一种用于本体排序的内容分析方法 *

徐德智, 刘怡静
(中南大学 信息科学与工程学院, 长沙 410083)

摘 要: 针对使用传统的用于本体排序的方法得到的排序结果不够准确的问题,提出了一种新的内容分析方法。首先通过构造本体的概念模型提取本体的主题词集合得到本体的主题相似度;然后通过对关键词所在的本体上下文进行分析,得到本体相对于关键词的上下文相关度;最后结合主题相似度和上下文相关度得到本体相对于关键词的综合评价值并进行排序。实验结果表明,该方法可以有效地提高本体排序的准确性。

关键词: 本体排序; 主题相似度; 上下文相关度
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2010)06-2127-03
doi:10.3969/j.issn.1001-3695.2010.06.038

Content analysis method used for ontology ranking

XU De-zhi, LIU Yi-jing
(*School of Information Science & Engineering, Central South University, Changsha* 410083, *China*)

Abstract: The ontology ranking result was not precise enough using traditional method. Aiming at this problem, this paper proposed a new content analysis method. First, obtained the topic similarity according to the keywords set which was extracted by constructing the ontology conceptual model. Second, got the context relatedness through analyzing the ontology context. Finally, combined the topic similarity and context relatedness to calculate the ranking value which was used to order the related ontologies. The experimental result shows that the method can improve the precision of the ontology ranking result.

Key words: ontology ranking; topic similarity; context relatedness

万维网中本体数目日益增长,大量表示相同或相似概念的本体相继出现,而寻找并重用已有的本体并不是一件容易的事情,导致其重用度低并产生了大量的冗余信息,为了帮助用户寻找并理解相关的本体,本体搜索引擎应运而生,而本体搜索引擎返回的搜索结果的排序直接影响到用户搜索的效率。为了提高本体搜索结果的准确性,本文提出了一种用于解决本体搜索中排序问题的内容分析方法,通过计算本体的主题相似度和上下文相关度得到本体综合评价值并进行排序。

1 相关工作

对语义 Web 中本体进行排序是语义 Web 众多排序问题中的一个关键问题[1]。随着本体技术的发展,不同的用于本体排序的评价方法被提出。OntoRank[2]是常用的本体搜索引擎 Swoogle 所使用的算法,它是一种类似网页排序的基于链接分析的方法,该算法考虑了可能存在于不同本体之间的多种语义关联,为不同关联指定相应的权值,再根据类似网页排序的基于链接分析的思想来评价本体的重要性。此外,文献[3]中提出了 AKTiveRank 算法,综合考虑了四种不同标准对本体的重要性进行评价,包括类匹配标准、密度标准、语义相似度标准和中介性标准。文献[4]中提出的 OntoQA 算法则是从本体模式和本体实例两个方面提出不同的标准来评价本体。

然而,OntoRank 算法没有考虑关键词在本体中的重要性对排序结果的影响和本体本身的结构,因此准确率比较低; AKTiveRank 和 OntoQA 算法虽然考虑了本体的结构信息和关键词对结果的影响,但是只考虑了对关键词描述的丰富程度,没有考虑其差异性,因此得到的结果均不理想。

2 本体的内容分析方法

2.1 主题相似度的计算

根据用户提供的关键词,本体搜索引擎可能搜索出大量的相关本体。本体对关键词的描述是否恰当与本体和本体所描述的主题是密不可分的。例如,用户想搜索关于“Person”的本体,关键词“Person”可能会以分类的形式出现在描述“Agent”的本体中,也有可能在描述“Publish”本体中,以与“Publish”概念具有“has Author”的关系出现。根据结构特征分析,这两个概念具有相似的重要性,但关键词与“Agent”的相似度要明显高于与“Publish”的相似度,显然用户会认为描述“Agent”的本体更加符合需要。因此,本体的内容分析应该考虑关键词与本体主题的相似度。

为了得到关键词与相关本体间的主题相似度,首先要得到本体的主题词信息。本体元素中,概念在表现本体主题信息方面所起到的作用显然要远远大于属性、实例等元素,因此本体的主题提取算法主要考虑本体的概念结构信息。本文根据本体的概念模型得到本体主题词集,本体的概念模型构造算法如算法 1 所示。

收稿日期: 2009-12-02; **修回日期**: 2010-01-31 **基金项目**: 国家自然科学基金资助项目(60970096);湖南省国土资源科技资助项目(200718);中南大学研究生学位论文创新资助项目(2009ssxt183)

作者简介:徐德智(1963-),男,湖南长沙人,教授,博士后,主要研究方向为 Web 计算、语义 Web;刘怡静(1984-),女,河北沧州人,硕士研究生,主要研究方向为语义 Web(liuyijing21@163.com).

算法 1 构造本体概念模型

输入:本体;输出:本体概念模型。

a)解析本体,得到本体中所有的概念及概念之间的父子关系、等价关系、组成关系,及以概念作为 domain 的宾语的属性数目与概念所具有的 onProperty 属性数目之和 n_{pd} 等。

b)将每个概念看做本体概念模型中的一个节点。

c)将具有等价关系的概念节点合并成一个节点。

d)对于任意概念 C_1,如果 C_1 是概念 C_2 的子概念,或 C_1 是 C_2 的组成部分,则增加一条 C_1 节点指向 C_2 节点的边。

e)将每个概念的 n_{pd} 更新成其子概念中 n_{pd} 的最大值。

f)如果所有的概念节点合并成一个有向图则完成;否则给定 Thing 概念作为公共父概念,为每个分图的根节点添加一条指向 Thing 概念的边。

本体的概念模型中包含多个概念节点,各个概念所起到的作用是不同的,并不是所有的概念都能描述本体的主题信息。在一个本体中,各个概念的分类信息分布得并不均匀,有些概念的子概念多一些,有些概念子概念少一些。为了描述本体中概念分类信息的不同,给出了定义 1。

定义 1 概念的子概念饱和度、宽度饱和度和深度饱和度

假设在本体概念模型中,概念 C_i 的父节点有 n 个孩子($C_1, C_2, \cdots, C_n$)。其中:$C_i(1 \leqslant i \leqslant n)$ 的子概念数用 s 表示,直接子概念数用 w 来表示,子树深度用 d 来表示,那么 $C_i.SAT_w = C_i.w / \max(C_k.w)$,$C_i.SAT_d = C_i.d / \max(C_k.d)$,$C_i.SAT_s = (C_i.s + C_i.n_{dp}) / \max(C_k.s + C_k.n_{dp})(1 \leqslant k \leqslant n)$ 分别叫做概念 C_i 的宽度饱和度、深度饱和度和子概念饱和度。

在本体的概念模型中,每个概念对于本体主题的贡献并不相同,因此应为主题词集中的每个概念分配不同的权值。而本体相对于关键词的重要性不仅要与关键词与相关主题词之间的相似度有关,还应与相关主题词在本体主题词集中的权重有关。一个概念是否能在某种程度上体现本体的主题信息,一方面与该概念的父概念分类的细致程度也就是该概念在兄弟概念中所处的位置有关;另一方面又与该概念本身分类的情况有关。根据以上的描述,本文给出两条本体主题词提取规则。

规则 1 一个本体概念模型中的任意节点 R 有 $n(n \geqslant 1)$ 个孩子(记为 $C_1, C_2, \cdots, C_n$),如果该本体中对 R 的孩子 C_i $(1 \leqslant i \leqslant n)$ 的描述远远多于对 R 的其他孩子的描述,那么可以认为节点 C_i 比 R 更能表现该本体的主题信息,而 R 的其他孩子对主题信息的贡献可以忽略不计。

规则 2 假设一个概念有非常细致的分类,即该概念的直接子概念的数目特别地多,那么其中的任何一个直接子概念都不能准确表现主题信息,除非该概念有少量的直接子概念相对于其他直接子概念来说具有非常详细的描述,那么这少量的直接子概念就能够代表部分主题的信息。

根据以上分析,本文给出在本体的概念模型中本体主题词权值的分配算法,如算法 2 所示。

算法 2 本体概念模型中概念处理算法。

输入:本体概念模型,当前概念。

输出:当前概念的子概念中可以作为主题词的概念及其权值分配。

a)如果 $C_2.SATs / C_1.SATs < \alpha$,则将 C_1 加入到集合 topic 和集合 wait 中,$C_1.f$ 的值置为 2,并忽略概念 C_1 的兄弟节点,即 $C_2.f \cdots C_n.f = -1$,同时,将 current 的权值赋给 C_1,降低 current 的权值,如式(1)(2)所示,转到 e)。dep(C_1)的定义如文献[5]所示。

$$C_1.\text{weight} = \text{current.weight} \tag{1}$$

$$\text{current.weight} = \text{current.weight} \times (1 - 1/2^{\text{dep}(C_1)}) \tag{2}$$

b)如果 $C_v.SATs < \alpha (1 \leqslant v \leqslant n)$,则忽略概念 $C_v \cdots C_n$,即将 $C_v.f \cdots C_n.f$ 的值置为 -1。

c)对于每个没被忽略的 current 的子概念,则

(a)如果 $C_i.SAT_w > \beta$ 且 $C_i.SAT_d < \gamma$,则将 C_i 加入到集合 topic 中,忽略 C_i,置 $C_1.f$ 值为 -1;

(b)如果 $C_i.SAT_w < \beta$ 且 $C_i.SAT_d < \gamma$,则通过值 $m = n_i / k_i$ 来选择操作,若 $m > 3$,则将 C_i 加入到集合 topic 和 wait 中,置 $C_1.f$ 值为 2;否则将 C_i 加入到集合 topic 中,忽略 C_i;

(c)如果 $C_i.SAT_d \geqslant \gamma$,则将 C_i 加入到集合 topic 和 wait 中置值 $C_1.f$ 为 2。

d)根据式(3)为 current 概念的子概念中抽取的主题词分配权值;

$$C_i.\text{weight} = \text{current.weight} \times (C_i.SATs / \sum_{j=1}^{v-1} C_j.SATs) \tag{3}$$

e)将概念 current 从集合 wait 中移除。

以上给出了本体概念模型中概念处理算法。本体主题词提取过程如算法 3 所示。

算法 3 本体主题词集提取算法

输入:本体概念模型;输出:本体的带权主题词集合。

a)初始化变量;

topic = {root}, wait = {root}, root.weight = 1, current = root

b)获得当前概念 current 的直接子概念 $C_1, C_2, \cdots, C_n$。

c)对每一个概念 C_i 计算其深度饱和度 $C_i.SAT_d$、宽度饱和度 $C_i.SAT_w$ 和子类饱和度 $C_i.SAT_s$。

d)根据 $C_i.SAT_s$ 将 current 的子概念进行排序,得到新的 $C_1, C_2, \cdots, C_n$。

e)根据算法 2 对 current 进行处理。

f)如果 wait = ∅, 转到 g),否则若集合 wait 中不存在概念 C 满足 $C.f = 1$,则对集合 wait 中的每一个概念 $C.f - -$,再在 wait 中选择概念 C 满足 $C.f = 1$ 赋给 current,转到 b)。

g)如果 Thing ∈ topic,那么将概念 Thing 从 topic 中移除。

h)输出 topic 集合中的元素及其权值。

在算法 2 和 3 中,n_i 为概念 C_i 的子概念的数目,k_i 为 C_i 的直接子概念数目;$C_i.f$ 为标志位,$C_i.f = 1$ 表示 C_i 与当前正在处理的概念为兄弟概念,即等待处理的概念;$C_i.f = 2$ 表示概念 C_i 为当前处理概念的下一级的概念,只有当 $C_i.f = 1$ 的概念全部处理完之后才会考虑这一级的概念;$C_i.f = -1$ 表示概念 C_i 已经处理完毕或者不需要进一步处理。α、β、γ 为选择因子。其中,α 的取值不能太大也不能太小,取值过大会导致某些主题信息被丢弃,取值太小会误将不能表示主题信息的概念加入到主题信息中,一般情况下取 $\alpha = 0.25$;β 和 γ 主要用来区分概念分类的深度与广度的不同对处理方式的影响,文中取 $\beta = 0.4, \gamma = 0.3$。

根据上述算法得到了本体的带权主题词集合。根据关键词与本体主题词的匹配关系,可以将对本体的主题相似度 rel_t 计算分为两种情况。当关键词与本体主题词集中元素完全匹配时,主题相似度即为主题词的权值;否则根据文献[6]中所给的方法,计算关键词与本体中权值最大的主题词之间的相似度来得到关键词与本体的主题相似度,如式(4)所示。

$$\text{rel}_t(k, O) = \begin{cases} \text{weight}(C) & \text{关键词与主题词 } C \text{ 完全匹配} \\ \text{sim}(k, \max(C)) & \text{不存在与关键词完全匹配的主题词} \end{cases} \tag{4}$$

其中:k 代表用户输入的关键词;$\max(C)$ 代表主题词集中权值最大的主题词。

2.2 上下文相关度计算

上下文相关度用来描述对本体中某个概念描述的认可程度,文中用向量来表示本体的上下文描述信息。首先根据所有本体对相关词的描述信息构造标准向量;然后分别对每个本体构造上下文描述向量;最后通过本体的描述向量与标准向量之间的相关度来表示本体的上下文相关度。

文中对上下文相关度的计算主要基于以下两个规则:

规则 3 本体中对关键词上下文的描述信息在越多的本体中出现,则该描述信息与关键词的相关度越大。

规则 4 如果一个描述信息只在极少数本体中出现,不能

单纯地认为该信息与本体的相关度小。

根据上述思想,用户给出的关键词的上下文相关度的计算可以描述为算法 4。

算法 4　本体上下文相关度计算

输入:相关本体,查询关键词;

输出:本体对查询关键词的上下文相关度。

a)依次遍历各个相关本体(O_i),得到各个本体中与关键词相对应的概念的描述信息集合(Dpt_i)。其中,Dpt_i 包括 O_i 中与关键词相对应的概念的子概念、属性等。

b)求所有本体对关键词描述信息集合的并集 Dpt,构造标准描述向量 $D(d_1 \cdots d_n)$。其中:$d_i \in \mathrm{Dpt} \wedge \forall i \neq j, d_i \neq d_j$,建立关键词描述矩阵 A。矩阵的 A 中的元素定义如下:

$$a_{ij} = \begin{cases} 0 & d_j \notin \mathrm{Dpt}_i \\ 1 & d_j \in \mathrm{Dpt}_i \end{cases}$$

c)统计各个描述信息在相关本体中出现的总次数,并计算各个描述信息在所有相关本体的流行度 p。

$$\mathrm{cnt}_j = \sum_{i=1}^{n} a_{ij} \qquad p_j = \mathrm{cnt}_j / N$$

其中:N 为候选本体的个数。

d)设定流行度阈值 $pt(0 < pt < 1)$,将关键词描述矩阵中出现的描述根据阈值分成两个集合 set_p、set_i。

$$\mathrm{set}_p = \{d_j \mid p_j \geqslant p_t, 1 \leqslant j \leqslant n\}, \mathrm{set}_i = \{d_j \mid p_j < p_t, 1 \leqslant j \leqslant n\}$$

e)分别计算每个描述的相关度,对于流行度大于阈值的描述,其相关度值即为流行度值;对于流行度小于阈值的描述,相关度则需要由计算得到。

$$tr_j = \begin{cases} p_j & d_j \in \mathrm{set}_p \\ \mathrm{rel}(k, d_j) & d_j \in \mathrm{set}_i \end{cases}$$

其中:$\mathrm{rel}(k, d_j)$ 的值可以根据文献[6]中的方法计算。

f)构建关键词描述相关度向量 $T = (t_1, t_2, \cdots, t_j, \cdots, t_n)$,向量中的元素 $t_j = tr_j$。

g)计算每个本体的上下文相关度:$\mathrm{rel}_c(O_i) = \mathrm{sim}(A_i, T)$。其中:$\mathrm{sim}(A_i, T)$ 为 A_i 与 T 的余弦相似度。

2.3　本体综合评价值的计算

根据上文所提出的用于本体排序的评价方法,本体的内容分析方法可由本体的主题相似度、上下文相关度来表示。对本体的综合评价值计算为

$$\mathrm{ranking}(O_i) = w_1 \times \mathrm{rel}_t(k, O_i) + w_2 \times \mathrm{rel}_c(O_i) \qquad (5)$$

其中:w_1 和 w_2 分别为主题相似度与上下文相关度的权值,可以对 w_1 和 w_2 进行调整,使其能够满足最终要求。

3　实验结果与分析

大量对语义 Web 中本体排序问题的研究只是停留在研究水平,并没有成熟的评价标准。本文采用文献[7]中的评价方法,使用 Swoogle 的排序结果做对比实验来评价本文的算法。以"Person"作为输入关键词,可得 Swoogle 的排序结果如下所示(去掉失效链接后的结果):

a. http://xmlns.com/foaf/0.1/index.rdf
b. http://rdfs.org/sioc/ns
c. http://inferenceweb.stanford.edu/2004/07/iw.owl
d. http://swrc.ontoware.org/ontology
e. http://morpheus.cs.umbc.edu/aks1/ontosem.owl
f. http://swrc.ontoware.org/ontology/portal
g. http://inference-web.org/2.0/pml-provenance.owl
h. http://www.aktors.org/ontology/portal
i. http://ebiquity.umbc.edu/ontology/publication.owl
j. http://ebiquity.umbc.edu/ontology/person.owl
k. http://lsdis.cs.uga.edu/projects/semdis/opus
l. http://swrc.ontoware.org/ontology/ontoware
m. http://ebiquity.umbc.edu/ontology/event.owl

以 Swoogle 所得到的结果作为输入,可得使用本文算法的排序结果以及人类主观排序结果,如表 1 所示。本文选取 Swoogle 中符合人类主观排序结果的前 10 个进行分析,并对本文中算法取 $w_1 = 0.7, w_2 = 0.3$;$w_1 = 0.8, w_2 = 0.2$;$w_1 = 0.85, w_2 = 0.15$ 三组数据进行比较。

表 1　排序结果的比较

主观结果	Swoogle 结果	$w_1=0.8$ $w_2=0.2$	$w_1=0.7$ $w_2=0.3$	$w_1=0.85$ $w_2=0.15$	主观结果	Swoogle 结果	$w_1=0.8$ $w_2=0.2$	$w_1=0.7$ $w_2=0.3$	$w_1=0.85$ $w_2=0.15$
1	1	3	4	3	6	8	2	3	2
9	3	7	6	7	8	9	10	10	9
3	4	4	2	4	2	10	1	1	1
5	6	6	7	6	10	11	8	8	8
7	7	9	9	10	4	12	5	5	5

使用皮尔森相关系数,分别对表 1 中 Swoogle 和本文排序结果与人类主观排序结果之间的关系进行评价,得出皮尔森相关系数如表 2 所示。皮尔森相关系数越接近于 1,说明所得到的结果与主观排序结果之间越符合线性关系。

表 2　不同排序方法的皮尔森相关系数比较

方法	Swoogle	$w_1=0.8$ $w_2=0.2$	$w_1=0.7$ $w_2=0.3$	$w_1=0.85$ $w_2=0.15$
皮尔森相关系数	0.270	0.758	0.721	0.745

根据表 2 所示,使用本文方法所得的结果与人类主观排序结果比较的皮尔森相关系数的绝对值要明显大于用 Swoogle 所得的结果,用本文方法所得的皮尔森相关系数更接近 1,因此本文提出的方法与主观的方法得到的结果更相近。根据表 2 结果所示,当取权值为 $w_1 = 0.8, w_2 = 0.2$ 时,所得的结果要略优于其他取值,因此一般选取 $w_1 = 0.8, w_2 = 0.2$。

4　结束语

排序问题是语义 Web 中的一个重要的研究课题,对搜索引擎而言具有很高的研究价值。对语义 Web 中的本体进行排序,有助于提高本体的重用度并节省了开发的消耗。本体的内容分析方法不局限于本体之间的链接关系,还为不同的本体主题词和不同的描述信息提供了不同的权值,使得排序结果能够更加符合人类的主观结果,提高本体排序的准确度。

笔者未来的工作主要包括:对本体的结构和内容信息进行更进一步的研究,得出多关键词搜索对本体评价值的影响;对本体的主题词集抽取算法进行进一步的完善,得到一种能够帮助用户快速理解本体的新方法。

参考文献:

[1] 张祥,瞿裕忠. 语义网中的排序问题[J]. 计算机科学,2008,35(2):196-200.

[2] FININ T, SACHS J, PARR C. Finding data, knowledge, and answers on the semantic Web[C]//Proc of the 20th International FLAIRS Conference. [S.l.]: AAAI Press,2007: 2-7.

[3] ALANI H, BREWSTER C, SHADBOLT N. Ranking ontologies with AKTiveRank[C]//Proc of the 5th International Semantic Web Conference. Berlin: Springer,2006: 5-9.

[4] TARTIR S, ARPINAR I B. Ontology evaluation and ranking using OntoQA[C]//Proc of the 1st IEEE International Conference on Semantic Computing. Washington DC: IEEE Computer Society, 2007: 185-192.

[5] 徐德智,郑春卉,PASSI K. 基于 SUMO 的概念语义相似度研究[J]. 计算机应用,2006,26(1):180-183.

[6] WU Zhi-biao, PALMER M. Verbs semantics and lexical selection[C]//Proc of the 32nd Conference on Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 1994: 133-138.

[7] RAJAPAKSHA S K, KODAGODA N. Internal structure and semantic Web link structure based ontology ranking[C]//Proc of the 4th International Conference on Information and Automation for Sustainability. 2008: 86-90.