

文章编号:1001-9081(2011)03-0677-03

doi:10.3724/SP.J.1087.2011.00677

带有高效索引的语义 Web 服务 I/O 匹配优化方法

冯 勇,方 欣,徐红艳

(辽宁大学 信息学院,沈阳 110036)

(fyxuh@163.com)

摘要:目前 Web 环境中蕴涵着大量的 Web 服务和 Web 服务请求,基于语义的 Web 服务匹配能够提高 Web 服务发现的准确性,但由于其复杂的语义计算,导致系统响应速度慢。首先,对语义 Web 服务过程进行了分析,确定大量的语义计算主要集中在输入/输出(I/O)匹配环节;然后,在研究现有 I/O 匹配算法和分析影响语义相似度的主要因素基础上,给出了一种带有高效索引的语义 Web 服务 I/O 匹配优化方法,包括:高效索引的建立和基于哈希二次探测再散列的启发式筛选机制的提出;最后,通过实例证明了该方法切实可行。该方法通过筛除无关 Web 服务,减少了语义计算量,提高了系统响应速度,进而带来了更好的用户体验。

关键词:Web 服务;I/O 匹配;语义;索引;本体

中图分类号:TP393.09 **文献标志码:**A

I/O matchmaking optimization method of semantic Web service with efficient index

FENG Yong, FANG Xin, XU Hong-yan

(Information College, Liaoning University, Shenyang Liaoning 110036, China)

Abstract: A great deal of Web services and requests exist in Web environment. Web services matchmaking based on semantic can improve accuracy of service discovery. Because of complicated semantic calculation, the reaction rate of Web service matchmaking was slow. Firstly, this paper analyzed the process of semantic Web service matchmaking to make clear that the large amount of semantic calculation existed in Inputs/Ouputs (I/O) matchmaking phase. Secondly, an I/O matchmaking optimized method of semantic Web services with efficient index was put forward on the basis of the studies on I/O matchmaking algorithms and main influence factors of semantic similarity, which included the creation of efficient index and the raise of the heuristic filter mechanism based on the re-hash secondary detection. Finally, the proposed method was proved to be feasible and rational via an instance. The proposed method can reduce semantic calculation and promote reaction rate by filtering some irrelevant Web services. Furthermore, the experience of users can be improved.

Key words: Web service; Inputs/Ouputs (I/O) matchmaking; semantic; index; ontology

0 引言

目前,开放的 Web 环境是人们获取信息的重要来源,Web 环境中蕴涵着海量信息资源。面向商务应用这些信息被组织成形式多样、复杂程度不同的 Web 服务^[1],如何从大规模的 Web 服务集合中快速而准确地查找到合适的服务已为学界和业界所关注。为此,有学者将语义信息添加到 Web 服务中,提出了基于语义的 Web 服务匹配^{[2][3]},即语义 Web 服务匹配。一般情况下,语义 Web 服务匹配过程是:遍历所有的发布 Web 服务,让每一个发布的 Web 服务与用户请求的 Web 服务的文本描述、输入、输出、前提条件、效果和服务质量进行语义匹配,计算它们的综合匹配度。最后依据综合匹配度对服务进行筛选和排序,将最后结果返回给用户。

上述匹配过程需要复杂的语义计算,计算量较大,当发布的服务数量很大时,会导致计算时间过长,严重影响了用户体验,因此对 Web 服务语义匹配的时间效率提出了新的要求^[3]。为提升 Web 服务语义匹配的时间效率,裴韶亮通过设置 3 个过滤器,即文本描述匹配过滤器、输入输出过滤器、服务质量匹配过滤器,逐步缩小候选集合,每经过一个过滤器剪

枝掉一部分不符合要求的候选 Web 服务^[4]。本文在此基础上,对计算量较大的输入输出过滤器(I/O 匹配环节)加以改进,提出了一种带有高效索引的语义 Web 服务 I/O 匹配优化方法,即通过建立高效索引和给出基于高效索引的筛选机制以削减候选 Web 服务的数量,减少语义计算量,从而提升语义 Web 服务 I/O 匹配效率。

1 语义 Web 服务中的 I/O 匹配

语义 Web 服务匹配过程包括两个方面的匹配:功能属性匹配主要包括输入、输出、前提、效果(Inputs、Outputs、Preconditions、Effects, IOPE)匹配和非功能属性匹配主要是服务质量属性(Quality of Service, QoS)匹配。在 IOPE 中 I/O 可以理解为数据的变换,主要描述 Web 服务提供什么样的功能,指明了服务所需要的输入参数和执行服务后产生的结果。因此,在语义 Web 服务匹配过程中 I/O 匹配占有重要地位,目前研究的重点主要集中于此。对 I/O 匹配研究具有代表性的成果有弹性匹配算法^{[2][3]},该算法查准率高,很好地利用了语义 Web 中基于描述逻辑的推理技术;基于几何距离算法^[5]具有简单直接的优点;基于信息容量算法^[6]能够较真实

收稿日期:2010-09-25;修回日期:2010-11-24。

基金项目:辽宁省自然科学基金资助项目(20102083);中国博士后科学基金面上资助项目(20100471474)。

作者简介:冯勇(1973-),男,辽宁沈阳人,副教授,博士,主要研究方向:商务智能、语义 Web; 方欣(1989-),女,江西抚州人,硕士研究生,主要研究方向:商务智能、语义 Web; 徐红艳(1972-),女,辽宁丹东人,副教授,硕士,主要研究方向:数据库、Deep Web。

地反映了客观事实;基于属性的算法则很好地模拟了人类大脑思维模式。这些算法从概念匹配的不同方面着手,有效地提升了 I/O 匹配的准确性和服务效率,但仍存在着一定的不足,如弹性匹配算法对于服务间的匹配结果定义得过于简单且没有考虑本体的融入;当概念间相距的边数相同时,基于几何距离算法就无法得出正确的结果;基于信息容量算法不能区别具有相同最近公共节点的概念对间的细微的语义区别;基于属性算法要求对每个概念的属性进行详细的描述,复杂程度高。

鉴于单一算法存在的不足,文献[7~9]融合基于几何距离算法和基于信息容量算法的优点,提出了基于语义距离的相似度算法,把概念间的语义距离用它们在本体层次结构中的距离来量化。一般说来,两个概念间的语义距离越小,概念相似程度越高,反之越低。影响语义距离的因素主要有以下 3 点:语义深度、最近公共节点、连接关系。

1)语义深度。语义深度是指概念在本体层次结构中的深度。如果在本体层次结构中,相距边数相同的概念语义距离随着语义深度的总和的增加而减少。

2)最近公共节点。最近公共节点是指本体层次结构中两概念之间深度最大的公共节点,表明了两个概念间的相同程度。在概念间相距边数相同的情况下,最近公共节点的深度越大,此时的概念间的相似度也随之增大。

3)连接关系。在本体层次结构中有很多二元关系,如 is_a 关系、part_of 关系、引用关系、反义关系、功能关系等。连接关系直接决定了连接强度,如 is_a 关系的语义相似度大于 part_of 关系的语义相似度。但 Rada 等认为对语义网相似性的评估实际上应该只限于 is_a 型的层次分类结构,而不考虑其他的连接类型^[10]。因此,本文将研究的语义相似度限定为基于 is_a 类型的连接关系。

在此基础上,Lin 等人提出了基于语义距离的语义相似度计算公式^{[7]873},见式(1),该式考虑了概念 c1 和 c2 间的相距边数和 L 与最近公共节点概念在本体层次结构中所处的深度 H:

$$\text{Sim}(c1, c2) = e^{-\alpha L} \cdot \frac{e^{\beta H} - e^{-\beta H}}{e^{\beta H} + e^{-\beta H}} \quad (1)$$

其中:两个概念的语义相似度关于 L 单调递减,关于 H 单调递增;α 和 β 用来调整 L 和 H 对概念相似度的影响程度。根据 Lin 等人的测试,当 α = 0.2, β = 0.6 时能够获得最佳度量效果^{[7]877}。

2 语义 Web 服务 I/O 匹配优化

本体构建时,可能会因本体不完善导致 Web 服务关键词在本体中无法定位,本文使用基于本体云影模型的本体进化进程^[11]来自动提升本体完善性,其通过语义评注、本体融合、本体重构、本体解体和本体回归等 5 个过程来实现未出现的 Web 服务关键词添加到本体中,完善本体的层次结构。在本体完善的前提下,为解决大规模 Web 服务集合环境下语义 Web 服务 I/O 匹配效率低下的问题,本文以上节所述基于语义距离的语义相似度计算为基础,首先依据服务请求建立 Web 服务的高效索引,然后给出基于高效索引的筛选机制,过滤掉无关的 Web 服务,从而减少语义计算量,提高语义 Web 服务 I/O 匹配效率。

2.1 高效索引的建立

根据式(1)可知,概念间相距边数和与最近公共节点深

度是影响概念相似度计算的关键。所以,在建立索引时着重考虑概念间相距边数和与最近公共节点深度两个主要的影响因子,并且最近公共节点深度越大,服务和请求概念间的语义相似度越大。而当最近公共节点深度确定时,服务和请求间的相距边数和越大,其语义相似度越小。故此,给出索引建立规则如下:1)以最近公共节点深度(根节点所在层次为 1)由大到小的顺序对 Web 服务进行索引;2)当最近公共节点深度相同时,以概念间相距边数和由小到大的顺序对 Web 服务进行索引;3)当最近公共节点深度与概念间相距边数都相同时,以 ID 号(字母序)由小到大的顺序对 Web 服务进行索引。构建的索引表结构如图 1 所示,包括:记录最近公共节点深度的顺序表(位于图的左侧)和在深度相同条件下按概念间相距边数和由小到大排列的 Web 服务表(位于图的右侧),它们之间通过指针链接。



图 1 索引表结构

为更好地说明索引表的建立,给出如下示例(见图 2),其中图 2(a)为给定的本体片段,虚线部分表示服务请求对应的概念(假设它与发布服务对应的概念 C 重合),其他概念 A~H 分别表示发布服务的 ID 号,根据上述索引建立规则和索引表结构可以构建如图 2(b)所示的索引。

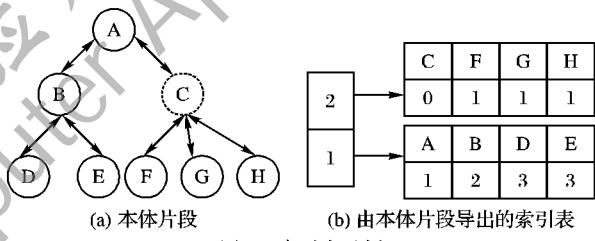


图 2 索引表示例

2.2 Web 服务筛选机制

索引表建立后可以根据其自身所带信息对大规模 Web 服务集进行初筛,过滤掉一些不相关的服务。由于服务概念按照它与请求概念的最近公共节点深度有序排列,它们之间的最近公共节点深度越深,相似度越大,当超过设定阈值时即找到匹配的服务。故此,采取的 Web 服务筛选策略是首先找到邻近满足请求但具有较小相似度的服务的置信区,该区域内的服务可能匹配请求也可能不匹配;然后在置信区内定位第二个不满足请求的服务的位置(简称第二服务点),将其后的服务过滤掉,这样在尽可能过滤掉不相关服务的同时,最大限度地保留满足请求的服务。在上述 Web 服务筛选策略指导下,本文给出了一种基于哈希二次探测再散列^[12]的启发式筛选机制,筛选机制被分为三个阶段,各阶段具体内容如下。

1)确定置信区。该阶段首先依据最近公共节点深度对顺序表进行折半查找,每次折半后计算对应服务表中最后节点的相似度,低于阈值向上折半,直到找到满足请求的服务,则上一次计算的服务点附近即为置信区;高于阈值向下折半,直到找到不满足请求的服务,则当前服务点附近即为置信区。

2)定位第二个不满足请求的服务的位置。上述确定置信区过程中既已包含了第一个不满足请求的服务的定位(简称第一服务点),接下来从第一服务点出发,采用启发式规则进行哈希二次探测再散列,启发式规则如下:

①向前进行探测,计算前一服务点的语义相似度,若低于设定阈值(表示该服务不能满足请求)则执行步骤②,否则执行步骤③;

②继续向前进行哈希二次探测再散列,保留最近计算的两个服务点的位置,当发现高于设定阈值的服务点时停止探测,按由近及远的顺序将保留的两个服务点分别设为第一和第二服务点;

③向后进行哈希二次探测再散列,当发现低于设定阈值的服务点时停止探测,则当前计算的服务点即为第二服务点。

3)删除掉第二服务点后的服务。

3 实例分析

为更好地说明索引的建立和 Web 服务的筛选,通过以下

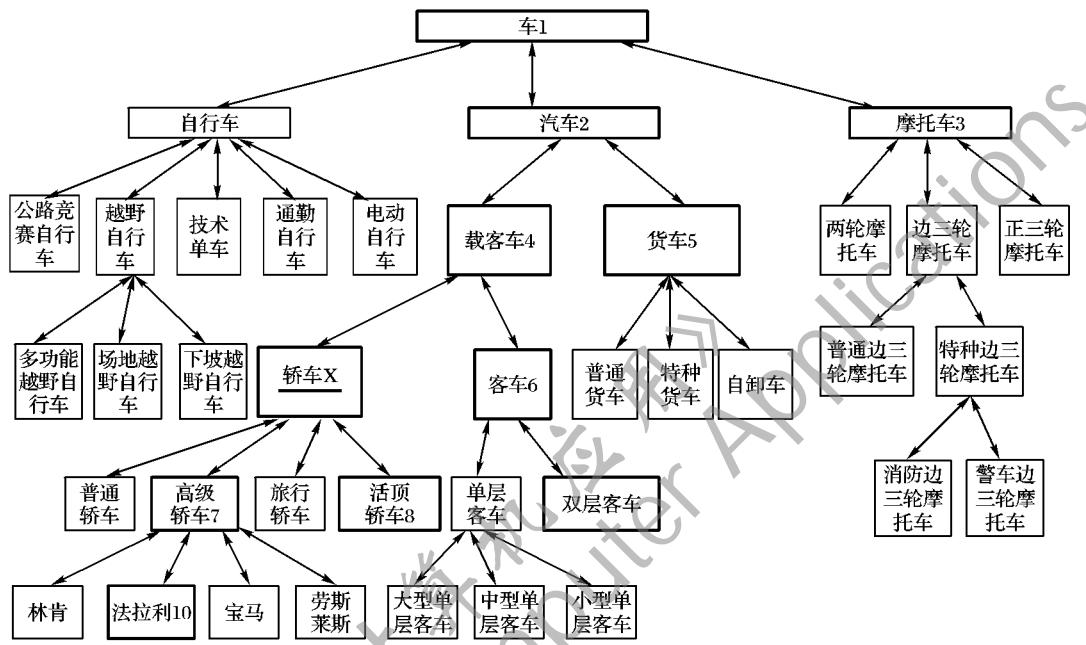


图 3 本体片段

表 1 匹配概念预处理表

Web 服务 ID 号	最近公共节点深度	边数和
7	4	1
8	4	1
10	4	2
4	3	1
6	3	2
9	3	3
2	2	2
5	2	3
1	1	3
3	1	4

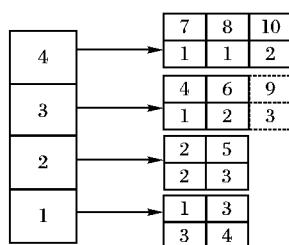


图 4 索引表图示

下面根据索引表和启发式筛选机制,为满足服务请求的输入概念 X,对 Web 服务进行 L/O 匹配,匹配过程如下。

1) 对应图 4 进行第一次折半查找,定位到服务 ID 为 9 的概念,代入式(1),计算其与 X 的语义相似度为 0.5196,低于

实例加以阐述。某网站为更加准确地向用户提供关于车辆信息的查询服务,采用上述所给方法建立索引,并对 Web 服务进行筛选。该网站使用的本体片段如图 3 所示,其中加粗框并带有下划线表示的是请求服务需要的概念,加粗框表示的是发布服务的概念,框内的数字表示 Web 服务 ID 号,其中:1~10 表示 Web 服务 ID 号为 1~10 的发布服务的输入概念,X 表示服务请求的输入概念。将本体片段中存在的服务概念可以遍历成如表 1 所示的匹配概念预处理表,然后根据索引表建立规则,构建如图 4 所示的索引表,这里将语义相似度的阈值设定为 0.6。

设定阈值,继续向上折半查找。

2) 找到服务 ID 为 10 的概念,代入式(1),计算其与 X 的相似度为 0.6595,高于设定阈值,则确定置信区在 ID 为 9 的概念附近。

3) 从概念 9 出发,首先向前探测,根据哈希二次探测再散列机理,计算概念 6 与 X 的相似度,得 0.6927,高于设定阈值。

4) 依照启发式筛选机制,向后进行探测,根据哈希二次探测再散列机理,计算概念 2 与 X 的相似度,得 0.5556,低于设定阈值,停止探测,确定服务 2 为第二服务点,其后服务不再进行匹配计算。

依照以往的 L/O 匹配方法^{[7]873},需要依次计算 10 个服务的概念与 X 的语义相似度,而依据上述的匹配优化方法只需要对 7 个发布的服务概念进行相似度计算,一定程度上减少了计算量。本文研究面向的是海量服务情况,当 Web 服务数量庞大,使用上述所给方法能够大幅减少计算量,提高系统响应速度,带来更好的用户体验。

4 结语

面向用户的请求在海量 Web 服务集中快速而准确地找到合适的 Web 服务具有挑战性,以往的 Web 服务匹配方法已经不再适应用户对系统响应速度和效率的要求。本文对 Web 服务匹配中计算量较大的 L/O 匹配环节加以优化,通过建立

(下转第 701 页)

后,分析了核函数在支持向量机的作用,提出了一种基于类别相关度量的词序列核,并将其应用于垃圾邮件过滤,取得了较好的过滤效果,然而该核也存在时间复杂度较高的问题,为此如何提高其计算效率将是下一步研究方向。同时垃圾邮件的动态特性决定了垃圾邮件过滤是一项长期而艰巨的任务,如何在保证正常邮件正常接收的情况下,尽量提高垃圾邮件识别一直是邮件过滤研究的目标,考虑将邮件更多的信息用于垃圾邮件过滤以及支持向量机在线过滤模型研究都是非常有益的研究方向。

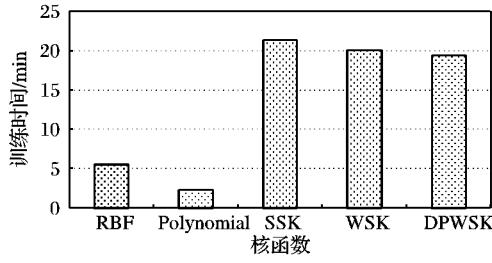


图4 SVM在各种核函数下的训练时间比较

参考文献:

- [1] CARRERAS X, MARQUEZ L. Boosting trees for anti-spam E-mail filtering [C]// Proceedings of the 4th International Conference on Recent Advances in Natural Language Processing. Tzgov Chark, Bulgaria: [s. n.], 2001: 58–64.
- [2] ANDROUTSOPoulos I, PALIouras G, KARKALETSIS V, et al. Learning to filter spam E-mail: A comparison of a naive Bayesian and a memory-based approach [C]// Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases. Lyon, France: [s. n.], 2000: 1–13.
- [3] ANDROUTSOPoulos I, KOUTSIAS J, CHANDRINOS K, et al. An evaluation of naïve Bayesian anti-spam filtering [C]// Proceedings of the 11th European Conference on Machine Learning. Barcelona, Spain: [s. n.], 2000: 9–17.
- [4] FU C, HUANG X, SCHUURMANS D, et al. Text classification in Asian languages without word segmentation [C]// Proceedings of the 6th International Workshop on Information Retrieval with Asian Languages. Sapporo, Japan: [s. n.], 2003: 44–48.
- [5] AMAYRI O, BOUGUILA N. A study of spam filtering using support vector machines [J]. Artificial Intelligence Review, 2010, 34(1): 73–108.
- [6] DRUCKER H, VAPNIK V, WU D. Support vector machines for spam categorization [J]. IEEE Transactions on Neural Networks, 1999, 10(5): 1048–1054.
- [7] TANTUG A, ERYIGIT G. Performance analysis of naïve Bayes classification, support vector machines and neural networks for spam categorization [M]// Applied Soft Computing Technologies: The Challenge of Complexity. Berlin: Springer, 2006: 495–504.
- [8] KOLCZ A, ALSPECTOR J. SVM-based filtering of E-mail spam with content specific misclassification costs [C]// Proceedings of the 2001 Workshop on Text Mining. California: [s. n.], 2001: 123–130.
- [9] 熊忠阳, 杜圣东, 张玉芳. 一种改进的支持向量机邮件过滤器 [J]. 计算机科学, 2007, 34(9): 90–92.
- [10] SCULLEY D, WACHMAN G. Relaxed online SVMs for spam filtering [C]// Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2007: 415–422.
- [11] 王祖辉, 姜维. 基于支持向量机的垃圾邮件过滤方法 [J]. 计算机工程, 2007, 35(13): 188–189.
- [12] LODHI H, SAUNDERS C, SHAWE-TAYLOR J, et al. Text classification using string kernels [J]. Journal of Machine Learning Research, 2002, 2(1): 419–444.
- [13] CANCEDEDA N, GAUSSIER E, GOUTTE C, et al. Word-sequence kernels [J]. Journal of Machine Learning Research, 2003, 3(6): 1059–1082.
- [14] GORDON V, THOMAS R. TREC 2007 public corpus [EB/OL]. [2010-09-20]. <http://plg1.cs.uwaterloo.ca/cgi-bin/cgi-wrap/gvcoramac/foo07>.
- [15] CORTES C, VAPNIK V. Support vector networks [J]. Machine Learning, 1995, 20(1): 273–329.

(上接第679页)

高效索引和给出一种基于哈希二次探测再散列的启发式筛选机制,过滤掉无关Web服务,减少了语义计算量,提高了系统响应速度。但运用本文所给方法带来的匹配效率的提升是以牺牲部分查全率为前提,所以给出的匹配优化方法比较适合Web服务集较大的情况。本文的局限主要是只能处理单概念的I/O匹配,为更好地适应实际需求,未来的工作重点是对多参数的I/O匹配优化开展研究。

参考文献:

- [1] 刘传昌, 陈俊亮. 目标Web服务描述本体和服务发现模型[J]. 计算机工程, 2007, 33(18): 187–189.
- [2] PAOLUCCI M, KAWAMURA T, PAYNE T R, et al. Semantic matching of Web service capabilities [C]// ISWC: Proceedings of the 1st International Semantic Web Conferences, LNCS 2342. Berlin: Springer-Verlag, 2002: 333–347.
- [3] 李婧, 陈旺虎, 冯百明. 提高Web服务匹配效率的服务过滤方法 [J]. 计算机应用, 2009, 29(11): 3139–3142.
- [4] 裴韶亮. 语义Web服务匹配框架模型研究与设计[J]. 计算机工程与设计, 2010, 31(2): 410–413.
- [5] LEWIS W, GARRETT M, BARKER J. Measuring conceptual distance using WordNet: The design of a metric of measuring the se-

mantic similarity of word substitution [R]. Tucson: University of Arizona, 2002.

- [6] RESNIK P. Using information content to evaluate semantic similarity in a taxonomy [C]// LJCIA'95: Proceedings of the 14th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers Inc, 1995: 448–453.
- [7] LIN Y H, BANDER 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.
- [8] 裴江南, 仲秋雁, 崔彦. 服务匹配模型中综合语义匹配方法研究 [J]. 大连理工大学学报, 2007, 47(6): 914–919.
- [9] 葛继科, 邱玉辉. 一种基于本体概念语义距离的服务相似度量方法 [J]. 计算机科学, 2009, 36(6): 182–184.
- [10] RADA R, MILI H, BIEKNELL E, et al. Development and application of a metric on semantic nets [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1989, 19(1): 17–30.
- [11] 朱江, 张国宁. 利用本体云影模型的本体进化进程 [J]. 计算机与数字工程, 2010, 38(8): 13–17.
- [12] 严蔚敏, 吴伟民. 数据结构: C语言版 [M]. 北京: 清华大学出版社, 1997: 253–262.