(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 101695082 B
(45)授权公告日 2012.08.22

(21)申请号 200910235615.1

(22)申请日 2009.09.30

(73)专利权人 北京航空航天大学
地址 100191 北京市海淀区学院路 37 号

(72)发明人 李建欣 孙海龙 黄子乘 曲先洋 林伟 刘旭东

(74)专利代理机构 北京同立钧成知识产权代理有限公司 11205
代理人 刘芳

(51)Int.Cl.
H04L 29/08(2006.01)
G06F 17/30(2006.01)

(56)对比文件
CN 101266603 A,2008.09.17,
CN 101452463 A,2009.06.10,
张荣清等．网格计算环境中的安全信任协商系统．《北京航空航天大学学报》.2006,
胡志刚，胡周君．计算服务网格中基于服务聚类的元任务调度算法．《小型微型计算机系统》.2009,
胡志刚，胡周君．计算服务网格中基于服务聚类的元任务调度算法．《小型微型计算机系统》.2009,

审查员 刘俊源

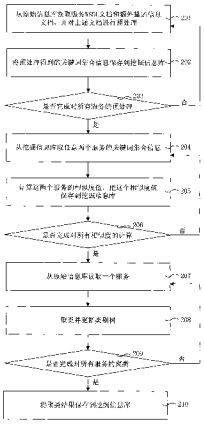
权利要求书 2 页 说明书 8 页 附图 6 页

(54)发明名称

基于关系挖掘的服务组织方法及装置

(57)摘要

本发明实施例涉及一种基于关系挖掘的服务组织方法及装置，其中，该基于关系挖掘的服务组织方法包括：对服务的描述信息进行预处理，并生成信息三元组；根据所述信息三元组计算所述服务的相似度；根据所述相似度对服务进行聚类，生成类别树，根据所述类别树完成服务定位。上述基于关系挖掘的服务组织方法及装置，利用服务相似度挖掘方法及根据上述服务相似度对服务进行聚类，有效地缩小了服务的搜索范围，提高了服务定位的效率。



CN 101695082 B

1. 一种基于关系挖掘的服务组织方法，其特征在于包括：

对服务的描述信息进行预处理，并生成信息三元组；

根据所述信息三元组计算所述服务的相似度；

根据所述相似度对服务进行聚类，生成类别树，根据所述类别树完成服务定位；

所述对服务的描述信息进行预处理，并生成信息三元组包括：

从原始信息库获取服务描述语言 WSDL 文档地址和额外描述信息文档地址；

根据所述服务描述语言 WSDL 文档地址载入所述服务描述语言 WSDL 文档，并对所述服务描述语言 WSDL 文档进行解析，获取服务名称和服务注释信息，并将所述服务名称和服务注释信息加入服务的标题信息中；

获取服务中所有方法的列表，将列表中每个方法的信息加入该服务的主题信息中，所述每个方法的信息包括方法名称、注释、输入消息名称、输出消息名称；

根据所述额外描述信息文档地址载入额外描述信息文档，并对所述额外描述信息文档进行解析，并将解析后的额外描述信息加入该服务的额外描述信息中；

所述标题信息、主题信息和额外描述信息构成了所述服务的信息三元组。

2. 根据权利要求 1 所述的基于关系挖掘的服务组织方法，其特征在于所述根据所述信息三元组计算所述服务的相似度包括：

采用如下公式计算第一服务和第二服务的相似度，

$$Sim(W_1, W_2) =$$
$$\alpha * SimSet(W_1.T, W_2.T)$$
$$+\beta * SimSet(W_1.B, W_2.B)$$
$$+\gamma * SimSet(W_1.A, W_2.A)$$

其中，W_1 表示第一服务，W_2 表示第二服务，$Sim(W_1, W_2)$ 表示第一服务和第二服务的相似度，T 表示标题信息，B 表示主题信息，A 表示额外描述信息，$SimSet(W_1.T, W_2.T)$ 表示第一服务标题信息单词集合和第二服务标题信息单词集合的词义相似度，$SimSet(W_1.B, W_2.B)$ 表示第一服务主题信息单词集合和第二服务主题信息单词集合的词义相似度，$SimSet(W_1.A, W_2.A)$ 表示第一服务额外描述信息单词集合和第二服务额外描述信息单词集合的词义相似度，α、β、γ 分别表示标题信息、主题信息和额外描述信息在第一服务和第二服务相似度中的权重。

3. 根据权利要求 1 所述的基于关系挖掘的服务组织方法，其特征在于所述根据所述信息三元组计算所述服务的相似度之前还包括：

对所述标题信息、主题信息和额外描述信息进行预处理，转换成符合标准的信息。

4. 根据权利要求 1 所述的基于关系挖掘的服务组织方法，其特征在于所述根据所述相似度对服务进行聚类，生成类别树包括：

将两服务间的相似度表示为两点间距离，根据任意两点间距离均小于预定聚类直径生成类别树。

5. 一种基于关系挖掘的服务组织装置，其特征在于包括：

生成单元，用于对服务的描述信息进行预处理，并生成信息三元组；

计算单元，用于根据所述信息三元组计算所述服务的相似度；

聚类单元，用于根据所述相似度对服务进行聚类，生成类别树，根据所述类别树完成服

务定位；

所述生成单元包括：

第一信息获取模块，用于根据服务描述语言 WSDL 文档地址载入所述服务描述语言 WSDL 文档，并对所述服务描述语言 WSDL 文档进行解析，获取服务名称和服务注释信息，并将所述服务名称加入服务的标题信息中；

第二信息获取模块，用于获取服务中所有方法的列表，将列表中每个方法的信息加入该服务的主题信息中，所述每个方法的信息包括装置名称、注释、输入消息名称、输出消息名称；

第三信息获取模块，用于根据额外描述信息文档地址载入额外描述信息文档，并对所述额外描述信息文档进行解析，并将解析后的额外描述信息加入该服务的额外描述信息中；

生成模块，用于根据所述标题信息、主题信息和额外描述信息，生成所述服务的信息三元组。

6. 根据权利要求 5 所述的基于关系挖掘的服务组织装置，其特征在于所述计算单元包括：

计算模块，用于采用如下公式计算第一服务和第二服务的相似度，

$$Sim(W_1, W_2) =$$
$$\alpha * SimSet \quad (W_1.T, W_2.T)$$
$$+\beta * SimSet \quad (W_1.B, W_2.B)$$
$$+\gamma * SimSet \quad (W_1.A, W_2.A)$$

其中，W_1 表示第一服务，W_2 表示第二服务，$Sim(W_1, W_2)$ 表示第一服务和第二服务的相似度，T 表示标题信息，B 表示主题信息，A 表示额外描述信息，$SimSet(W_1.T, W_2.T)$ 表示第一服务标题信息单词集合和第二服务标题信息单词集合的词义相似度，$SimSet(W_1.B, W_2.B)$ 表示第一服务主题信息单词集合和第二服务主题信息单词集合的词义相似度，$SimSet(W_1.A, W_2.A)$ 表示第一服务额外描述信息单词集合和第二服务额外描述信息单词集合的词义相似度，α、β、γ 分别表示标题信息、主题信息和额外描述信息在第一服务和第二服务相似度中的权重。

7. 根据权利要求 5 所述的基于关系挖掘的服务组织装置，其特征在于还包括：

转换单元，用于对生成单元生成的标题信息、主题信息和额外描述信息进行预处理，转换成符合标准的信息。

8. 根据权利要求 5 所述的基于关系挖掘的服务组织装置，其特征在于所述聚类单元包括：

聚类模块，用于将两服务间的相似度表示为两点间距离，根据任意两点间距离均小于预定聚类直径生成类别树。

基于关系挖掘的服务组织方法及装置

技术领域

[0001] 本发明实施例涉及数据挖掘技术领域，尤其涉及一种基于关系挖掘的服务组织方法及装置。

背景技术

[0002] 随着计算机网络应用的不断发展，信息系统的交互模式已由网络层系统互联向应用层服务集成迁移，网络 (Web) 技术的进一步发展和软件工程技术的进化相结合产生了面向服务的体系结构 (Service OrientedArchitectures，SOA)；随着 SOA 应用的普及，Web 服务数目与日俱增，如何从大量已有服务中高效地定位所需的目标服务是 Web 服务急需解决的一个重要问题。目前 Web 基于关系挖掘的服务组织方法主要分为两类：第一类是语法级匹配，采用基于服务名称的字符串匹配，典型的系统有统一描述、发现和集成协议 (Universal Description，Discovery and Integration，UDDI) 系统，语法级服务发现实现相对简单，但查准率较低；第二类是语义级匹配，服务描述采用本体论的方法，增强了对 Web 服务的功能、行为的语义描述，在匹配算法上，依赖于逻辑演绎和推理，虽然查准率高，但匹配效率低、实用性差。由此可见，现有的基于关系挖掘的服务组织方法在实现难度、查询效率或者查询准确率等方面还有较大局限性。

[0003] 随着计算机的广泛应用，数据大量增加，运用数据挖掘技术可以从这些数据中提取出对决策有潜在价值的知识；把传统的数据挖掘技术引入服务发现领域可以为服务发现带来新的突破，目前，将数据挖掘技术引入服务发现领域的技术有 UDDI 技术，该技术定义了 Web 服务的发布与发现的方法，所谓“Web 服务”，是指由企业发布的完成其特别商务需求的在线应用服务，其它 公司或应用软件能够通过因特网 (Internet) 来访问并使用该项在线服务，Web 服务将逐渐成为电子商务应用构建的基础体系架构，但是，当需要找出哪些企业可以提供某种服务时，快速地发现并找到答案仍然十会困难；其中一个可选的方法是使用电话和每个合作伙伴进行联系找出合适的对象，另一个解决该问题的办法是在公司的每个网站上放置一个 Web 服务的描述文件，这样，那些依靠已经注册的统一资源定位符 (URL) 来工作的网络爬虫程序能够发现并为它们建立索引。可是这种定位 Web 服务的方法完全依赖爬虫程序的能力，且缺少一种机制来保证服务描述格式的一致性，无法便捷地跟踪不断发生的变化。UDDI 提供了一种基于分布式的注册中心的方法，该注册中心维护了一个企业和企业提供的 Web 服务的全球目录，而且其中的信息描述格式是基于通用的可扩展标记语言 (XML) 格式的。UDDI 计划的核心组件是 UDDI 商业注册，它使用一个 XML 文档来描述企业及其提供的 Web 服务，UDDI 商业注册所提供的信息包含三个部分：“白页 (White Page)”包括了地址、联系方法和已知的企业标识；“黄页 (Yellow page)”包括了基于标准分类法的行业类别；“绿页 (Green Page)”则包括了关于该企业所提供的 Web 服务的技术信息，其形式可能是一些指向文件或是 URL 的指针，而这些文件或 URL 是为服务发现机制服务的，所有的 UDDI 商业注册信息存储在 UDDI 商业注册中心中。

[0004] 另外，语义级服务定位技术是将语义融合到 Web 服务技术中去，对于该技术最重

要的是要有一个强有力的描述Web服务的语言，德帕代理标记语言(DAML)组织制定的德帕代理标记语言服务(Darpa Agent Markup Language for Service，DAML-S)是一个在未来语义Web中使用Web服务的标准。DAML-S作为一个本体模型，它用基于DAML和本体推理层(OIL)的构造去定义Web服务；同时作为一种语言，DAMLS-S支持更强大的Web服务描述。此外，DAML-S还集成了过程模型(process model)，不仅可以控制Web服务的控制流和数据流，而且可以控制Web服务的初始条件和处理结果。将DAML-S加入到Web服务之后，可以把Web服务的协议层次进行改造，DAML-S应用由过程模型、服务描述(service profile)、服务基础(service grounding)三个部分组成，其中，Service Profile说明了指定的Web服务能做什么的问题，Service profile可以替代UDDI中描述的部分来完成对Web服务的表达，DAML-S支持的一些特性，比如对Web服务性能的表达等等，都不是UDDI所能达到的。另外，还有一个不同点就是UDDI“绿页”中的绑定描述（如服务端口号）等信 息，在DAML-S结构中是由grounding来完成的。process model记录Web服务的初始条件、处理结果、控制流和工作流，即process model就是说明指定的Web服务是如何工作的：它的任务是什么；它按哪些步骤来完成；各个步骤的预期子结果是什么；需要哪些输入，什么时候需要；会报告哪些输出，什么时候报告等等。DAML-S process model可以说是process-mode和工作流(workflow)语言的一个超集，集建模语言、人工智能语言和类及其关系描述语言于一身，再加上良好的语义规范，使它能够更好地表述Web服务的工作性能。同时，DAML-S同样支持用WSDL来规范和说明Web服务接口，用报文(SOAP)来传递消息。

[0005] 但发明人在实施上述技术方案的过程中发现现有技术存在一些缺陷，例如，基于服务名称的字符串匹配，查找准确度较低，逐个遍历服务，效率很低；目前大多数已经存在的服务没有语义描述信息，如何把这些已存在的服务加上语义信息工作量庞大，同时，Web服务语义描述语言过于复杂，技术实现难度大，且缺乏灵活有效的服务匹配算法，不利于其实际应用。

[0006] 发明内容

[0007] 本发明实施例提供一种基于关系挖掘的服务组织方法及装置，以提高服务定位效率。

[0008] 本发明实施例提供了一种基于关系挖掘的服务组织方法，该方法包括：

[0009] 对服务的描述信息进行预处理，并生成信息三元组；

[0010] 根据所述信息三元组计算所述服务的相似度；

[0011] 根据所述相似度对服务进行聚类，生成类别树，根据所述类别树完成服务定位；

[0012] 所述对服务的描述信息进行预处理，并生成信息三元组包括：

[0013] 从原始信息库获取服务描述语言WSDL文档地址和额外描述信息文档地址；

[0014] 根据所述服务描述语言WSDL文档地址载入所述服务描述语言WSDL文档，并对所述服务描述语言WSDL文档进行解析，获取服务名称和服务注释信息，并将所述服务名称和服务注释信息加入服务的标题信息中；

[0015] 获取服务中所有方法的列表，将列表中每个方法的信息加入该服务的主题信息中，所述每个方法的信息包括方法名称、注释、输入消息名称、输出消息名称；

[0016] 根据所述额外描述信息文档地址载入额外描述信息文档，并对所述额外描述信息文档进行解析，并将解析后的额外描述信息加入该服务的额外描述信息中；

[0017] 所述标题信息、主题信息和额外描述信息构成了所述服务的信息三元组。

[0018] 上述基于关系挖掘的服务组织方法，利用服务相似度挖掘方法及根据上述服务相似度对服务进行聚类，有效地缩小了服务的搜索范围，提高了服务 定位的效率。

[0019] 本发明实施例提供了一种基于关系挖掘的服务组织装置，该装置包括：

[0020] 生成单元，用于对服务的描述信息进行预处理，并生成信息三元组；

[0021] 计算单元，用于根据所述信息三元组计算所述服务的相似度；

[0022] 聚类单元，用于根据所述相似度对服务进行聚类，生成类别树，根据所述类别树完成服务定位；

[0023] 所述生成单元包括：

[0024] 第一信息获取模块，用于根据服务描述语言 WSDL 文档地址载入所述服务描述语言 WSDL 文档，并对所述服务描述语言 WSDL 文档进行解析，获取服务名称和服务注释信息，并将所述服务名称加入服务的标题信息中；

[0025] 第二信息获取模块，用于获取服务中所有方法的列表，将列表中每个方法的信息加入该服务的主题信息中，所述每个方法的信息包括装置名称、注释、输入消息名称、输出消息名称；

[0026] 第三信息获取模块，用于根据额外描述信息文档地址载入额外描述信息文档，并对所述额外描述信息文档进行解析，并将解析后的额外描述信息加入该服务的额外描述信息中；

[0027] 生成模块，用于根据所述标题信息、主题信息和额外描述信息，生成所述服务的信息三元组。

[0028] 上述基于关系挖掘的服务组织装置，利用生成单元生成信息三元组，利用计算单元计算服务的相似度，并利用聚类单元对上述服务进行聚类，有效地缩小了服务的搜索范围，提高了服务定位的效率。

[0029] 下面通过附图和实施例，对本发明实施例的技术方案做进一步的详细描述。

[0030] 附图说明

[0031] 图 1 为本发明基于关系挖掘的服务组织方法实施例的流程图；

[0032] 图 2 为本发明服务挖掘过程实施例的流程图；

[0033] 图 3 为本发明文档预处理过程实施例的流程图；

[0034] 图 4 为本发明文本预处理过程实施例的流程图；

[0035] 图 5 为本发明类别树生成方法实施例的流程图；

[0036] 图 6 为本发明基于关系挖掘的服务组织装置实施例的结构示意图。

[0037] 具体实施方式

[0038] 如图 1 所示，为本发明基于关系挖掘的服务组织方法实施例的流程图，该方法包括：

[0039] 步骤 101、对服务的描述信息进行预处理，并生成信息三元组；

[0040] 首先对服务的各种描述信息进行预处理，从这些信息中提取出有意义的关键词并构造信息三元组；

[0041] 其中，该步骤可以包括：

[0042] 从原始信息库获取服务描述语言 (WSDL) 文档地址和额外描述信息文档地址；

[0043] 根据上述服务描述语言 WSDL 文档地址载入上述服务描述语言 WSDL 文档，并对上述服务描述语言 WSDL 文档进行解析，获取服务名称和服务注释，并将上述服务名称和服务注释加入服务的标题信息中；

[0044] 获取服务中所有方法的列表，将列表中每个方法的信息加入该服务的主题信息中，上述信息包括方法名称、注释、输入消息名称、输出消息名称；

[0045] 根据上述额外描述信息文档地址载入额外描述信息文档，并对上述额外描述信息文档进行解析，并将解析后的额外描述信息加入该服务的额外描述信息中；

[0046] 上述标题信息、主题信息和额外描述信息构成了上述服务的信息三元组；

[0047] 步骤 102、根据上述信息三元组计算上述服务的相似度；

[0048] 在获得标题信息、主题信息和额外描述信息后，需对上述三类信息进行预处理，转换成符合标准的单词，然后利用如下公式计算第一服务和第二服务的相似度，

[0049] $Sim(W_1, W_2) =$

[0050] $\alpha * SimSet(W_1.T, W_2.T)$

[0051] $+\beta * SimSet(W_1.B, W_2.B)$

[0052] $+\gamma * SimSet(W_1.A, W_2.A)$

[0053] 其中，W_1 表示第一服务，W_2 表示第二服务，$Sim(W_1, W_2)$ 表示第一服务和第二服务的相似度，T 表示标题信息，B 表示主题信息，A 表示额外描述信息，$SimSet(W_1.T, W_2.T)$ 表示第一服务标题信息单词集合和第二服务标题信息单词集合的词义相似度，$SimSet(W_1.B, W_2.B)$ 表示第一服务主题信息单词集合和第二服务主题信息单词集合的词义相似度，$SimSet(W_1.A, W_2.A)$ 表示第一服务额外描述信息单词集合和第二服务额外描述信息单词集合的词义相似度，α、β、γ 分别表示标题信息、主体信息和额外描述信息在第一服务和第二服务相似度中 的权重。

[0054] 步骤 103、根据上述相似度对服务进行聚类，生成类别树，根据上述类别树完成服务定位。

[0055] 将两服务间的相似度表示为两点间距离，根据任意两点间距离均小于预定聚类直径生成类别树。

[0056] 上述基于关系挖掘的服务组织方法，利用服务相似度挖掘方法及根据上述服务相似度对服务进行聚类，有效地缩小了服务的搜索范围，提高了服务定位的效率。

[0057] 如图 2 所示，为本发明服务挖掘过程实施例的流程图，该过程包括：

[0058] 步骤 201、从原始信息库获取服务 WSDL 文档和额外描述信息文档，并对上述文档进行预处理；

[0059] 步骤 202、将预处理得到的关键词集合信息保存到挖掘信息库；

[0060] 步骤 203、判断是否完成对所有服务的预处理，若是，执行步骤 204，否则，转向步骤 201；

[0061] 步骤 204、从挖掘信息库取任意两个服务的关键词集合信息；

[0062] 步骤 205、计算这两个服务的相似度值，把这个相似度值保存到挖掘信息库；

[0063] 步骤 206、判断是否完成对所有相似度的计算，若是，执行步骤 207，否则，转向步骤 204；

[0064] 步骤 207、从原始信息库读取一个服务；

[0065] 步骤 208、聚类并更新类别树；

[0066] 步骤 209、判断是否完成对所有服务的聚类，若是，执行步骤 210，否则，转向步骤 207；

[0067] 步骤 210、将聚类结果保存到挖掘信息库。

[0068] 其中，上述步骤 201 中对文档进行预处理的过程如图 3 所示，该过程包括：

[0069] 步骤 301、从原始信息库获取服务 WSDL 文档地址和额外描述信息文档地 址；

[0070] 步骤 302、载入服务的 WSDL 文档并解析；

[0071] 步骤 303、获取服务名称和服务注释信息，并将服务名称和服务注释信息加入服务标题信息 T 中；

[0072] 步骤 304、获取服务中所有方法的列表；

[0073] 步骤 305、把列表中每个方法的名称、注释、输入消息名称、输出消息名称等加入服务的主题信息 B 中；

[0074] 步骤 306、载入服务的额外描述信息文档并解析；

[0075] 步骤 307、把所有的额外描述信息都加入服务的额外描述信息 A 中；

[0076] 步骤 308、保存解析后的服务功能描述信息三元组到挖掘信息库。

[0077] 通过上述步骤 301-308，生成了信息三元组，生成三元组之后还需对三元组中的三类信息进行文本预处理，其过程如图 4 所示，该过程包括：

[0078] 步骤 401、输入字符串；

[0079] 步骤 402、按标点符号分词；

[0080] 由于英文单词用空格分开，故分词只需把非字母符号替换成空格；

[0081] 步骤 403、拆除连接词；

[0082] 在 WSDL 文档中，服务名称、方法、参数含有重要的服务功能信息且一般采用 Pascal 或 Camel 大小写命名方式，需要进一步拆分，如 RealTimeMarketData 需拆分成 real time market data。

[0083] 步骤 404、过滤停用词；

[0084] 停用词 (stopword) 指句子中一些无描述功能作用的词，如 a，the 以及一些服务常用词如“http”、“post”、“soap”、“get”等，这些词需要被过滤，以提高相似度计算的效率和精度；

[0085] 步骤 405、修正词形；

[0086] 由于一些词是以复数、过去式等非标准形态出现，需把这些词还原成标准形态；

[0087] 步骤 406、过滤停用词；

[0088] 步骤 407、提取名词；

[0089] 名词已基本可描述服务的功能信息，为了提高效率，只利用名词计算相似度；

[0090] 步骤 408、输出单词集合。

[0091] 经过文本预处理后的三元组中的三类信息转换成标准形式，上述步骤 205 计算两服务间的相似度需要计算两关键词集合的相似度，计算关键词集合的相似度目前有多种方法，例如有基于编辑距离的方法、基于规则的方法、基于向量模型的方法、基于交集的方法、基于词频 - 文档频率 (TF-IDF) 的方法等，该实施例采用了马克 (Mailk) 等提出的词性相似度 (Part-of-SpeechSimilarity) 计算方法，在该方法中，给定两个关键词集合 S1 和 S2，首

先把 S1 和 S2 中的单词按词性分类，然后计算 S1 中的每个单词 W1i 到 S2 的距离并累加，再计算 S2 中的每个单词 W2i 到 S1 的距离并累加，最后把这两个累加值相加后除以 S1 和 S2 所含有单词数目的总和，即为集合 S1 和 S2 的相似度，具体计算公式如下：

[0092] $$SimSet_{ps}(S_1,S_2)=\frac{\sum_{w\in S_1}Sim_m(w,S_2)+\sum_{w\in S_2}Sim_m(w,S_1)}{|S_1|+|S_2|}$$

[0093] 其中，$Sim_m(w,S)$ 为单词 W 到词集合 S 的距离，这个距离的定义为词 W 和集合 S 中与词 W 词性相同且最为相似的词 Wi 的相似度值；对于词到词集合相似度的计算可以转换成两个单词相似度的计算，其计算公式如下：

[0094] $$Sim(w_1,w_2)=-\log\frac{\left[\min_{c_1\in sen(w_1),c_2\in sen(w_2)}len(c_1,c_2)\right]}{2d_{max}}$$

[0095] 其中，sen(w) 是指单词 w 所有可能的词义集合，d_{max} 指 WordNet 中名词层次结构树的最大深度，本实施例中只考虑 WordNet 中名词的上下位关系，$len(c_1,c_2)$ 为 c_1、c_2 在这个上下位关系层次结构树中 c_1、c_2 两个节点的最短距离。

[0096] 由于已知计算集合相似度的计算公式，那么采用如下公式可以进一步计算服务间的相似度：

[0097] $Sim(W_1,W_2)$ =

[0098] α *SimSet(W_1.T, W_2.T)

[0099] +β *SimSet(W_1.B, W_2.B)

[0100] +γ *SimSet(W_1.A, W_2.A)

[0101] 其中，W_1 表示第一服务，W_2 表示第二服务，$Sim(W_1,W_2)$ 表示第一服务和第二服务的相似度，T 表示标题信息，B 表示主题信息，A 表示额外描述信息，SimSet(W_1.T, W_2.T) 表示第一服务标题信息单词集合和第二服务标题信息单词集合的词义相似度，SimSet(W_1.B, W_2.B) 表示第一服务主题信息单词集合和第二服务主题信息单词集合的词义相似度，SimSet(W_1.A, W_2.A) 表示第一服务额外描述信息单词集合和第二服务额外描述信息单词集合的词义相似度，α、β、γ 分别表示标题信息、主体信息和额外描述信息在第一服务和第二服务相似度中的权重。

[0102] 在计算完服务间的相似度后，可把服务看成空间中的点，服务间的相似度看成两点间的距离，采用如图 5 所示的类别树生成方法可将距离最近的点聚类，该聚类过程包括：

[0103] 步骤 501、输入类别树树根 T；

[0104] 步骤 502、输入服务 Wi；

[0105] 步骤 503、从服务集合中寻找与 Wi 最相近的服务 Wj 及所属类 Tj，次相近的服务 Wk 及所属类 Tk；

[0106] 步骤 504、判断 Tj 是否存在，若不存在执行步骤 505，若存在，执行步骤 506；

[0107] 步骤 505、构造一个类别 Ti，将 Wi 加入 Ti，将 Ti 加入树根 T，转向步骤 516；

[0108] 步骤 506、判断 Tk 是否存在或 Tk 是否等于 Tj，若 Tk 不存在或 Tk 等于 Tj，执行步骤 507，若 Tk 存在或 Tk 不等于 Tj，执行步骤 508，

[0109] 步骤 507、将 Wi 加入到 Tj，更新 Tj 决定是否分裂，转向步骤 516；

[0110] 步骤 508、将 Wi 加入到 Tj，更新 Tj；

[0111] 步骤 509、判断 Wi 是否为 Tj 的中心点，若不是，执行步骤 510，若是， 执行步骤 511 ；

[0112] 步骤 510、更新 Tj 决定是否分裂，转向步骤 516 ；

[0113] 步骤 511、判断 Tk 中服务数目是否为 1，若是执行步骤 512，否则，执行步骤 513 ；

[0114] 步骤 512、合并 Tj 和 Tk 为新的 Tj，转向步骤 510 ；

[0115] 步骤 513、判断 Wk 是否为 Tk 的中心点，若是转向步骤 510，否则，执行步骤 514 ；

[0116] 步骤 514、将 Wk 加入到 Tj ；

[0117] 步骤 515、更新 Tk 决定是否分裂，转向步骤 510 ；

[0118] 步骤 516、判断是否处理完所有服务，若是聚类结束，否则转向步骤 502。

[0119] 通过上述步骤 501-516，较好地实现了自顶而下的递增式聚类，当读入第一个数据时，将其分为一类，后续读入的数据插入已有的一个合适类中，再跟据类别效应决定是否分裂或者合并相应的类，重复这样的聚类操作直到处理完所有数据，就可以得到一个合适的类别树。

[0120] 但是，在该聚类过程中有三个问题需要处理 ：数据读入顺序对聚类结果的影响 ；类别效应的计算 ；过度拟合的预防 ；本实施例对这三个问题的解决方法如下 ：(1) 通过使用分裂和合并算法消除数据输入顺序的影响 ；(2) 把一个类别看作一个球体，用球体的直径当作类别效应，如果球体的直径越小，则该类的类别效应越好 ；(3) 设置一个球体直径的最大值来限制过度拟合，当球体直径小于这个最小值时，该类不再分裂。

[0121] 另外，在上述实施例中对类别直径与中心点的定义如下 ：一个类别中的所有点 $\{n_1, n_2, .. n_n\}$ 中若以点 n_i 为球心算出球体直径 d_i，这个值比以其他任何点为球心球体直径都小，那么称 d_i 为该类别的直径，称 n_i 为中心点 ；同时，对允许的最大聚类直径 D_{max} 的定义如下 ：如果 $D > D_{max}$，则选择当前聚类中距离最大的两个点，以这两个点为种子，把其余点按距离远近分成两个类，分别计算这两个类的类直径，如果直径大于 D_{max}，则继续对这个类进行分裂直到类直径小于 D_{max}。

[0122] 采用上述聚类方法可有效缩小服务的搜索范围，提高服务的定位效率，假设要从 M 个服务中寻找 1 个与服务 W 最为相似的服务，如果事先没有进行聚类，则查找次数为 M，如果事先已经聚类 (假设有 N 个类，每个类中有 Q_i 个服务，聚类准确度为 a)，查找算法按首先与每个类的中心点服务距离最近确定待寻找服务所在的类，然后再顺序查找这个类中的其他服务，那么平均查找次数 K 为 ：$K = N + \left[\sum_{i=1}^{N} \frac{(Q_i - 1)^2}{M} \right]$，由此可见，搜索效率显著提高。

[0123] 如图 6 所示，为本发明基于关系挖掘的服务组织装置实施例的结构示意图，该装置包括 ：生成单元 1，用于对服务的描述信息进行预处理，并生成信息三元组 ；计算单元 2，用于根据上述信息三元组计算上述服务的相似度 ；聚类单元 3，用于根据上述相似度对服务进行聚类，生成类别树，根据上述类别树完成服务定位。

[0124] 其中，上述生成单元可以包括 ：第一信息获取模块，用于根据上述服务描述语言 WSDL 文档地址载入上述服务描述语言 WSDL 文档，并对上述服务描述语言 WSDL 文档进行解析，获取服务名称，并将上述服务名称加入服务的标题信息中 ；第二信息获取模块，用于获取服务中所有方法的列表，将列表中每个方法的信息加入该服务的主题信息中，上述信息包括装置名称、注释、输入消息名称、输出消息名称 ；第三信息获取模块，用于根据上述额外

描述信息文档地址载入额外描述信息文档，并对上述额外描述信息文档进行解析，并将解析后的额外描述信息加入该服务的额外描述信息中；生成模块，用于根据上述标题信息、主题信息和额外描述信息，生成上述服务的信息三元组。计算单元可以包括：计算模块，用于采用如下公式计算第一服务和第二服务的相似度，

[0125] $Sim(W_1, W_2) =$

[0126] $\alpha * SimSet(W_1.T, W_2.T)$

[0127] $+\beta * SimSet(W_1.B, W_2.B)$

[0128] $+\gamma * SimSet(W_1.A, W_2.A)$

[0129] 其中，W_1 表示第一服务，W_2 表示第二服务，$Sim(W_1, W_2)$ 表示第一服务和第二服务的相似度，T 表示标题信息，B 表示主题信息，A 表示额外描述信息，$SimSet(W_1.T, W_2.T)$ 表示第一服务标题信息单词集合和第二服务标题信息单词集合的词义相似度，$SimSet(W_1.B, W_2.B)$ 表示第一服务主题信息单词集合和第二服务主题信息单词集合的词义相似度，$SimSet(W_1.A, W_2.A)$ 表示第一服务额外描述信息单词集合和第二服务额外描述信息单词集合的词义相似度，α、β、γ 分别表示标题信息、主体信息和额外描述信息在第一服务和第二服务相似度中的权重。

[0130] 另外，上述基于关系挖掘的服务组织装置还可以包括：转换单元，用于对生成单元生成的标题信息、主题信息和额外描述信息进行预处理，转换成符合标准的信息。

[0131] 进一步地，上述聚类单元还可以包括：聚类模块，用于将两服务间的相似度表示为两点间距离，根据任意两点间距离均小于预定聚类直径生成类别树。

[0132] 上述基于关系挖掘的服务组织装置，利用生成单元生成信息三元组，利用计算单元计算服务的相似度，并利用聚类单元对上述服务进行聚类，有效地缩小了服务的搜索范围，提高了服务定位的效率。

[0133] 最后应说明的是：以上实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

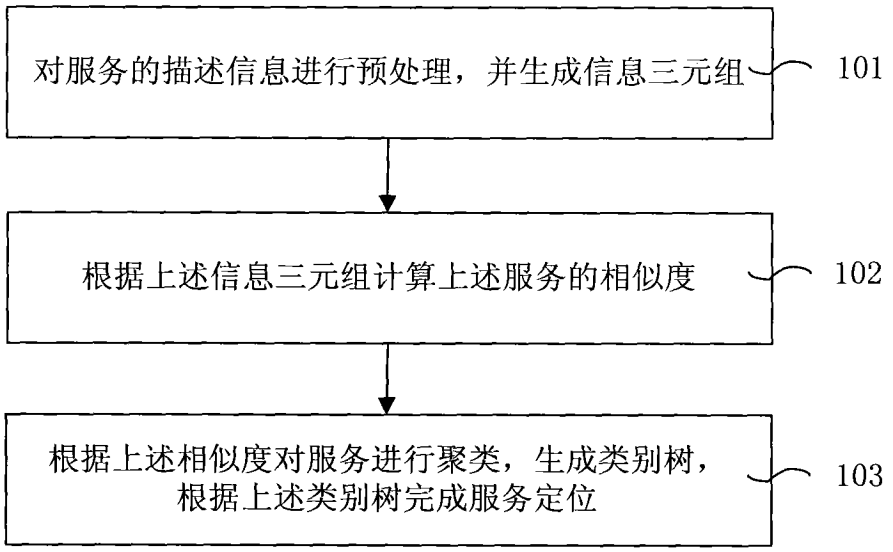

对服务的描述信息进行预处理，并生成信息三元组
101
根据上述信息三元组计算上述服务的相似度
102
根据上述相似度对服务进行聚类，生成类别树，
根据上述类别树完成服务定位
103

图 1

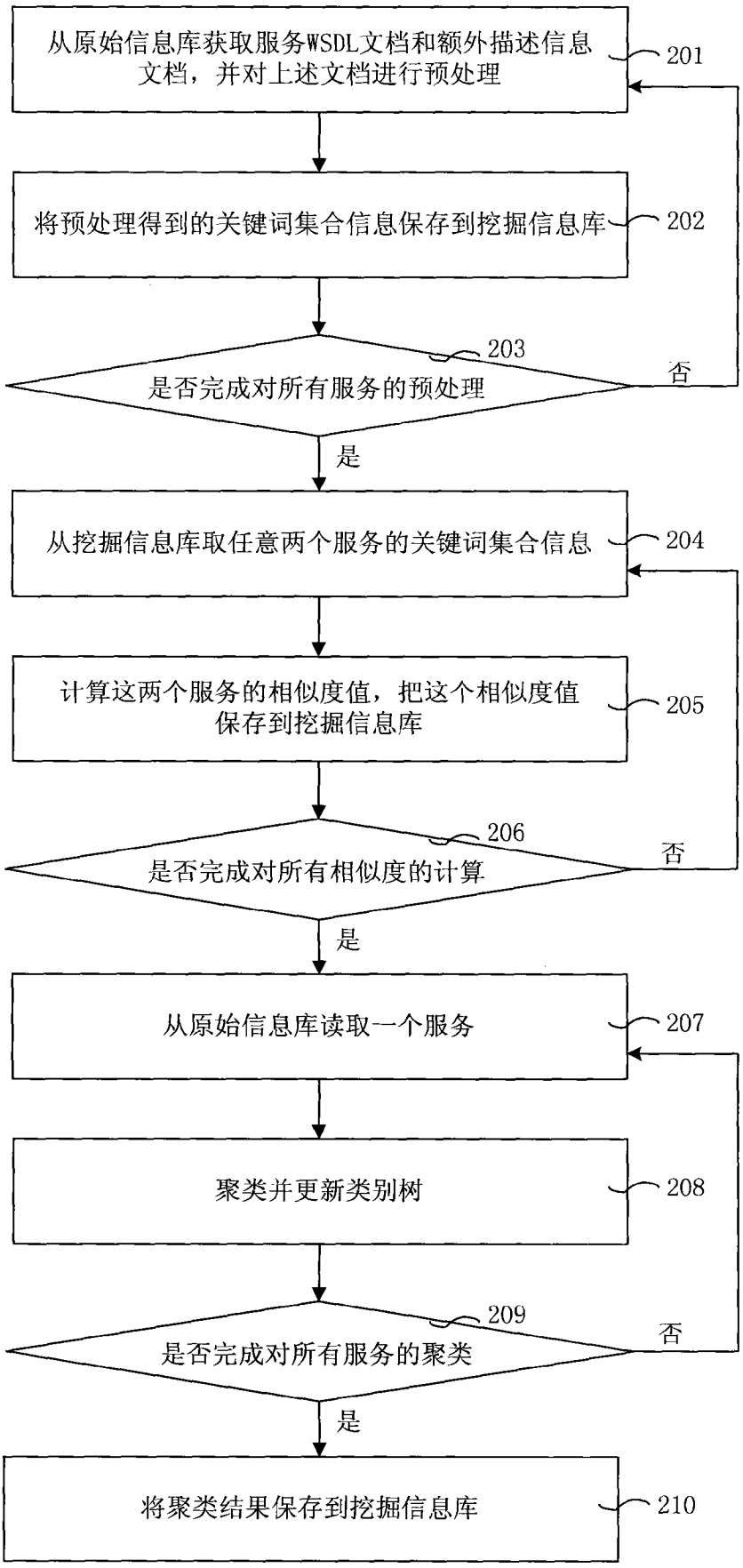

从原始信息库获取服务WSDL文档和额外描述信息文档，并对上述文档进行预处理
201
将预处理得到的关键词集合信息保存到挖掘信息库
202
是否完成对所有服务的预处理
203
否
是
从挖掘信息库取任意两个服务的关键词集合信息
204
计算这两个服务的相似度值，把这个相似度值保存到挖掘信息库
205
是否完成对所有相似度的计算
206
否
是
从原始信息库读取一个服务
207
聚类并更新类别树
208
是否完成对所有服务的聚类
209
否
是
将聚类结果保存到挖掘信息库
210

图 2

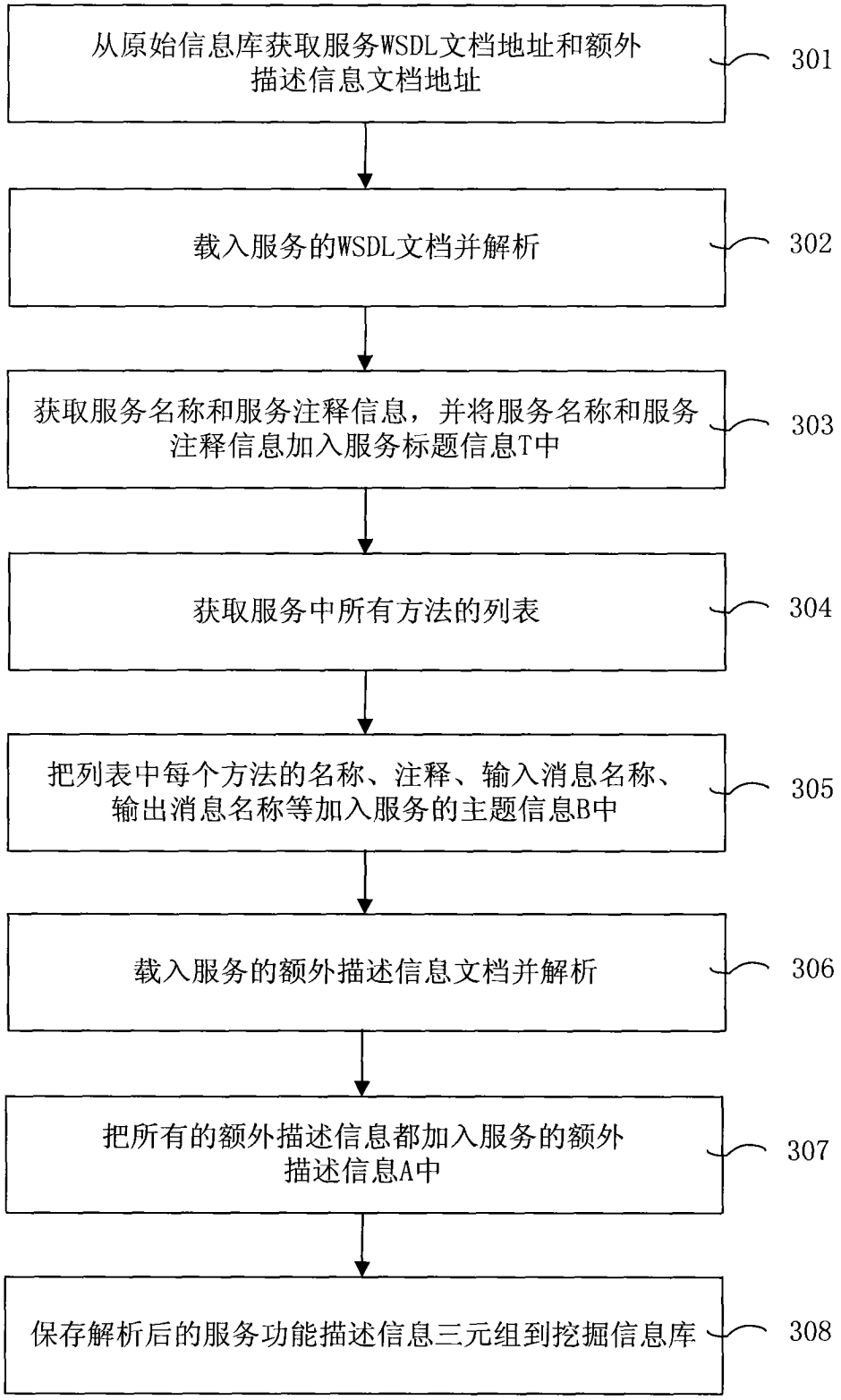

从原始信息库获取服务WSDL文档地址和额外描述信息文档地址
301
载入服务的WSDL文档并解析
302
获取服务名称和服务注释信息，并将服务名称和服务注释信息加入服务标题信息T中
303
获取服务中所有方法的列表
304
把列表中每个方法的名称、注释、输入消息名称、输出消息名称等加入服务的主题信息B中
305
载入服务的额外描述信息文档并解析
306
把所有的额外描述信息都加入服务的额外描述信息A中
307
保存解析后的服务功能描述信息三元组到挖掘信息库
308

图 3

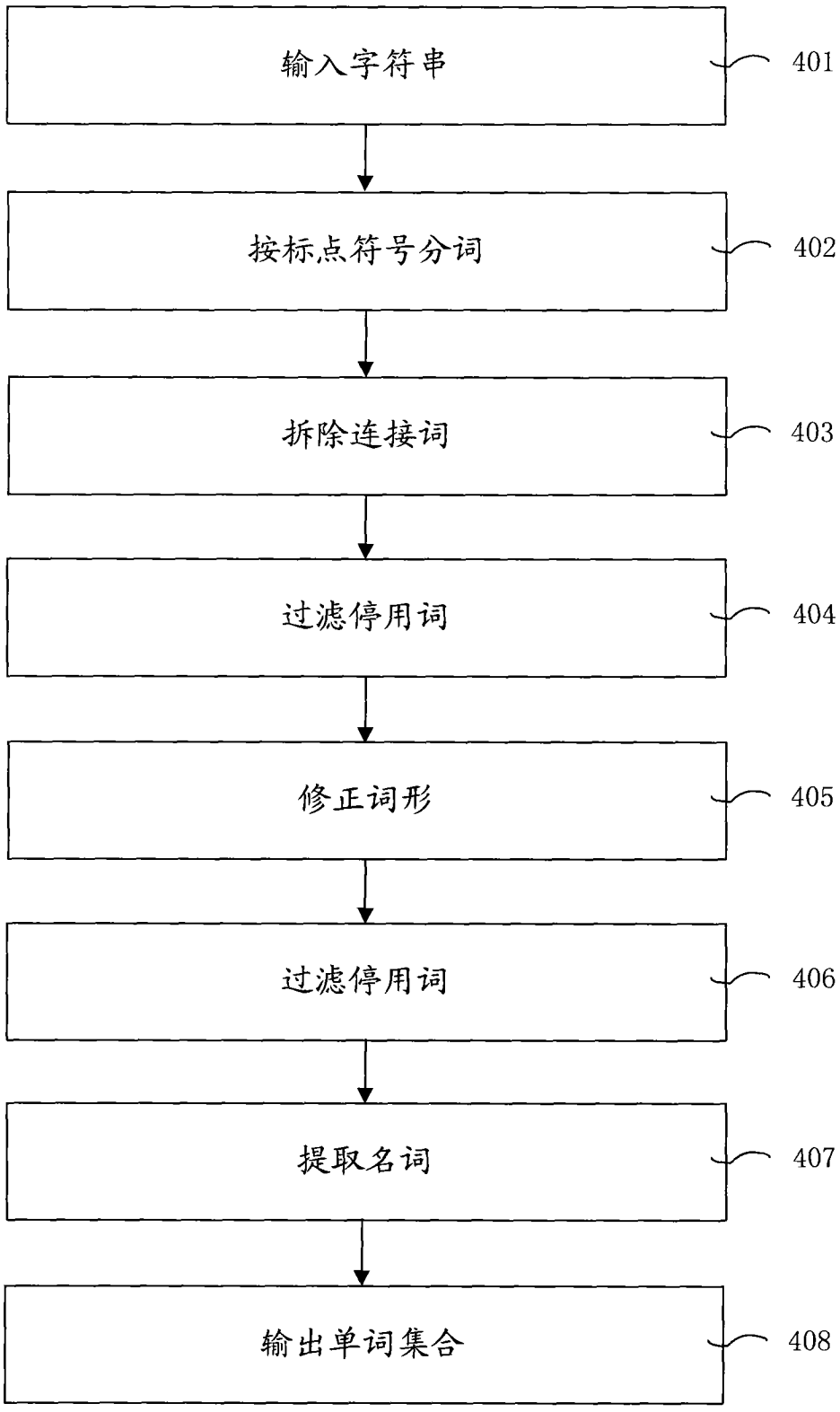

输入字符串
401
按标点符号分词
402
拆除连接词
403
过滤停用词
404
修正词形
405
过滤停用词
406
提取名词
407
输出单词集合
408

图 4

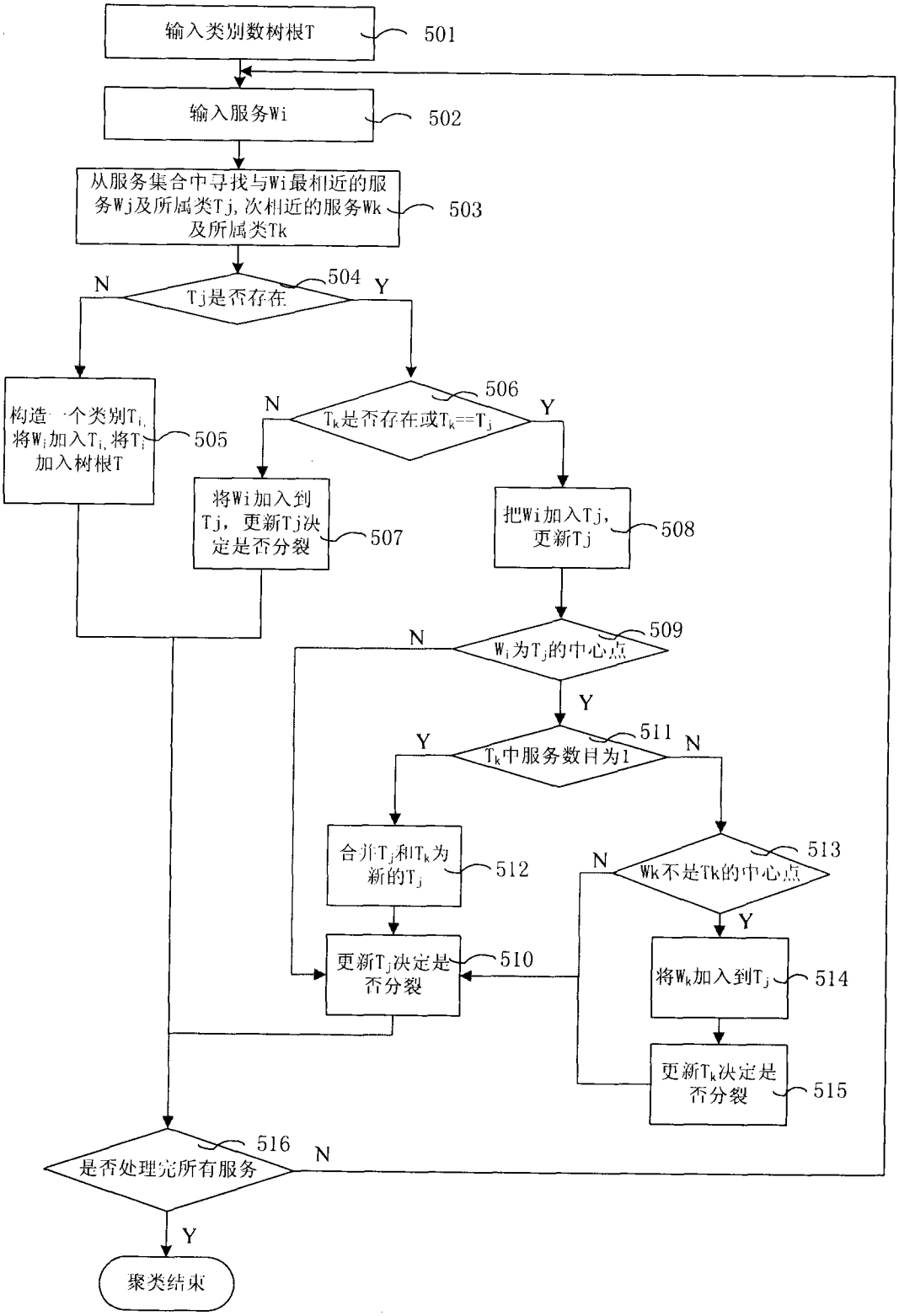

输入类别数树根T
501
输入服务Wi
502
从服务集合中寻找与Wi最相近的服务Wj及所属类Tj，次相近的服务Wk及所属类Tk
503
Tj是否存在
504
N
Y
构造一个类别Ti，将Wi加入Ti，将Ti加入树根T
505
Tk是否存在或Tk==Tj
506
将Wi加入到Tj，更新Tj决定是否分裂
507
把Wi加入Tj，更新Tj
508
Wi为Tj的中心点
509
Tk中服务数目为1
511
合并Tj和Tk为新的Tj
512
Wk不是Tk的中心点
513
更新Tj决定是否分裂
510
将Wk加入到Tj
514
更新Tk决定是否分裂
515
是否处理完所有服务
516
聚类结束

图 5

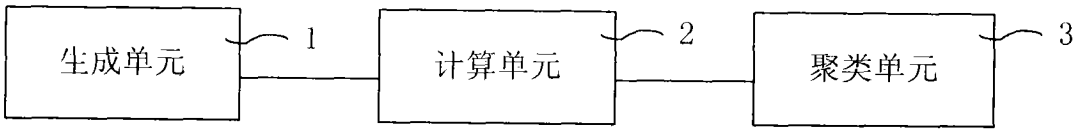


图 6