

$$T_{\text{value}} = \min\left\{\frac{N_s}{m \times W}, \frac{N_t}{n \times W}\right\} \tag{3}$$

2.1.3　关系特征置信度计算

本体通过关系来描述其特征. 关系组成中原语是极其重要的描述方式. $\overrightarrow{V_n}$ 和表示原语的出现次数.

本文定义关系特征的置信度如下：

$$S_{\text{value}} = \frac{\vec{V}_1 \times \vec{V}_2}{|\vec{V}_1 \| \vec{V}_2|} \tag{4}$$

2.1.4　关系属性特征置信度计算

由于关系属性及其复杂性，基于关系属性的特征置信度也是一个重要的衡量标准. $|P_s|$ 和 $|P_t|$ 表示不同的本体包含关系属性的数目. 关系属性特征置信度如下：

$$P_{\text{value}} = \begin{cases} 0; & \text{if } |P_s| = |P_t| = 0 \\ (1 - \frac{1}{|P_s| + |P_t|}) \times \frac{\min(|P_s|, |P_t|)}{\max(|P_s|, |P_t|)}; & \text{others} \end{cases} \tag{5}$$

2.1.5 实例特征置信度计算

实例对应本体的多个概念，所以实例相对于其他本体组成来说是动态的. 实例的丰富程度由实例特征描述. 本文定义本体实例特征置信度如下：

$$I_{\text{value}} = \begin{cases} 0; & \text{if } |I_s| = |I_t| = 0 \\ (1 - \frac{1}{|I_s| + |I_t|}) \times \frac{\min(|I_s|, |I_t|)}{\max(|I_s|, |I_t|)}; & \text{others} \end{cases} \tag{6}$$

$|I_s|$ 和 $|I_t|$ 表示实例特征中的实例信息数目.

2.2　映射策略实现方法

2.2.1　基于概念的映射策略

该映射策略将概念当作字符串，通过比较字符串的差异，计算相似度. 现给出一种计算方法：I-SUB[6] 方法. 其公式定义如下：

$$\text{Sim}_{\text{L-value}}(s_1, s_2) = \text{Comm}(s_1, s_2) - \text{Diff}(s_1, s_2) + \text{Winkler}(s_1, s_2) \tag{7}$$

$$\text{Comm}(s_1, s_2) = \frac{2 * \sum_i \text{length}(\text{maxCommSubString}_i)}{\text{length}(s_1) + \text{length}(s_2)} \tag{8}$$

$$\text{Diff}(s_1, s_2) = \frac{d\text{Len}_{s1} * d\text{Len}_{s2}}{p + (1-p) * (d\text{Len}_{s1} + d\text{Len}_{s2} - d\text{Len}_{s1} * d\text{Len}_{s2})} \tag{9}$$

式中，s_1、s_2 分别表示了两个概念的字符串.

2.2.2　基于概念属性的映射策略

基于概念属性的映射策略，首先解析各个概念对应的虚拟文档，将虚拟文档转换成相对应的向量. 然后使用余弦公式，算出概念间的相似度.

$$\text{Sim}_{T-\text{value}}(C_1, C_2) = \cos(\overrightarrow{N_1}, \overrightarrow{N_2}) = \frac{\sum_{k=1}^{d} n_{ki} n_{kj}}{\sqrt{(\sum_{k=1}^{d} n_{ki}^2)(\sum_{k=1}^{d} n_{kj}^2)}} \tag{10}$$

对此公式进行简单的解释. $\overrightarrow{N_1}$ 表示概念 C_1 对应的向量，$\overrightarrow{N_2}$ 表示概念 C_2 对应的向量，n_{ki}、n_{kj} 分别表示相对应的元素.

2.2.3　基于结构的映射策略

本方法是利用本体的结构特征，进行本体映射，在此选用常用的基于结构的映射方法 SF 方法. 该方法将本体的结构表示为 RDF 图，然后用相似度迭代，将邻居点的信息结合，计算最终的基于结构特征的相似度.

$$\text{Sim}_{s-\text{value}}(C_1, C_2) = \sum_{n=0}^{n} \sigma_n(a_n, b_n) \tag{11}$$

设映射初始值为 $\sigma_0(x_0,y_0)=1.0$. 该方法运用到邻居节点进行一个综合计算，利用周围的邻居节点信息，综合初始相似度，产生新的相似度结果[7-8].

2.2.4 基于属性特征的映射策略

计算属性特征的相似度，采用树结构的计算方法，其定义如下：

$$\text{Feature}(C)=\begin{cases}\text{Feture0}; & \text{if } C=\text{root}\\ \text{Feature}(\text{Parent}(C))\cup F(C); & \text{otherwise}\end{cases} \tag{12}$$

相似度计算公式为：

$$\text{Sim}_{P-\text{value}}(C_1,C_2)=\frac{|\text{Feature}(C_1)\cap\text{Feature}(C_2)|}{|\text{Feature}(C_1)\cup\text{Feature}(C_2)|} \tag{13}$$

2.2.5 基于实例特征的映射策略

在本文中，选用基于实例特征的方法是 Jaccard 方法[8]，该方法是实例特征中最典型及最广泛的用法. A 表示概念 C_1 中含有的实例集的数目，B 表示概念 C_2 中含有的实例集的数目，其公式如下：

$$\text{Sim}(A,B)=\frac{P(A\cap B)}{P(A\cup B)} \tag{14}$$

3 AHP -熵权决策优化算法

3.1 优化算法详细步骤

本文提出的层次熵权优化算法，将层次分析法和熵权决策法有机的结合起来. 在求取综合权重时，将两种方法的中间计算过程结合起来，得到主观客观都合理的权值分析. 熵权法和层次分析法已经是常识性算法，具体计算步骤不再解释，本节着重讨论结合后的优化算法. 详细步骤如下：

Step1：层次分析法求解出各个权值. 利用层次分析法的判断矩阵，求出 m 个上层权值 $B=\{\beta_1,\beta_2,\cdots,\beta_m\}$，求出 n 个子层权值 $\Phi=\{\varphi_1,\varphi_2,\cdots\varphi_n\}$.

Step2：利用熵权决策算法，求出各个权重分别为 $A=\{\alpha_1,\alpha_2,\cdots,\alpha_n\}$.

Step3：将子层权值与上层权值分开，子层权值结合熵权决策，重新计算子层的综合权值 $T=\{\tau_1,\tau_2,\cdots,\tau_n\}$，并进行更新. 每一步子层求解的权值计算公式为：

$$\tau_i=\varphi_i\alpha_i/\left(\sum_{i=1}^{n}\varphi_i\alpha_i\right). \tag{16}$$

Step4：求出新的子层权值之后，将子层与上层权值综合计算，再次更新子层权值，求出新权重为 $T=\{\tau_{11},\tau_{12},\cdots,\tau_{1n_1},\tau_{1n_2}\}$，在此基础上，并分别对每一上层准则下子准则的综合权重归一化得：$\Omega''=\{\omega''_{11},\omega''_{12},\cdots,\omega''_{mm}\}$ 计算公式为：

$$\omega''_{ij}=\tau_{ij}/\sum_{j=1}^{k}\tau_{ij},k=n_1,n_2\cdots,n_m;i=1,2,\cdots,m. \tag{16}$$

Step5：将上层准则权重 B 与所求得的综合权重 Ω'' 对应相乘，得权重

$$\Omega'=\{\omega'_{11},\omega'_{12},\cdots,\omega'_{mm}\},\ \omega'=\beta_i\omega''_{ij},i=1,2,\cdots,m;j=1,2,k;k\in\{n_1,n_2,\cdots,n_m\} \tag{17}$$

Step6：将 $\Omega'=\{\omega'_1,\omega'_2,\cdots,\omega'_n\}$ 归一化表示为 $\Omega=\{\omega_1,\omega_2,\cdots,\omega_n\}$，其中计算公式为：

$$\omega_i=\omega'_i/\sum_{i=1}^{n}\omega'_i,i=1,2,\cdots,n. \tag{18}$$

经过以上六个步骤，优化了层次分析法和熵权决策法给予的权值分配，将这两者权值分配合理的结合在一起，最终得到各个策略的综合权值.

3.2 实验参考样本数据

通过以上提出的改进方法，给出参考数据如表 1 所示，画出层次分析法实验结构图如图 1 所示，进行计算并给出计算结果，验证本方法的优化作用. 在参考实验数据理论计算完成后，设计系统实验，将设计思路运用到实际本体映射中.

表 1 实验参考样本

置信度 / 本体	*L*-value	*T*-value	*S*-value	*P*-value	*I*-value
O_1/O_2	0.97	0.92	0.71	0.23	0.51
O_2/O_3	0.38	0.39	0.21	0.13	0.38
O_1/O_3	0.11	0.21	0.19	0.17	0.12

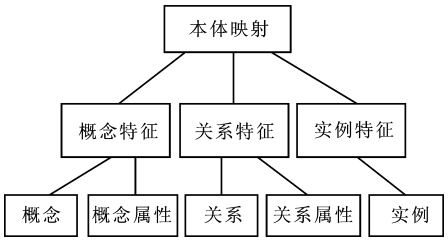

图 1 基于层次分析结构图

基于实验参考样本，利用层次分析法得到判断矩阵如表 2 所示，由此计算出层次分析法的权值分配情况. 再依据熵权决策的定义，计算出熵权分配的权值. 最后，依据本文提出的优化方法，计算各个权值分配. 给出层次分析法，熵权法，本文优化方法的各个权值分配如表 3 所示. 利用计算出的权值分配结果，求出总的相似度. 计算结果如表 4 所示，其计算公式为：$\text{sim}=e_l*l+e_t*t+e_s*s+e_p*p+$

$e_i * i$

表 2　层次分析法判断矩阵

置信度	L-value	T-value	S-value	P-value	I-value
L-value	1	1/2	1/3	1/5	1/7
T-value	2	1	1/4	5	2
S-value	3	4	1	2	1/3
P-value	5	1/5	1/2	1	1/7
I-value	7	1/2	3	7	1

表 3　总相似度对比汇总表

方法 / 本体	熵权法	层析分析法	本文优化方法
O_1/O_2	0.815	0.552	0.687
O_2/O_3	0.347	0.315	0.319
O_1/O_3	0.159	0.164	0.170

式中，l,t,s,p,i 分别表示上述的相似度值 $\mathrm{sim}_{L-\mathrm{value}}$，$\mathrm{sim}_{T-\mathrm{value}}$，$\mathrm{sim}_{S-\mathrm{value}}$，$\mathrm{sim}_{P-\mathrm{value}}$，$\mathrm{sim}_{I-\mathrm{value}}$，$e_l,e_t,e_s,e_p,e_i$ 分别表示了通过优化算法后得到的各个相似度权值分配值. 计算结果如表 4 所示.

表 4　各个相似度权值计算结果

权值 / 方法	e_l	e_t	e_s	e_p	e_i
熵权法	0.324	0.313	0.175	0.027	0.161
层次分析法	0.051	0.240	0.247	0.103	0.359
优化方法	0.075	0.340	0.195	0.127	0.263

分析实验参考数据，每一种方法计算所得的结果都有一定差异. 但是由单一的一种方法来计算权值，就有一定的局限性. 由实验数据可以看出，在熵权决策和层次分析法得出的总相似度差异较大的时候，优化方法可以得到一个比较综合的数据，例如本体 O_1/O_2，但当前两种方法得到的总相似度差异不明显的时候，例如本体样本 O_2/O_3 和 O_1/O_3，本文的优化方法得到的效果就不明显. 但综合来说，本优化方法结合了两者的优势，那么可以得到较为准确的数据，这样使本体映射的多策略选择更为合理.

4　实验与分析

4.1　实验描述及实验系统

本次实验，将层次分析法和熵权决策优化算法运用到源本体和目标本体的映射过程中间[9-10]. 采用 Benchmark 集作为实验本体. 将优化算法导入到 FAMS 系统的 selector 中，输入目标本体，给出权值分配，最后得到映射的结果[11]. F 系统基本原理如图 2 所示.

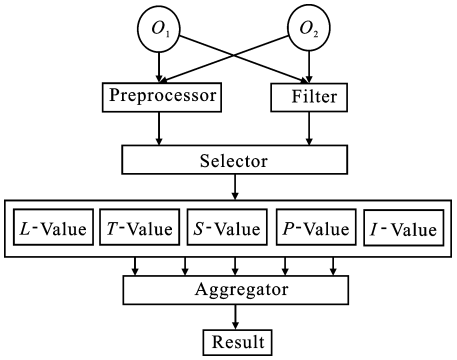

图 2　FAMS 系统原理示意图

4.2　评价标准

在本体映射的系统中，业界使用的评价标准是查全率和查准率，其定义如下：

定义 1　查全率(Recall)：

$$\mathrm{Recall}(A,R)=\frac{|A\cap R|}{|R|} \tag{20}$$

定义 2　查准率(Precision)：

$$\mathrm{Precision}(A,R)=\frac{|A\cap R|}{|A|} \tag{21}$$

式中，A 表示最终映射结果，R 为参考值.

4.3　实验结果

系统使用了优化算法，经过映射权值主客观综合，得到的 F 系统映射结果为表 5 所示.

表 5　FAMS 系统映射结果

	#100-104	#201-210	#221-247	#248-266	ALL
Prec.	1.00	0.95	0.98	0.91	0.96
Rec.	1.00	0.88	0.98	0.65	0.87

其中 Prec. 对应为查准率，Rec. 为查全率.

与其他系统的其他方法进行查准率与查全率的比较，结果比较如图 3 所示.

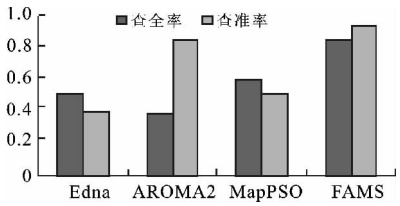

图 3　F 系统与其他系统实验结果比较图

由图 3 可知，F 系统利用了提出的优化方法，合理的分配各个映射策略权值，在查全率和查准率都基本满足的情况下，能够得到比较好的实验结果. 与之相比较的系统，在时间复杂度上都基本一致. 由此也说明了本方法的有效性[12-13].

5 结束语

本文在本体映射多策略权值分配问题下[14]，提出了一种基于层次分析法和熵权决策的改进优化方法.该方法结合了层次分析法的主观运用，也使用了熵权决策的客观关联，在查准率和查全率能够满足预期的情况下，该算法提高了映射效率，并且能够更准确全面的得到映射结果.做到主客观综合评价，将主客观两种方法的中间过程相结合求出综合权重.使得本体映射结果反映数据本身，并且符合实际应用.

参考文献：

[1] Helou M A, Palmonari M, Jarrar M. Effectiveness of automatic translations for cross-lingual ontology mapping[J]. J. Artif. Intell. Res. (JAIR), 2016, 55(1): 165-208.

[2] Zhang L Y, Ren J D, Li X W. OIM-SM: A method for ontology integration based on semantic mapping [J]. Journal of Intelligent & Fuzzy Systems, 2017, 32 (3): 1983-1995.

[3] Wang Z, Bie R, Zhou M. Hybrid ontology matching for solving the heterogeneous problem of the IoT[C]// Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference on. Liverpool, United Kingdom, IEEE, 2012: 1799-1804.

[4] Wang R, Wang L, Liu L, et al. Combination of the improved method for ontology mapping[J]. Physics Procedia, 2012, 25(22):2167-2172.

[5] Shvaiko P, Euzenat J. Ontology matching: state of the art and future challenges[J]. Knowledge and Data Engineering, IEEE Transactions on,2013,25(1):158-176.

[6] Euzenat J, Shvaiko P. Ontology matching[M]. Heidelberg: Springer, 2007.

[7] Li Y, Li J Z, Zhang D, et al. Result of ontology alignment with RiMOM at OAEI'06[C]// Proceedings of the ISWC'2006 Workshop on Ontology Matching. USA, 2006, 173-182.

[8] Isaac A, Van Der Meij L, Schlobach S, et al. An empirical study of instance-based ontology matching[M]. Berlin, Heidelberg, Springer,2007: 253-266.

[9] 赖雅，王润梅，徐德智. 基于参考点的大规模本体分块与映射[J]. 计算机应用研究，2013，30(2)：469-471.

[10] 邹亮，徐德智，郭维. 基于参考点的大规模本体扩散映射算法[J]. 小型微型计算机系统，2013，34(7)：1507-1513.

[11] Kandpal A, Goudar R H, Chauhan R, et al. Effective ontology alignment: an approach for resolving the ontology heterogeneity problem for semantic information retrieval[M]. India, Springer, 2014: 1077-1087.

[12] 唐雅媛，徐德智，赖雅. 基于概念特征的语义相似度计算方法[J]. 计算机工程，2012，38(5)：170-172.

[13] 廖晖寰. 基于本体特征的自适应映射方法研究[D]. 长沙:中南大学，2014.

[14] 郭金维，蒲绪强，高祥，等. 一种改进的多目标决策指标权重计算方法[J]. 西安电子科技大学学报，2014，41(6)：118-125.

作者简介：

徐德智　男，(1963-)，博士后，教授.研究方向为 Web 计算、语义数据库.

易晓媛(通讯作者)　女，(1991-)，硕士研究生.研究方向为语义网、本体映射. E-mail:624906884@qq.com.

汤　哲　男，(1977-)，副教授.研究方向为人工智能.