第34卷 第24期 计 算 机 工 程 2008年12月
Vol.34 No.24 Computer Engineering December 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)24—0186—04 文献标识码：A 中图分类号：TP18

基于AIS的服务匹配模型

王 磊，刘戈峰，李园园
(西安理工大学计算机科学与工程学院，西安 710048)

摘 要：针对在当前服务发现系统的服务匹配过程中存在的系统自学习能力差的缺点，借鉴人工免疫系统中细胞变异、演化和二次响应等基本原理，通过模拟抗体-抗原识别机制来解决实际匹配问题，提出一种基于人工免疫系统的服务匹配模型。理论分析与仿真实验结果表明，该模型不仅在查全率、匹配速度等方面较传统服务发现系统有一定的提高，而且实现了由已知服务请求推测出相似服务请求，进而搜寻到最佳匹配服务的功能，提高了服务匹配过程中系统的适应学习、记忆和动态演化的能力。

关键词：人工免疫系统；服务发现；自学习；二次响应

AIS-based Service Match Model

WANG Lei, LIU Ge-feng, LI Yuan-yuan
(School of Computer Science & Engineering, Xi'an University of Technology, Xi'an 710048)

【Abstract】 With regarded to the problem that the current system's self-learning ability usually appears weak during service discovery, a novel Artificial Immune System(AIS) based service match model is proposed, which is based on the reference to the principles of cell's mutation, evaluation and the secondary response abilities, as well as simulation on the antibody-antigen identification mechanism. Theoretical analysis and simulations show that the model can increase the recall ratio and match speed, and realize the function that similar services can be obtained by known services. Furthermore, this service match model is able to improve a system's self-learning, memory and dynamic evaluation capabilities.

【Key words】 Artificial Immune System(AIS); service discovery; self-learning; secondary response

1 概述

随着电子商务的大规模应用，Web服务逐渐成为企业级开发和应用的热点技术之一，而服务发现技术作为服务请求者请求服务的一种手段也受到了广泛的关注。其中，服务匹配是服务发现系统架构的一个重要部分，是服务发现系统研究的一个重要内容。早期的服务匹配技术主要有基于关键字的匹配、基于框架的匹配和演绎检索技术等。笔者熟悉的UDDI就是采用了“基于框架”的改进方法以及“演绎检索”匹配方式来进行服务匹配的，但由于该匹配方式在匹配过程中需要进行复杂的逻辑表示和逻辑推理，从而在一定程度上限制了该系统的应用推广。目前服务匹配技术所关注的主要方向是系统对语义的理解问题，在此基础上提出了大量基于语义网络和本体论的Web服务匹配技术，如DAML以及后续版本OWL-S等，这在很大程度上解决了语义识别问题。但上述这些方法很少涉及系统的自学习问题，即针对同一个用户请求的二次查询只不过是前一次或几次查询的简单重复时，系统不能表现出查询速度和质量等方面的提高；同时，目前的系统也不能根据一次服务匹配过程，再通过相应的学习机制，推测、联想出类似服务请求及其所请求的服务[1]。

另一方面，随着人们对智能信息处理模型研究的不断深入，一些新颖的、智能化程度相对较高的系统不断涌现，并已经表现出潜在的应用价值，这其中就包括人工免疫系统(Artificial Immune System, AIS)[2]。AIS是在生物免疫系统研究基础上发展起来的一门新兴交叉科学，目前已被广泛应用到计算机安全、模式识别、机器学习、调度控制、故障诊断、数据挖掘、联想记忆、优化计算等许多领域[3-4]。作为一种新的智能计算方法，AIS具有强大的自组织、自学习和自适应等能力，这些特点可以弥补当前服务匹配过程在这方面的不足。有鉴于此，AIS有可能是解决服务匹配中自学习问题的一个好的选择。

基于以上考虑，本文依据AIS基本原理，提出了一种新的服务匹配模型，以图探索完善系统在服务匹配过程中的学习与判断的能力。理论分析和仿真实验表明，该模型在一定程度上解决了传统的服务匹配过程中自学习方面能力明显不足的问题；运用该模型，系统不但能匹配服务请求，而且还能针对已知服务请求，生成与其相类似的服务请求与相应的服务，供请求者参考与筛选。

2 基于AIS系统的服务匹配技术

一般而言，服务匹配过程具有2方面的特性：(1)针对相同的服务请求，不同用户的请求描述不会完全相同；(2)针对相同的服务请求，不同的请求描述之间具有很大的相似性。本文正是从服务发现系统的这2个一般特点出发，探讨在性能上更优的服务发现模型的存在性与可行性。具体而言，就是研究能否将与某次服务请求匹配的服务分为2种类型，其中匹配程度较高的一类服务作为此次服务请求的最终匹配服务；而对匹配程度较低的一类服务，则按照一定的规则对其进行变异、检验，进而生成与此次服务请求类似的虚拟服务，以便当系统再次检测到类似服务请求时作出快速响应，提高系统整体的工作效率。

基金项目：国家自然科学基金资助项目(60603026)
作者简介：王 磊(1972－)，男，教授、博士，主研方向：人工免疫，智能计算，普适计算；刘戈峰、李园园，硕士研究生
收稿日期：2008-04-15 **E-mail：**leiwang@xaut.edu.cn

2.1 基本概念

假设问题讨论范围为集合 D，D 为长度为 $l(l\in N)$的二进制串集合，则

$$D=\{0,1\}^l \tag{1}$$

定义 1 抗原

在一般的 AIS 系统中，抗原多指问题的描述或约束条件。与此类似，本文所指的抗原即为用户请求经过分词工具分词后所得到的关键词的集合。假设抗原集合用 Ag 表示，则有

$$Ag=\{ag \mid ag\in D\} \tag{2}$$

定义 2 抗体

与通用 AIS 模型中的定义类似，本文所指的抗体也对应于问题的求解。在服务发现系统中，WSDL 文件是描述 Web 服务的标准的 XML 格式，它用一种和具体语言无关的抽象方式定义了给定 Web 服务收发的有关操作和消息。由此可见，Web 服务发现的实质就是准确、高效地找到 Web 服务所对应的 WSDL 文件，所以描述 WSDL 文件的相关信息，则成为潜在抗体的组成元素。本文中，抗体元素用符号 ab 来表示；抗体集合用 Ab 来表示，则

$$Ab=\{ab\} \tag{3}$$

定义 3 亲和力

抗体与抗原之间的亲和力定义为抗体抗原之间的相似程度，即亲和力越高，用户请求的服务与系统提供服务的一致性越高。常见的亲和力计算方法有距离、结合强度、匹配度等几种方式，而每一种方式中又包含多种不同的方法。根据服务发现中请求匹配的特点，本文采用 r-连续位匹配规则来衡量亲和力的大小。假设抗原 ag 与抗体 ab 之间的最大连续位为 num，则规定当连续位超过 $r(r\in N\wedge r>0)$时，认为 ag 与 ab 匹配；反之，则不匹配。r-连续位匹配规则满足如下关系：

$$f(ab,ag)=\begin{cases}1 & num\geq r\\ 0 & num<r\end{cases} \tag{4}$$

其中，$f(ab,ag)=1$ 表示抗原与抗体匹配；$f(ab,ag)=0$ 表示两者不匹配。

定义 4 细胞变异

由定义 1 与定义 2 可知，抗原和抗体是由二进制字符串来表示的，因此细胞的变异将围绕表示抗体的二进制字符串的某一位的翻转进行。若变异的某位起初是 0，则变异后该位翻转为 1；反之，则翻转为 0，如图 1 所示。

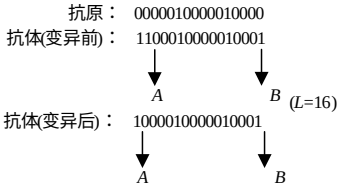

图 1 细胞变异操作示意图

根据需要，细胞进行变异时，保留细胞中与抗原相同的部分，对不同的部分选取一位进行变异。变异位置的确定方法(如图 1 所示)为：若抗原编码的长度为 Str，A 为抗原与抗体最大连续位的左端，B 为连续位的右端，最大连续位为 r 位(即二进制串中从 A 到 B 的部分)，则有

$$\begin{aligned}&f(ab)=\begin{cases}A-1 & A\geq 1\\ B+1 & A=1\end{cases}\\ &1\leq A\leq Str-r\\ &r\leq B\leq Str\\ &r=B-A+1\end{aligned} \tag{5}$$

由上式可知，若 A=3(如图 1 所示)，因为 A>1，所以细胞的变异为 A-1=2。

2.2 系统模型

如前文所述，当前服务发现系统中服务匹配过程的自学习能力还普遍比较薄弱。针对这一问题，本文结合人工免疫系统的自学习、模式识别和二次响应等机制，提出一种新的服务匹配模型，如图 2 所示，旨在提高原系统的自学习能力。

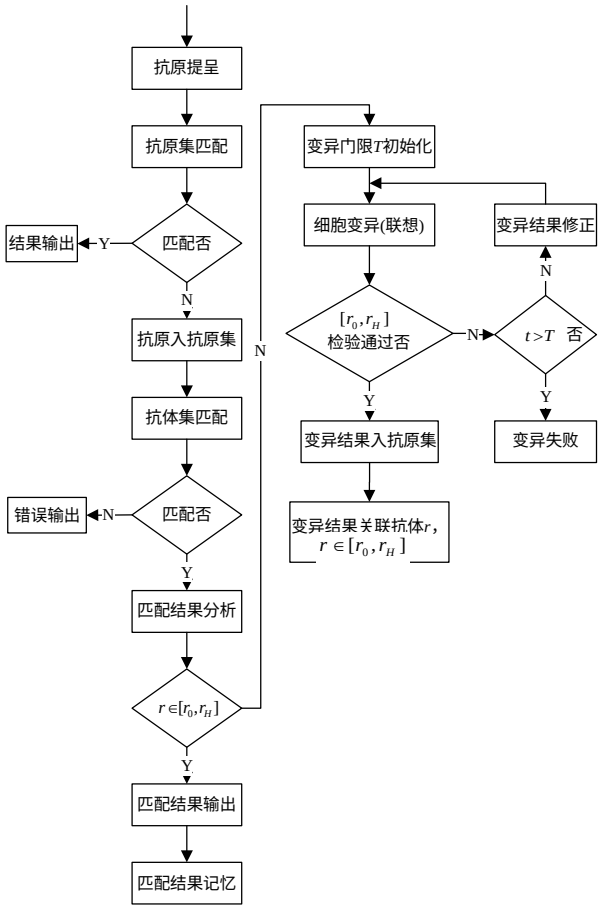

图 2 基于 AIS 的服务匹配模型结构示意图

该模型设计的基本思想是将系统服务的匹配过程进行分类处理(如图 3 所示)，从而可以使系统工作的重点放在对非匹配信息的处理上，以强化系统对不确定信息的处理能力。

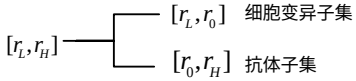

图 3 抗原初次匹配结果的分类分析

在模型中，对服务的初次匹配操作是系统进化的主要过程。其中，初次匹配是指系统从未匹配响应过的，并且其他响应过的服务请求经变异、演化[5]后，也未能联想出的服务请求信息。而抗体与抗原之间的初次匹配操作，则需经过以下过程：

(1)系统将抗原添加到抗原集中。

(2)系统从抗体集中找到所有与抗原相匹配的抗体(如图 3 所示，抗体与抗原的匹配位数属于区间$[r_L,\ r_H]$。其中，r_0是该区间上用于划分细胞变异子集和抗体子集的临界值)。

(3)系统将区间$[r_0,\ r_H]$内的抗体视为对用户请求所响应的服务，并输出结果，同时建立抗原与该区间内抗体的关联信息，即两者之间形成的一种映射关系。这样当系统对该抗原

进行再次匹配时，无须到抗体集合中进行针对所有元素的匹配操作，而只需要从抗原与抗体之间的映射关系集中找到与之相对应的抗体即可，从而匹配操作的规模大大缩小。

(4)系统对区间[r_L, r_0]内的抗体，依据定义 4 的方法进行细胞变异，并利用[r_0, r_H]内的抗体对变异结果进行检测。若某个细胞的检测结果(即平均匹配度)达到指定的阈值，则该细胞停止变异，此时将变异后的细胞添加到抗原集中，同时也将变异细胞与[r_0, r_H]内抗体之间的关联信息予以相应保存。如果细胞没有达到变异临界值，它将继续变异直至变异细胞的特性达到指定的阈值；当变异细胞的特性达到临界值但仍没有通过区间[r_0, r_H]内抗体的检验时，细胞变异失败。

通过上述过程的初次匹配操作，一方面，系统对抗原及其匹配结果建立了关联信息，从而形成对两者匹配关联过程的记忆；另一方面，通过对细胞进行变异，系统产生了一系列与此抗原相似的虚拟抗原(即一类相似关键词，这也是一种联想信息)，并添加到抗原集中，同时将这些虚拟抗原与[r_0, r_H]区间内的抗体建立了映射关系。

在初次匹配过程结束后，若系统遇到该过程已经匹配响应过的，或这类服务请求所联想出的类似信息，则会激发二次匹配过程(如图 4 所示)。所谓的二次匹配过程并不是一般意义上的第二次匹配，而是对经历过初次匹配，或在某次匹配过程中由细胞变异、联想产生的一些抗原，在其生存周期内每一次匹配过程的统称。当这类抗原进入系统，系统首先从抗原集中搜索与本次服务请求性质相同的抗原，若存在，则此次匹配过程被视为二次匹配，系统根据初次匹配中抗原与抗体的关联记忆结果，快速地找到相应抗体并输出结果；反之，系统则进行初次匹配过程，并对匹配结果进行记忆、变异、联想和演化等一系列操作。由此可见，二次匹配过程充分利用了初次匹配中细胞变异和演化的结果，通过记忆与联想等机制，一定程度上提高了系统执行的效率，从而加快了服务响应的速度。

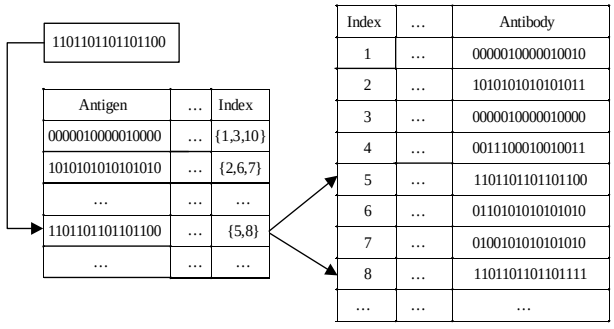

图 4　抗原二次匹配过程示意图

经过对抗原的初次匹配，系统对相同或类似的抗原，则由针对抗体的匹配转换为对抗原的匹配。因为针对抗体的匹配每次都要搜索整个抗体库来识别出所有与提呈抗原对应的抗体，而针对抗原的匹配则由于在细胞联想变异过程中已经产生了与该细胞类似的抗原，所以，对抗原的匹配只需要按照完全匹配来搜索抗原集，不需要再检索整个抗原库。另一方面，加之抗原细胞的进化功能，频繁匹配的抗原总是会优先被匹配到，而不同用户对相同服务请求的描述信息多是类似的，这使得系统的工作效率整体上有较大程度的提高。

3　性能分析

针对文中给出的服务匹配模型，着重从以下几个方面对其性能进行分析。

3.1　查全率

假设界定与实际服务请求能够匹配的服务所具有的连续相同字符的位数为$n(0<n<Str$，满足这类条件的服务数为$2^{Str-n})$，则与服务请求有$m(n<m<Str)$位连续相同字符的服务视为系统可以提供的服务(满足这类条件的服务数为 2^{Str-m})。据此，模型系统的查全率为

$$R_{\text{recall}}=\frac{2^{Str-m}}{2^{Str-n}} \tag{6}$$

以上论述的是服务请求只有一个关键字的情况，当服务请求可以分解为多个关键字时，本文提出的模型在查全率上的优势将更加明显。因此，从总体上讲，本文提出的服务匹配模型查全率一般要高于基于关键字的匹配方式。

3.2　匹配速度

由服务匹配的基本原理可知，无论是基于关键字的服务匹配，还是本文提出的基于 AIS 的服务匹配模型进行的服务匹配，待匹配的服务的数目都是影响服务匹配速度的决定性因素，因此对服务匹配速度的比较就转变为对服务匹配个数的比较。

3.2.1　初次匹配与基于关键字的的服务匹配的比较

假设抗体集中的细胞个数为 NUM，且抗体集中存在与抗原完全相匹配的抗体，位置为 $P(P\le NUM)$，则基于关键字和本文模型的初次服务匹配速度之比如下：

$$\frac{S_{\text{KEY}}}{S_{\text{FIRST}}}=\frac{P}{NUM} \tag{7}$$

由于 $P\le NUM$，因此

$$\frac{S_{\text{KEY}}}{S_{\text{FIRST}}}\le 1 \tag{8}$$

可见，初次匹配速度较基于关键字匹配速度低。这主要是因为根据 r-连续位匹配的特点，初次匹配要遍历整个抗体集，而基于关键字的匹配只要找到完全匹配的抗体，匹配立即中止。

3.2.2　二次匹配与基于关键字的服务匹配的比较

假设抗体集中细胞个数为NUM_{Ab}，抗原集中细胞个数为NUM_{Ag}，从统计学上分析有，基于关键字的服务匹配的平均匹配次数为$NUM_{Ab}/2$，而基于本文模型二次匹配时的平均匹配次数为$NUM_{Ag}/2$。由于抗原集的规模要远小于抗体集，$NUM_{Ab}>NUM_{Ag}$，故两者的速度满足如下关系：

$$\frac{S_{\text{KEY}}}{S_{\text{SECOND}}}>1 \tag{9}$$

因此，系统二次匹配的速度较基于关键字的匹配速度要快。这主要是由于二次匹配的机制与初次匹配不同，二次匹配不需要遍历抗体集，而且只需要从比规模远比抗体集小的抗原集中运用基于关键字的匹配方式匹配到抗原，然后利用初次匹配或演化生成的结果直接匹配抗体。

3.3　匹配度

匹配度关系到系统服务匹配结果的准确程度，即服务匹配度过低时，系统固然能匹配到合理的服务，但同时也匹配到了大量的垃圾信息；而服务匹配度过高时，服务匹配结果准确程度较高，但很多时候可能根本匹配不到服务，或者由于表述的原因，系统可能将一些本应匹配的服务信息过滤掉，基于关键字的服务匹配就存在这种问题。

假设界定本文匹配的服务与服务请求的连续相同位的位数为$m(0<m<Str)$，则有概率统计知识可知，满足条件的服务的数目为 2^{Str-m}，则系统的匹配度M如下：

$$M=\sum_{i=m}^{Str}\frac{2^{Str-i}-2^{Str-i+1}}{2^{Str-m}}\cdot\frac{i}{Str} \tag{10}$$

3.4 细胞变异联想

本文模型充分利用每一次初次匹配过程，一方面系统将服务匹配结果与抗原集中抗原进行关联，对匹配结果进行记忆；另一方面，利用服务匹配所得的匹配度较低(连续位在区间[r_L, r_0]内)的部分服务进行细胞变异联想，得出相似抗原及其服务，省去了经编译联想的部分抗原的初次匹配过程，充分提高了服务匹配速度。

4 仿真实验结果

根据人工免疫系统的特点，本文采用 16 位二进制字符串来表示服务请求和服务。首先在 0~216 之间随机取 10 000 个二进制串作为系统能够提供的服务(抗体集)，用户请求(抗原)将从抗体集中进行服务初次匹配、进化，得到相匹配的服务。

4.1 查全率实验

利用本文提出的模型进行查全率实验，得出对比结果如表 1 所示。

表 1 查全率对比

抗原	关键字	本文模型
0001001100010010	0001001100010010	0001001100010010 0011001100010010 1011001100010010
0000000000010100	无	1111111110010100
1110001111111111	无	1100001111111111 1010001111111111
0100101001000100	无	无

由上述实验数据可以看出，利用本文提出的服务匹配模型所做出的仿真系统不但可以匹配基于关键字匹配的服务，而且可以匹配到基于关键字不能匹配到的服务。由此可以证明，基于免疫的服务发现模型在服务(抗体)的查全率上总体上优于基于关键字的匹配方式，取得了较高的查全率。

4.2 匹配速度

针对基于关键字的服务匹配方式以及本文提出的基于人工免疫的服务发现模型的系统做仿真实验，各进行 20 次服务匹配，时间对比如图 5 所示，其中横坐标上对象“1”代表基于关键字服务匹配所花的时间，对象“2”代表基于免疫的服务初次匹配所需时间，对象“3”代表基于免疫的二次匹配所需时间。由该图可以看出，本文所提的模型初次匹配所花费的时间要比基于关键字花费的时间稍多，这主要是因为初次匹配一方面要遍历所有服务，另一方面还要对匹配结果进行分析和变异。而二次匹配与前 2 种匹配相比，匹配速度明显优于前两者，从而达到了预期的效果。

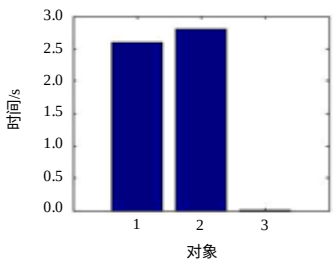

图 5 3 种匹配方式工作时间对比

4.3 匹配度

r_0与匹配度之间的关系如图 6 所示，针对r_0在 11～16 之间的每一个值，随机抽取 10 个 16 位二进制串进行抗原匹配，然后得到每一组的平均匹配度，即图 6 所示曲线。同时随着r_0的增大，系统的匹配时间明显变长，因此，能否合理地选取r_0是关系系统工作效率的一个重要因素。例如，由实验发现，本例中当r_0取值 13 时，系统的综合性能达到最佳，此时系统的平均匹配度达 87.0%。

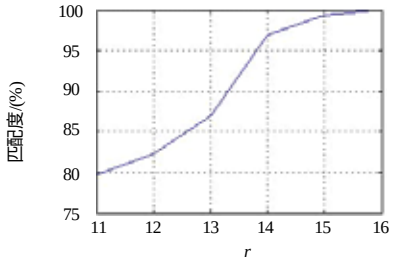

图 6 匹配度与 *r* 的关系

4.4 细胞变异联想

细胞的变异联想功能是本文提出服务匹配模型的一个重要功能，随机取 4 组抗原进行细胞变异与联想操作，其实验结果如表 2 所示。

表 2 细胞变异与联想结果测试表

抗原	变异演化类似抗原
0100100010100011	0010100010100011 1010100010100011 0000100010100011 1100100010100011 1101100010100011
1111111111110000	1010111111110000 0010111111110000 0110111111110000 0100111111110000 1100111111110000 0001111111110000 0000111111110000
1111010010100011	无
1101101101101100	0100101101101100 0001101101101100 1010101101101100 0101101101101100 1110101101101100 1100101101101100 1101101101101100 1101101101101100 0111101101101100

由上述数据可以看出，系统在初次匹配时，除了能够准确地匹配到请求的服务外，还能够通过免疫细胞的变异机制，产生请求服务的一些类似服务。这样，系统对这些新生成的服务进行匹配时，将不必再进行初次匹配，直接进入二次匹配，大大缩短了匹配时间。根据已有服务匹配的结果，推出相似服务的匹配结果，从而使系统具备了一定的学习联想功能，这正是笔者设计之初所希望看到的。

5 结束语

通过以上分析，可以看出本文给出的基于免疫机制的服务匹配模型，在查全率上较基于关键字的服务匹配方式有了一定的提高；在服务匹配速度方面，系统经过初次匹配在速度方面的微弱损失后，二次匹配过程中系统的速度有了明显的改善，进而使系统的总体性能得到提升；在细胞联想能力方面，根据理论分析与仿真验证，能够看出系统基本上可以实现一个智能化系统所特有的自学习、自演化和自适应的设计目标。当然，本系统的设计与理论分析还相对粗浅，特别是在匹配速度相对于问题规模的条件约束下进行理论计算或估计还没有涉及；另外，抗原集中细胞的进化机制尚未给出，而这在现实的自然界系统中却是普遍存在的。

(下转第 193 页)