

一种基于加权领域本体的语义检索方法

张 亮^{1,2} 屈振新¹ 丁 菰¹ 唐胜群¹

(武汉大学软件工程国家重点实验室 武汉 430072)¹ (武汉工程大学管理学院 武汉 430074)²

摘 要 提出了新方法 WOSR, 以对已经被本体概念标注的领域信息资源进行语义检索。WOSR 方法首先建立领域本体, 再采用均等概率分布方法为概念赋权, 然后通过概念的权重求出概念相似度, 最后计算用户检索请求和信息资源之间的语义相似度, 并根据相似度的大小排序输出检索结果。实验结果表明, WOSR 方法比其他经典方法的检索效果更好。

关键词 语义检索, 领域本体, 相似度

中图法分类号 TP391 **文献标识码** A

Semantic Retrieval Based on Weighted Domain Ontology

ZHANG Liang^{1,2} QU Zhen-xin¹ DING Song¹ TANG Sheng-qun¹

(State Key Lab of Software Engineering, Wuhan University, Wuhan 430072, China)¹

(School of Management, Wuhan Institute of Technology, Wuhan 430074, China)²

Abstract This paper proposed an approach called WOSR for the semantic retrieval of domain resources that have been previously annotated with ontology concepts. Firstly WOSR developed domain ontology and associated weights with concepts through uniform probability distribution. Secondly the similarity between concepts was computed by using their weights. Finally calculated semantic similarities between a user request and the domain resources, and the answer was returned in the form of a ranked list ordered by semantic similarity. Experiments show that WOSR yields a better performance than other classical methods.

Keywords Semantic retrieval, Domain ontology, Similarity

1 引言

如何在浩如烟海的互联网信息海洋中检索到所需的信息资源至今仍是一个研究热点。传统的信息检索方法都是基于关键词的检索。基于关键词的检索在查全率和查准率方面有明显的缺点, 比如不能进行泛化、细化等语义操作, 无法计算出检索需求与信息资源之间的符合程度等。

针对这些问题, 本文提出了基于加权领域本体的语义检索方法 WOSR。本体具有良好的概念层次结构且支持逻辑推理, 可以表述特定领域的知识体系。利用本体概念之间的相似度能够更准确、更全面地描述用户的查询请求和领域资源的语义, 实现领域信息资源的语义检索。

WOSR 方法的核心是计算概念向量的相似度, 因此也是以组成两个向量的本体概念的相似度计算为基础的。WOSR 方法包括 3 个阶段: 第一个阶段是建立领域本体, 并采用均等概率分布方法为概念赋权。第二个阶段是计算配对的两个概念之间的相似度。第三个阶段是求出两个概念向量之间的相似度, 也就是用户查询请求和信息资源之间的相似度, 并根据相似度的大小排序输出检索结果。WOSR 方法以信息资源

的标注向量为基础进行检索, 而不是直接检索目标资源, 这样可以在检索时避免信息资源的固有结构的影响。WOSR 方法的另一个特点是将加权领域本体作为信息资源标注的基础。

本文第 2 节介绍相关研究工作; 第 3 节给出相关定义和术语; 第 4 节详细说明 WOSR 方法的算法; 第 5 节对 WOSR 方法进行实验评价; 最后是总结。

2 相关研究

关于语义检索的研究文献^[1-4]已有很多。本文关注的重点是以两个概念向量之间的相似度计算为基础的方法。

大多数现有方法利用的 WordNet 是通用的词典本体, 不局限于某一特定领域。而且 WordNet 只包含简单的术语, 不包含多词汇术语, 比如“Business Intelligence”。而针对特定领域的语义检索需要特定的领域本体, WOSR 方法是以包含加权概念的领域本体为基础的, 与这些使用通用本体的方法不一样。

文献[5]提出的方法使用 NLP 技术对本体进行了强化, 附加了词频和倒排文档频率(tf-idf), 但是这些附加信息并未

到稿日期: 2009-08-14 返修日期: 2009-11-03 本文受 973 国家重点基础研究发展计划(2007CB310800)资助。

张 亮(1973-), 男, 博士生, 讲师, 主要研究方向为语义 Web、商务智能等, E-mail: seperson@163.com; 屈振新(1972-), 男, 博士生, 副教授, 主要研究方向为软件工程、语义 Web 等; 丁 菰(1980-), 男, 博士生, 主要研究方向为软件工程等; 唐胜群(1946-), 男, 教授, 博士生导师, 主要研究方向为软件工程、电子政务等。

真正体现概念的语义内容。另外,在给与本体元素有关联的术语赋权时,该方法提出了5个相关度水平对应于5个常数值:直接相关(1.0),强相关(0.7),一般相关(0.4),弱相关(0.2)和不相关(0.0)。而在WOSR方法中,概念的权重可以在区间[0,1]上取任何值。

文献[6]定义了词语之间的相似度,其中每一个词语和给定的IS-A层次结构中的一个概念相联系。这个相似度综合考虑了词语之间的路径长度、层次结构中词语包含的深度以及词语的局部语义密度。文献[7]根据本体的拓扑结构和本体中概念的位置得出概念节点的权重,通过建立权重的向量求出概念相似度。但是这两个方法都采用了Miller&Charles实验^[8]进行实验评价,而这个实验方法是以WordNet为基础的,并不适用于特定领域的检索效果分析。WOSR方法考虑得更加全面,将概念之间的相似度仅仅作为更加全面的一个步骤,最终目标是求出两个概念向量的相似度。而且在实验中采用了人工判断参照组的结果作为评价依据,以便更真实地反映领域资源的语义。

其他研究工作与WOSR方法类似,都是关注于两个概念的集合或者向量之间的相似度。文献[9-12]的方法经常被用来比较概念向量。然而,这几个方法的概念相似度是通过严格的匹配计算出来的,两个概念向量的匹配是基于它们的交集,没有考虑概念所包含的信息量。而WOSR方法采用更加精确的语义匹配方法,两个概念的匹配考虑了它们共享的信息量,而且向量相似度的计算是基于最优概念配对的。

文献[13]采用了新算法来计算属于同一个本体的两个概念集合之间的语义距离,是基于Dijkstra算法的扩展在图中搜索最短路径。相对于WOSR方法,这里的相似度是基于概念之间的路径距离,而不是每个概念具有的信息量。而且相邻概念之间的相似度是开发者在设计本体时就确定的,因此对结果有一定影响。

综上所述,WOSR方法综合考虑了概念的信息量、本体拓扑结构等信息,适用于某个特定领域的语义检索,第5节的对比实验也表明WOSR方法比一些经典方法的检索效果更好。

3 定义

本节介绍相关的定义和术语,并给出了形式化定义。

定义1(本体) 本体是共享概念模型的明确的形式化的规范说明^[14]。在我们的研究中将本体 O 定义为一个使用IS-A包含关系将概念组织起来的树状层次结构,即

$$O = \langle C, H \rangle$$

式中, C 是概念集合, H 是 C 中的概念对的集合,定义为

$$H = \{ \langle c_i, c_j \rangle \in C \times C \mid \text{subs}(c_i, c_j) \}$$

式中, $\text{subs}()$ 是成对的概念之间的包含关系。因为已假定本体 O 是一个树状结构,那么给定两个概念 $c_i, c_j \in C$,则 c_i 和 c_j 的上确界 $\text{lub}(c_i, c_j)$ 就是在本体中同时包含 c_i 和 c_j 的层次最低的那个概念。

定义2(加权领域本体) 给定一个本体 $O = \langle C, H \rangle$,一个加权领域本体WDO定义为

$$\text{WDO} = \langle O, w \rangle$$

式中, w 是在概念集合 C 上定义的函数,给定一个概念 $c \in C$,则 $w(c)$ 是区间[0,1]上的有理数,表示与概念 c 相关的权重。

定义3(领域资源库DR) 领域资源库是所有使用WDO的概念标注的领域资源的集合。

定义4(资源标注向量) 给定一个本体 $O = \langle C, H \rangle$ 和一个领域资源 $dr_i \in DR$,资源 dr_i 的资源标注向量 rav_i 是本体概念的集合,定义为

$$\text{rav}_i = (c_{i,1}, \dots, c_{i,m})$$

式中, $c_{i,j} \in C, j=1, \dots, m$ 。

定义5(概念外延) 给定一个本体 $O = \langle C, H \rangle$ 和一个概念 $c \in C$,概念 c 的外延 $E(c)$ 定义为

$$E(c) = \{ dr \in DR \mid c \in \text{rav}_{dr} \}$$

因此,给定一个概念 c ,概念外延 $E(c)$ 提供了领域资源DR中所有被概念 c 标注的资源,也就是说,这些资源的标注向量 rav 包含概念 c 。

定义6(相似概念外延) 给定一个本体 $O = \langle C, H \rangle$ 和一个概念 $c \in C$,概念 c 的相似概念外延 $S(c)$ 定义为

$$S(c) = \{ dr \in DR \mid \exists c' \in \text{rav}_{dr} \text{consim}(c, c') > k \}$$

式中, consim 是概念相似度,将在第4节详细介绍,变量 k 是根据实际情况确定的阈值。给定概念 c ,则它的相似概念外延 $S(c)$ 包含的领域资源的标注向量 rav 包含概念 c' ,它和概念 c 的相似度大于某一给定的阈值。换句话说,领域资源中被概念 c 及其相似概念 c' 所标注的资源都属于相似概念外延 $S(c)$ 。

定义7(请求向量的外延) 给定一个本体 $O = \langle C, H \rangle$,请求向量 rv 定义为

$$rv = (c_1, \dots, c_n)$$

式中, $c_i \in C (i=1, \dots, n)$,请求向量 rv 的外延 $E(rv)$ 定义为

$$E(rv) = \bigcup_{i=1, \dots, n} (E(c_i) \cup S(c_i))$$

式中, $E(c_i)$ 和 $S(c_i)$ 分别是概念 $c_i (i=1, \dots, n)$ 的概念外延和相似概念外延。因此 $E(rv)$ 包含的领域资源的标注向量 rav 包含 rv 的至少一个概念,或者 rav 包含的概念与 rv 中的至少一个概念相似。

从语义相似度的角度可以将请求向量 rv 的外延 $E(rv)$ 定义为

$$E(rv) = \{ dr \in DR \mid \text{semsim}(\text{rav}_{dr}, rv) > h \}$$

式中, $\text{semsim}(\text{rav}_{dr}, rv)$ 是 rav_{dr} 与 rv 之间的语义相似度,将在第4节详细介绍。 h 是根据实际情况确定的阈值。因此,一个请求向量的外延所包含资源的标注向量 rav 和请求向量 rv 的语义相似度大于给定的阈值。

4 语义检索算法

本文提出的WOSR方法首先建立领域本体并为本体的概念赋权,再利用概念相似度求出查询请求向量和资源标注向量的语义相似度,最后根据语义相似度的大小将检索结果筛选、排序返回给用户。为便于说明,采用了电子商务行业中服装商品领域的一个案例。

4.1 权重的设置

电子商务营销的商品种类繁多,图1演示的是服装领域本体的局部。这是一个用IS-A包含关系将概念组织起来的树状层次结构,其中的概念已经设置了权重。我们采用均等概率分布方法为概念确定权重。这个树状结构的根称为Thing,代表最抽象的概念,它的概率 $w(\text{Thing})$ 等于1。假设某个概念 c 的概率为 $w(c)$,它的子概念 $c_i (i=1, \dots, n)$ 的数目

为 n , 则每一个子概念的概率 $w(c_i) = w(c)/n$ 。这些概率就是为概念设置的权重, 表示概念的信息量的多少。例如, 概念“Cotton”的权重 $w(\text{Cotton}) = 0.03$, 这是因为 $w(\text{Natural}) = 0.08$ 以及概念“Natural”有 3 个子概念。

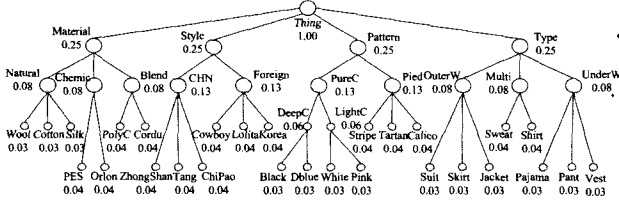


图1 已经赋权的服装商品领域本体

利用该服装领域本体的概念, 已经对一些服装商品进行了标注。每一个标注就是一个商品的资源标注向量 rav , 所有标注向量共同构成了领域资源库 DR , 以便于用户的检索。

4.2 概念相似度的计算

假设某用户的查询请求是“想购买一件棉外衣, 颜色深蓝, 中式风格”, 则可以把该请求形式化为请求向量:

$$rv = (\text{OuterW}, \text{Cotton}, \text{Dblue}, \text{CHN}) \quad (1)$$

一旦确定了查询请求向量 rv , WOSR 方法就能够在 rv 和领域资源库的每一个 rav 之间求出语义相似度 $semsim$ 。正如上面介绍过的, 语义相似度 $semsim$ 的计算需要先求出构成向量的概念之间的概念相似度 $consim$ 。

对概念相似度的计算借鉴了文献[15]定义的概率方法。按照信息量理论, 概念 c 的信息量 $I(c)$ 可定义为

$$I(c) = -\log w(c)$$

式中, $w(c)$ 是概念 c 的权重。概念越抽象, 越靠近概念层次结构的顶端, 权重越大, 它的信息量就越多。

给定两个概念 c_i 和 c_j , 它们的相似度 $consim(c_i, c_j)$, 被定义为两个概念共享的最大信息量除以两个概念的信息量^[11], 因为本体的概念层次是一个树状结构, 概念 c_i 与 c_j 的上确界 $lub(c_i, c_j)$ 就是该树状结构中两个概念节点的共同的父节点中层次最低的那一个, 该父节点提供了两个子概念在分类学方面共享的最大信息量。概念相似度计算公式为

$$consim(c_i, c_j) = 2\log w(lub(c_i, c_j)) / (\log w(c_i) + \log w(c_j))$$

例如在本服装商品领域的案例中, 概念“Cotton”和“Wool”的概念相似度计算过程为

$$consim(\text{Cotton}, \text{Wool}) = 2\log w(\text{Natural}) / (\log w(\text{Cotton}) + \log w(\text{Wool})) = 0.70$$

这是因为概念“Natural”是概念“Cotton”和“Wool”的上确界, 提供了被两个子概念共享的最大信息量。

4.3 语义相似度的计算

本部分介绍如何使用概念相似度 $consim$ 求出用户的查询请求向量 rv 与服装商品的资源标注向量 rav 的语义相似度 $semsim$ 。简单地讲, 就是在两个向量之间找到最佳的概念配对, 使得这些概念对的概念相似度的总和最大, 这个最大的总和就是向量的语义相似度。

两个向量的笛卡尔积形式化定义为

$$rv \otimes rav = \{(c_i, c_j)\}$$

式中, $i = 1, \dots, n, j = 1, \dots, m, n = |rv|, m = |rav|, c_i \in rv, c_j \in rav$ 。

如前所述, 对于每一对概念都能够计算出概念相似度

$consim$, 然而, 我们不需要考虑所有可能的概念对。这是因为有些配对没有实际意义, 例如概念“Cotton”和“ChiPao”的配对。因此采用独占式的配对, 即某一概念只能参与一次概念配对, 也就是说不允许概念的一配多。语义相似度的计算公式为

$$semsim(rv, rav) = \max(\sum consim(c_i, c_j)) / n$$

式中, n 为查询请求向量的基数。

对于语义相似度 $semsim(rv, rav)$ 的计算, 借鉴了权重二分图的最大匹配问题的思想^[16], 采用的算法是以著名的匈牙利算法^[17] (Hungarian Algorithm) 为基础的。

例如, 对于式(1)中的请求向量 rv 和资源标注向量 $(\text{Shirt}, \text{PolyC}, \text{DeepC}, \text{Korea})$, 概念对的集合 $\{(\text{OuterW}, \text{Shirt}), (\text{Cotton}, \text{PolyC}), (\text{Dblue}, \text{DeepC}), (\text{CHN}, \text{Korea})\}$ 有最大的概念相似度 $consim$ 的总和 0.58, 其它所有的配对方案都将导致一个更小的总和, 因此语义相似度为 0.58。计算过程为

$$\begin{aligned} consim(\text{Shirt}, \text{OuterW}) &= 0.49 \\ consim(\text{Cotton}, \text{PolyC}) &= 0.41 \\ consim(\text{Dblue}, \text{DeepC}) &= 0.89 \\ consim(\text{CHN}, \text{Korea}) &= 0.53 \\ semsim(rv, rav_1) &= (0.49 + 0.41 + 0.89 + 0.53) / 4 = 0.58 \end{aligned}$$

4.4 检索结果的排序

用户检索到的结果将以语义相似度的减序排列的形式返回给用户, 但只有语义相似度高于给定的阈值 h 的检索结果才返回给用户。

5 实验评价

5.1 语义相似度实验评价

我们对 WOSR 方法进行了初步的实验评价, 实验采用的加权领域本体和领域资源库分别见图 1 和表 1, 这些领域资源包括的服装商品 $R_i (i = 1, \dots, 12)$ 都由资源标注向量 rav_i 进行了标注。用户的查询请求向量如式(1)所示。评价依据是所有对比方法和人工参照组的结果的相关度。人工参照组由 30 名服装设计专业的学生组成, 每位组员对用户查询要求和那些服装商品资源之间的相似度做出估值, 并取其平均值。

表1 服装领域资源库的案例

资源	向量名	向量元素
R1	rav1	(Shirt, PolyC, DeepC, Korea)
R2	rav2	(Jacket, Cotton, Pied, CHN)
R3	rav3	(OuterW, Silk, Calico, ChiPao)
R4	rav4	(Suit, Black, Wool)
R5	rav5	(Shirt, PureC, Cowboy)
R6	rav6	(UnderW, Cotton, White)
R7	rav7	(OuterW, Silk, Pied, Tang)
R8	rav8	(Skirt, PES, Calico, Lolita)
R9	rav9	(Jacket, Blend, LightC, Cowboy)
R10	rav10	(Pant, LightC)
R11	rav11	(OuterW, Blend, Black, ZhongShan)
R12	rav12	(Sweat, Wool, Stripe)

表 2 展示的语义相似度计算结果包括了人工参照组、WOSR 方法以及具有代表性的经典方法 Dice 方法^[10]、Jaccard 方法^[12] 和 Salton 方法^[11] 的计算结果。实验结果表明, 相比于其他经典方法, WOSR 方法的相似度计算结果和人工参照组的判断结果相关度最高。

表2 语义相似度计算结果对比

资源	参照组	WOSR	Jaccard	Dice	Salton
R ₁	0.63	0.61	0.14	0.25	0.06
R ₂	0.91	0.96	0.60	0.75	0.19
R ₃	0.66	0.73	0.14	0.25	0.06
R ₄	0.50	0.58	0.00	0.00	0.00
R ₅	0.48	0.44	0.00	0.00	0.00
R ₆	0.30	0.54	0.17	0.29	0.08
R ₇	0.82	0.80	0.33	0.50	0.13
R ₈	0.52	0.41	0.00	0.00	0.00
R ₉	0.58	0.62	0.00	0.00	0.00
R ₁₀	0.64	0.50	0.50	0.67	0.25
R ₁₁	0.74	0.76	0.14	0.25	0.06
R ₁₂	0.55	0.48	0.17	0.29	0.08
相关度	1.00	0.81	0.64	0.62	0.50

5.2 检索效果评价

为评价返回给用户的检索结果,我们仍然以人工参照组的结果为标准来评估 WOSR 方法。在表 3 中对比展示了人为判断方法和 WOSR 方法的检索结果,商品资源已经按照语义相似度进行了排序,根据阈值 h 确定的基准线将资源分成两部分,基准线以上的资源将返回给用户。在这里阈值 h 定为 0.55。

表3 检索结果排序列表

人工参照组		WOSR 方法	
检索结果	相似度	检索结果	相似度
R ₂	0.91	R ₂	0.96
R ₇	0.82	R ₇	0.80
R ₁₁	0.74	R ₁₁	0.76
R ₃	0.66	R ₃	0.73
R ₁₀	0.64	R ₉	0.62
R ₁	0.63	R ₁	0.61
R ₉	0.58	R ₄	0.58
R ₁₂	0.55	(基准线)	
R ₈	0.52	R ₆	0.54
R ₄	0.50	R ₁₀	0.50
R ₅	0.48	R ₁₂	0.48
R ₆	0.30	R ₅	0.44
		R ₈	0.41

检索性能的常用衡量指标是查准率、查全率以及 F-measure,其计算方法不再赘述。根据表 3 的结果,如果将人工参照组的结果确定为有效资源,则求出查全率为 75%,查准率为 86%,F-measure 值为 80%,WOSR 方法的检索效果总体表现良好。

结束语 本文主要研究基于加权领域本体的语义检索方法,该方法采用均等概率分布方法为领域本体的概念赋权,再使用信息量理论的方法求出概念之间的相似度,并以此为基础得出请求向量和资源标注向量之间的语义相似度,最后根据语义相似度的大小将检索结果筛选、排序返回给用户。实验表明,与其他经典方法相比较,本文方法的相似度计算结果和人工参照组结果的相关度最高,并且检索效果方面的衡量指标也表现良好。

在后续的研究中我们准备进一步研究使用非均等概率分布方法为本体概念赋权,另外再考虑将权重和请求向量的元素联系起来,允许用户指定查询词的重要程度。

参 考 文 献

[1] Shvaiko P, Euzénat J. A Survey of Schema-based Matching Ap-

proaches[C]// Spaccapietra S, et al., eds. Journal on Data Semantics IV. Heidelberg: Springer-Verlag Berlin, 2005: 146-171

- [2] Madhavan J, Halevy Alon Y. Composing Mappings among Data Sources[C]// Proceedings of the 29th International Conference on VLDB, Berlin, Germany, 2003: 572-583
- [3] Maguitman A G, Menczer F, Roinestad H, et al. Algorithmic Detection of Semantic Similarity[C]// Proceedings of the 14th international conference on World Wide Web. Chiba, Japan, 2005: 107-116
- [4] 宋二伟, 刘宗田, 徐力斌, 等. 基于领域本体的 Web 信息检索实现机制研究[J]. 计算机科学, 2007, 34(5): 104-106
- [5] Fang W D, Zhang L, Wang Y X, et al. Towards a Semantic Search Engine Based on Ontologies[C]// Proceedings of 4th International Conference on Machine Learning and Cybernetics. Guangzhou, China, 2005: 1913-1918
- [6] Li Y, Bandar Z A, McLean D. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources[J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(4): 871-882
- [7] Kim J W, Candan K S. CP/CV: Concept Similarity Mining without Frequency Information from Domain Describing Taxonomies [C]// Proceedings of the 15th ACM international conference on information and knowledge management. Virginia, USA, 2006: 483-492
- [8] Miller G A, Charles W G. Contextual Correlates of Semantic Similarity[J]. Language and Cognitive Processes, 1991, 6(1): 1-28
- [9] Castano S, Antonellis V D, Fugini M G, et al. Conceptual Schema Analysis: Techniques and Applications[J]. ACM Transactions on Databases Systems, 1998, 23(3): 286-333
- [10] Frakes W B, Baeza-Yates R. Information Retrieval, Data Structure and Algorithms [M]. Englewood Cliffs: Prentice Hall, 1992: 225-236
- [11] Maarek Y S, Berry D M, Kaiser G E. An Information Retrieval Approach for Automatically Constructing Software Libraries [J]. IEEE Transactions on Software Engineering, 1991, 17(8): 800-813
- [12] Jaccard P. Bulletin del la Société Vaudoise des Sciences Naturelles 37[M]. 1901: 241-272
- [13] Cordi V, Lombardi P, Martelli M, et al. An Ontology-Based Similarity between Sets of Concepts[C]// Proceedings of WOA 2005. Italy, 2005: 16-21
- [14] Gruber T R. A Translation Approach to Portable Ontologies [J]. Knowledge Acquisition, 1993, 5(2): 199-220
- [15] Lin D. An Information-Theoretic Definition of Similarity[C]// Proceedings of the 15th international conference on Machine Learningpp. Wisconsin, USA, 1998: 296-304
- [16] Formica A, Missikoff M. Concept Similarity in SymOntos: an Enterprise Ontology Management Tool[J]. Computer Journal, 2002, 45(6): 583-594
- [17] Munkres J. Algorithms for the Assignment and Transportation Problems[J]. Journal of the Society of Industrial and Applied Mathematics, 1957, 5(1): 32-38