

KMP 算法在 Web 服务语义标注中的应用

曾 诚^{1,2}, 李 兵², 何克清²

(1 湖北大学 数学与计算机科学学院, 湖北 武汉 430062; 2 武汉大学 软件工程国家重点实验室, 湖北 武汉 430072)

摘 要: 对 Web 服务进行语义标注是指基于领域本体对 Web 服务进行语义描述,使其能够被计算机所理解,因此标注算法的优劣直接会影响到 Web 服务标注的准确率。目前人工标注方法繁琐,而全自动标注方法又无法保证其准确性,因此研究半自动语义标注是很有必要的,提出了一种概念相似度标注算法,重点介绍了 KMP 算法在其中的应用,大大提高了 Web 服务语义标注的效率和准确率。

关键词: Web 服务;语义标注;KMP 算法

中图分类号: TP311.5

文献标识码: A

文章编号: 1000-7180(2010)08-0001-03

Application of KMP Algorithm in Web Service Annotation

ZENG Cheng^{1, 2}, LI Bing², HE Ke-qing²

(1 College of Mathematics & Computer Science, Hubei University, Wuhan 430062, China;

2 State Key Laboratory of Software Engineering, Wuhan University, Wuhan 430072, China)

Abstract: Semantic annotation means to add semantic information to Web service based on Domain Ontology in order to make Web service be understood by the computer. So the annotation algorithm will directly affect the accuracy of Web service annotation. At present manual annotation methods are tedious, while automatic annotation methods can not guarantee its accuracy, so the research on semi-automatic semantic annotation is necessary. This paper presents a conceptual similarity algorithm, focusing on the application of KMP algorithm, which has greatly enhanced the efficiency and accuracy of annotation.

Key words: Web service; semantic annotation; KMP algorithm

1 引言

对 Web 服务进行语义标注是指基于本体对 Web 服务进行语义描述。由于 WSDL 描述文件中不帶有任何语义信息,因此,有必要对 WSDL 描述文件进行语义标注,让服务描述信息可被计算机所理解,从而提高服务发现的准确率和效率。语义标注算法的输入是待标注的 Web 服务和本体 Ontology(一般是指领域本体),输出是经过语义标注的 Web 服务描述文档(即带有语义信息的 Web 服务)。一般而

言,标注的内容主要是针对 Web 服务的功能接口信息,如 Operation, Inputs 和 Outputs。

目前关于语义标注的研究中,文献[1]构建的 METEOR-S 标注构架,提供了一种在当前工业标准中加入语义的模式,其基本原理是利用相应的本体来半自动化标注 WSDL 文件。文献[2-3]也提出相应的语义标注方法:前者提出的 Web 服务的语义标注方法是通过分析 WSDL 文件中数据类型的 XML schema 信息,生成了具有类和类之间关系的信息结构。通过对这些信息的标注,将其表示为领域本体中

收稿日期: 2010-05-08;修回日期: 2010-06-08

基金项目: 国家“九七三”计划项目(2007CB310801);国家自然科学基金项目(60873083,60803025,60970017,60903034);湖北省青年杰出人才基金项目(2008CDB351);湖北省自然科学基金项目(2008ABA358,2008ABA379);教育部博士点基金项目(20070486065,20090141120020);中央高校基本科研业务费专项资金(武汉大学)(6082008);湖北省教育厅科学技术研究计划重点项目(D20091012)

的概念和关系,从而能被计算机理解.后者提出了一种基于概念频率的 Web 服务语义标注方法.它通过统计 WSDL 文件中概念属性出现的频率,对不同属性区分概念的能力进行加权,从而计算两个概念之间的匹配度.

语义标注 Web 服务的方法分为人工标注和(半)自动标注.手工标注方法需要大量的人工参与,不仅非常繁琐,而且还可能由于误操作而导致标注错误,因此研究(半)自动语义标注是很有必要的.本文利用概念在本体中的相似度匹配来过滤掉大量的不相关的概念,再应用 KMP 算法来进一步提高标注的准确率,从而实现高效率的语义标注过程.

2 Web 服务语义标注的基本思想

本文将研究如何采用半自动的方式对 Web 服务进行语义标注,基本思想如图 1 所示.

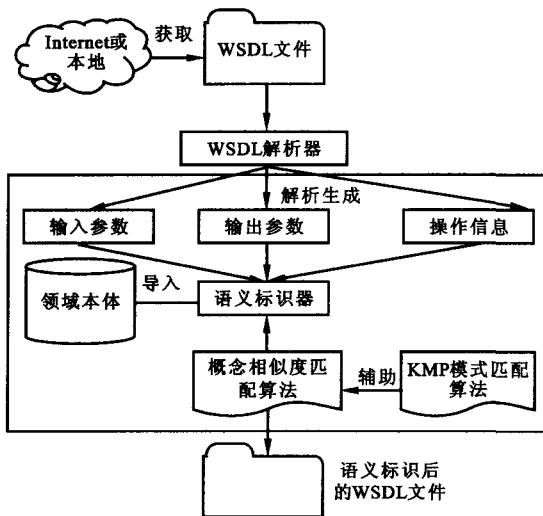


图 1 Web 服务语义标注基本思想

首先导入 Web 服务的 WSDL 文件,利用 WSDL 解析器读出 Web 服务的功能接口信息,如服务的输入、输出或操作等;然后从领域本体库中选择该服务所面向的领域,导入相应的领域本体,利用领域本体概念对接口信息需要进行半自动化语义标注,即根据 Web 服务的实际输入输出(IO)信息,利用概念相似度匹配算法从领域本体概念中选出相似度较高的概念作为候选标注对象供用户选择,算法核心是利用 Paolucci 等人在文献[4]中描述的匹配等级 {Exact, Plug-in, Subsume, Fail} (匹配度级别依次减弱)来对候选概念进行排序.

在匹配过程中,可能存在部分 Web 服务功能接口信息无法与领域本体概念对应,此时可以利用 KMP 算法对 Web 服务的这些功能接口信息进行子

串模式匹配,以寻求子串概念的匹配级别,然后利用子串匹配后的结果集在领域本体中再去寻找相关候选概念,这样会大大增强标注的有效性和准确性.下面详细介绍一下语义标注过程.

3 Web 服务语义标注的基本步骤

3.1 WSDL 文件解析

首先获取 Web 服务的 WSDL 文件地址,无论是本地的还是互联网上的 Web 服务,只要是真实的能够访问成功的 WSDL 地址均可.然后利用 IBM 研发的 WSDL 解析器 WSDL4J(它提供了标准的 WSDL 解析接口),对即将注册的 Web 服务的 WSDL 文件进行解析.通过此工具可以提取 Web 服务的功能接口信息,如服务的名称,服务的输入、输出或操作等.

3.2 概念相似度匹配算法

完成 WSDL 文件的解析后,需要将解析得到的功能接口信息映射到领域本体中的相应概念上.在本文中,我们将领域本体划分为领域操作本体和领域实体本体.前者包含了领域本体中的动词概念,往往用于标注服务的操作;后者包含了领域本体中的名词概念,往往用于标注服务的输入和输出信息.具体实现过程中我们应用概念相似度匹配算法来对 Web 服务进行语义标注,基本思想是:首先将 Web 服务的实际输入和输出等功能信息和领域本体概念进行逐一匹配,计算出它们之间的匹配度级别.然后将所有符合匹配级别的领域概念搜索出来并按照匹配度的高低排序,以供用户选择.该算法的思路是借鉴 Paolucci 等人^[4]的算法思想,利用领域概念之间包含和被包含的关系来体现匹配级别高低的差异.具体算法如下(以输入为例):

算法:概念相似度匹配算法

输入: Web 服务的输入参数(简称 WS_{input});

领域本体的概念 ($OntoConcept_j$)

输出: 匹配度级别;

$degreeOfMatch(WS_{input}, OntoConcept)$

```

{
    if  $OntoConcept_j = WS_{input}$ 
        then return exact
    if  $WS_{input}$  subclassOf  $OntoConcept_j$ 
        then return exact
    if  $OntoConcept_j$  subsumes  $WS_{input}$ 
        then return plugIn
    if  $WS_{input}$  subsumes  $OntoConcept_j$ 
        then return subsumes
    return fail
}
  
```

Web服务的输出参数和操作信息的匹配级别算法和输入算法类似,这里就不再重复描述。

从上述标注过程可以看出,这是一种典型的半自动语义标注过程,因为一方面我们的算法在标注过程中对领域本体概念进行了筛选,缩小了用户的寻找范围,避免了用户从海量本体概念中去费时费力地寻找标注对象;另一方面也给了用户选择的权利,而并非全自动地由算法一次性标注完毕。虽然我们的算法给出了匹配级别的高低,但选择权掌握在用户手中,用户完全有权力在我们所提供的候选概念列表中去选择他最满意的本体概念去标注。

概念相似度匹配算法是半自动语义标注Web服务的核心算法,但正如前面所说,也有可能出现Web服务功能接口信息和领域本体概念无法匹配、进而无法标注的问题,下一节的KMP算法可以用来解决这个问题。

3.3 KMP模式匹配算法

因为本体往往是由领域专家构建的,他们所描述的领域本体概念一般都是描述本领域内的标准概念,而描述Web服务功能接口信息的字符串则可能存在一定程度的主观性和随意性,因此我们的匹配算法也要能够保证这样的功能接口信息能够尽可能地匹配到,因此我们提出应用KMP算法^[5]作为补充,下面重点描述KMP算法在标注过程中的应用。我们可以把领域本体的概念作为主串,而把Web服务功能接口信息描述串作为模式串,应用KMP算法进行匹配。KMP算法是一种长字符串中匹配短字符串的无回溯算法,其基本思想为:设主串为 $S_1 S_2 S_3 \cdots S_n$,模式串为 $P_1 P_2 P_3 \cdots P_m$,主串从左往右扫描对模式串逐一匹配,如果主串扫描到第 i 个字符时与模式串的第 j 个字符匹配失效,则 $S_{i-j+1} S_{i-j+2} \cdots S_{i-1} = P_1 P_2 \cdots P_{j-1}$, $S_i \neq P_j$;此时 i 不用回溯保持不变,只需要模式串的当前扫描位置回退 $j-1-k$ 步, k 即为在这里需要定义一个模式串的next函数的函数值:

$$\text{next}(j) = \begin{cases} 0, & j = 1 \text{ 时} \\ \max\{k \mid 0 < k < j, P_1 \cdots P_{k-1} = P_{j-k+1} \cdots P_{j-1}\}, & \text{该集合不空时} \\ 1, & \text{其他情况} \end{cases}$$

这个函数表示当模式串匹配到第 j 个字符失败,此时已匹配到的模式串前缀和后缀能匹配的最大长度,它描述的是模式串自身的性质。因为

$S_{i-j+1} S_{i-j+2} \cdots S_{i-1} = P_1 P_2 \cdots P_{j-1}$ 并且 $P_1 \cdots P_{k-1} = P_{j-k+1} \cdots P_{j-1}$, 所以 $P_1 \cdots P_{k-1} = S_{i-k+1} S_{i-k+2} \cdots S_{i-1}$, 所以已匹配的模式串的前 k 个字符是与主串扫描位置为 i 时的前 k 个字符是匹配的,所以模式串没有必要又移动到串首,只需回退 $j-1-k$ 步即可。KMP算法本身的实现代码见参考文献[5],因篇幅原因不再赘述。

KMP算法主要解决的是大规模长串来匹配短串的问题,它的时间复杂度为 $O(n)$,效率是比较高的。而且KMP算法要求模式串是主串的子串就能匹配,并不要求模式串和主串完全相等。当用户输入关键字错输或者少输某些字母时,KMP算法也能找到能与模式串相匹配的本体概念主串,从而提高了匹配结果的准确率,同时也提高了整个过程的标注效率。整个匹配过程类似于搜索引擎的智能提示功能,当在搜索引擎中输入搜索关键词时,包含该关键词的相关搜索句子都提示出来供用户选择。

KMP算法匹配的是模式串完全是主串子串的情况,但用户可能输入关键字的时候多输了一个字符,这样导致得不到匹配结果的情况。因此,在这里可以对KMP算法做些改进。首先由用户自定义一个相似度 α ,当采用KMP算法进行匹配时,遇到匹配失配的情况模式串不是立刻根据当前的next值进行回退,而是先求出模式串已匹配的字符占模式串所有字符的比例 $r = \text{matchedChars}/\text{allChars}$ 。如果 $r \geq \alpha$,则主串符合条件返回,无需完全匹配模式串。这样匹配就不至于因为用户的错误多输字符而导致没有结果,匹配结果的遗漏容易导致失去最佳的标注概念,因此这样的改进可以进一步提高标注过程的准确率和效率。

经过上述改进算法的实施后,均会为Web服务的每一个输入输出产生一个候选标注概念列表。列表中给出了候选的概念以及匹配的级别,用户可以选择满足Web服务接口信息的本体概念进行语义标注,从而实现了本体对Web服务的标注。

4 结束语

本文针对Web服务标注过程中出现的问题,提出了一种半自动化语义标注技术,通过“WSDL解析—概念相似度匹配—KMP模式匹配”等一系列步骤实现了领域本体概念对Web服务的语义标注。其中KMP算法的应用使得整个Web服务标注过程的准确率和效率较传统语义标注技术有了明显的提高。

(下转第8页)

扩展性和可维护性;采用基于全局 ID 的服务注册和动态调用机制,增强了 SOA 系统的敏捷性和松耦合性.同时,该方案的实现采用了业界主流的 J2EE

技术和开源成熟的 Kerberos 协议,对企业异构且复杂的遗留环境具有较强的通用性和适应能力,在企业工程时间中展现了良好的经济和实用价值.

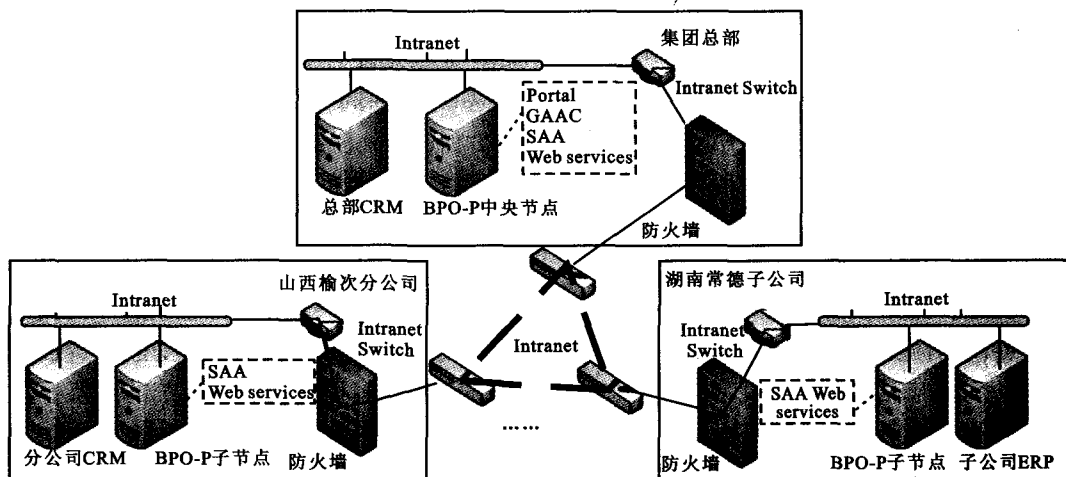


图 4 工程应用系统物理架构

参考文献:

- [1] Komoda N. Service oriented architecture (SOA) in industrial systems[C]// 2006 IEEE International Conference. Taiwan: Taipei, 2006: 1-5.
- [2] Kanneganti R, Chodavarapu P. SOA security[M]. Greenwich: Manning Publications Co, 2008.
- [3] 邵晓燕, 邵贝恩. 基于 SOA 的企业应用跨安全域访问控制[J]. 清华大学学报, 2009, 49(7): 1050-1053.
- [4] Bill Jones, Kevin Minder. Planning for service management within a service-oriented computing infrastructure [EB/OL]. [2009-11-19]. <http://java.sys-con.com/read/45783html>
- [5] 呼斯乐, 邵贝恩. 集团型企业异构系统的中间件集成方案[J]. 微计算机信息, 2009, 25(9): 3-5.
- [6] Sheth A, Verma K, Miller J, et al. Enhancing web service descriptions using WSDL-S[J]. Lecture Notes in Computer Science, 2005(3387): 55-68.
- [7] Li H, Du X, Tian X. A WSMO-based semantic web services discovery framework in heterogeneous ontologies environment[J]. Lecture Notes in Computer Science, 2007 (4798): 617-622.
- [8] 邵晓燕. 企业服务资源共享中跨安全域访问控制研究[D]. 北京: 清华大学自动化系, 2008.

作者简介:

吴靖 男, (1985-). 研究方向为系统集成、企业计算等.
 彭启瑞 男, (1986-). 研究方向为网络安全.
 李清 男, (1970-), 博士, 副教授. 研究方向为系统建模、系统评价体系、航空飞行器控制等.
 刘永青 男, (1979-), 工程师. 研究方向企业信息化应用.

(上接第 3 页)

参考文献:

- [1] Verma K, Sivashanmugam K, Sheth A. Meteor-S WSDI: a scalable infrastructure of registries for semantic publication and discovery of web services[J]. Journal of Information Technology and Management, 2005, 6(1): 17-39.
- [2] 许斌, 李涓子, 王克宏. Web 服务语义标注方法[J]. 清华大学学报: 自然科学版, 2006, 46(10): 1784-1787.
- [3] 张大陆, 吕韬. 基于概念频率的 Web 服务语义标注[J]. 同济大学学报: 自然科学版, 2008, 36(1): 1395-1400.
- [4] Paolucci M. Semantic matching of Web service capabilities [M]. Berlin: Springer Verlag, 2002.
- [5] 严蔚敏, 吴伟民. 数据结构[M]. 北京: 清华大学出版社, 2002.

作者简介:

曾诚 男, (1976-), 博士, 副教授. 研究方向为网络化软件工程.
 李兵 男, (1969-), 教授, 博士生导师. 研究方向为网络化软件工程.
 何克清 男, (1947-), 教授, 博士生导师. 研究方向为网络化软件工程.