(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 105373530 A
(43)申请公布日 2016.03.02

(21)申请号 201510881661.4

(22)申请日 2015.12.03

(71)申请人 北京锐安科技有限公司
地址 100044 北京市海淀区西小口路66号中关村东升科技园北领地B-2号楼七层

(72)发明人 敬星 刘鹏

(74)专利代理机构 北京品源专利代理有限公司 11332
代理人 胡彬 孟金喆

(51)Int.Cl.
G06F 17/27(2006.01)

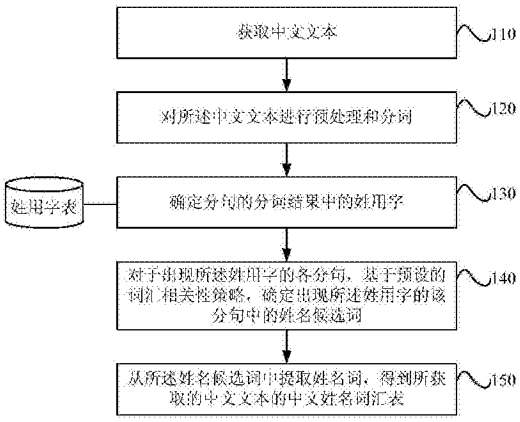
权利要求书2页 说明书7页 附图3页

(54)发明名称

中文姓名的识别方法和装置

(57)摘要

本发明实施例提供一种中文姓名的识别方法和装置。该方法包括：获取中文文本；对所述中文文本进行预处理和分词，得到预处理后的中文文本所包含的分句，以及所包含分句的分词结果；对于各分句，利用姓用字表，确定该分句的分词结果中的姓用字；对于出现所述姓用字的各分句，基于预设的词汇相关性策略，确定出现所述姓用字的该分句中的姓名候选词；从所述姓名候选词中提取姓名词，得到所获取的中文文本的中文姓名词汇表。本方案，能实现“有姓无名”的中文姓名的识别，并避免中文姓名内部成词以及中文姓名与上下文成词的影响，可以最大程度的提取姓名候选词，从而提高了从姓名候选词中提取到的姓名词的识别率。



CN 105373530 A

1. 一种中文姓名的识别方法，其特征在于，包括：

获取中文文本；

对所述中文文本进行预处理和分词，得到预处理后的中文文本所包含的分句，以及所包含分句的分词结果；

对于各分句，利用姓用字表，确定该分句的分词结果中的姓用字；

对于出现所述姓用字的各分句，基于预设的词汇相关性策略，确定出现所述姓用字的该分句中的姓名候选词；

从所述姓名候选词中提取姓名词，得到所获取的中文文本的中文姓名词汇表。

2. 根据权利要求 1 所述的方法，其特征在于，对于出现所述姓用字的各分句，基于预设的词汇相关性策略，确定出现所述姓用字的该分句中的姓名候选词，包括：

对于出现所述姓用字的各分句，判断出现所述姓用字的该分句中是否出现预设的姓名相关特征词；

若是，对所出现的姓用字向后扩展，得到扩展词组；

统计所述扩展词组在所述预处理后的中文文本中的出现次数；

判断所述出现次数是否大于设定阈值；

若大于，则将出现次数大于设定阈值的所述扩展词作为姓名候选词，否则，将向后扩展操作之前的词组作为姓名候选词；

对向后扩展得到的扩展词组进行向后扩展，得到新的扩展词组，并返回执行统计操作。

3. 根据权利要求 1 或 2 所述的方法，其特征在于，在对于各分句，利用姓用字表，确定该分句的分词结果中的姓用字之后，在从所述姓名候选词中提取姓名词之前，所述方法还包括：

利用预先配置的上下文信息表，从出现所述姓用字的各分句中提取姓名候选词。

4. 根据权利要求 3 所述的方法，其特征在于，利用预先配置的上下文信息表，从出现所述姓用字的各分句中提取姓名候选词，包括：

对于出现所述姓用字的各分句，对所出现的姓用字向后扩展，得到多元扩展词组；

利用预先配置的上下文信息表，从所得到的多元扩展词组中提取姓名候选词。

5. 根据权利要求 1 或 2 所述的方法，其特征在于，在从所述姓名候选词中提取姓名词之前，所述方法还包括：

对所述姓名候选词进行去重处理。

6. 一种中文姓名的识别装置，其特征在于，包括：

文本获取模块，用于获取中文文本；

文本处理模块，用于对所述中文文本进行预处理和分词，得到预处理后的中文文本所包含的分句，以及所包含分句的分词结果；

姓用字确定模块，用于对于各分句，利用姓用字表，确定该分句的分词结果中的姓用字；

第一姓名候选词确定模块，用于对于出现所述姓用字的各分句，基于预设的词汇相关性策略，确定出现所述姓用字的该分句中的姓名候选词；

姓名词提取模块，用于从所述姓名候选词中提取姓名词，得到所获取的中文文本的中文姓名词汇表。

7. 根据权利要求 6 所述的装置，其特征在于，第一姓名候选词确定模块具体用于：

对于出现所述姓用字的各分句，判断出现所述姓用字的该分句中是否出现预设的姓名相关特征词；

若是，对所出现的姓用字向后扩展，得到扩展词组；

统计所述扩展词组在所述预处理后的中文文本中的出现次数；

判断所述出现次数是否大于设定阈值；

若大于，则将出现次数大于设定阈值的所述扩展词作为姓名候选词，否则，将向后扩展操作之前的词组作为姓名候选词；

对向后扩展得到的扩展词组进行向后扩展，得到新的扩展词组，并返回执行统计操作。

8. 根据权利要求 6 或 7 所述的装置，其特征在于，所述装置还包括：

第二姓名候选词确定模块，用于在对于各分句，利用姓用字表，确定该分句的分词结果中的姓用字之后，在从所述姓名候选词中提取姓名词之前，利用预先配置的上下文信息表，从出现所述姓用字的各分句中提取姓名候选词。

9. 根据权利要求 8 所述的装置，其特征在于，第二姓名候选词确定模块具体用于：

对于出现所述姓用字的各分句，对所出现的姓用字向后扩展，得到多元扩展词组；

利用预先配置的上下文信息表，从所得到的多元扩展词组中提取姓名候选词。

10. 根据权利要求 6 或 7 所述的装置，其特征在于，所述装置还包括：

姓名候选词过滤模块，用于在从所述姓名候选词中提取姓名词之前，对所述姓名候选词进行去重处理。

中文姓名的识别方法和装置

技术领域

[0001] 本发明实施例涉及信息识别技术领域，尤其涉及一种中文姓名的识别方法和装置。

背景技术

[0002] 汉语的自身特点使得中文信息自动处理大多是先对要处理的文本进行自动分词，如加入显式分割符，然后再在分词的基础上进行词法、语法、以及语义等方面的深入分析。而在分词阶段，中文文本中的中文姓名大多被切分成单字词，在这种情形下如不能很好地解决中文文本中中文姓名的识别问题，将给其后的中文文本的深入分析带来难以逾越的障碍。中文姓名的自动识别问题就是在这种背景下提出来的。对这一问题的研究目前采用的技术有规则方法、统计方法以及规则与统计相结合的方法。

[0003] 其中，规则方法一般是，获取中文文本，并进行分词，根据中文姓名的构成原则得到姓名候选词，从姓名候选词中提取姓名词，得到所获取的中文文本的中文姓名词汇表。其中，中文姓名的构成原则是指：中文姓名一般由二字或三字组成，第一字为姓用字（而复姓则为前两字），其后的一到两个汉字为名用字。

[0004] 统计方法一般包括：使用姓名语料库来训练某个字作为姓名组成部分的概率值；依据其概率值计算某个候选字段作为姓名的概率；其中概率值大于一定阈值的字段为识别出的中文姓名。

[0005] 规则与统计相结合的方法，可以通过概率计算减少规则方法的复杂性与盲目性，而且可以降低统计方法对语料库规模的要求。目前的研究基本上都是采取规则与统计的方法，不同之处仅仅在于规则与统计的侧重不同而已。

[0006] 现有解决方案本身存在着固有的一些不足：

[0007] 首先，扫描到姓氏用字这种具有明显姓名特征的字段时，才将前后的几个字列为姓名候选词进行中文姓名的识别，使得不具备明显姓名特征的中文姓名往往会被丢失，如“有姓无名”的中文姓名，例如“李称杨已离开上海”，“张和刘是好朋友”。其次，姓名候选词大都是选取切分后的碎片，在这种选取机制的作用下，中文姓名内部成词以及中文姓名与上下文成词的情况导致得到的姓名候选词的识别率低，从而导致从其中提取的中文姓名的识别率低。例如：[王国]维，由于内部成词，姓名候选词为切分后的碎片“王国”，这样并不会提取到中文姓名“王国维”。根据对 80,000 条中文姓名的统计，内部成词的比例高达 8.49%，由于这两种成词机制所引起的识别率损失将不小于 10%。

发明内容

[0008] 本发明实施例提供一种中文姓名的识别方法和装置，以提高中文文本中中文姓名的识别率。

[0009] 第一方面，本发明实施例提供了一种中文姓名的识别方法，包括：

[0010] 获取中文文本；

[0011] 对所述中文文本进行预处理和分词，得到预处理后的中文文本所包含的分句，以及所包含分句的分词结果；

[0012] 对于各分句，利用姓用字表，确定该分句的分词结果中的姓用字；

[0013] 对于出现所述姓用字的各分句，基于预设的词汇相关性策略，确定出现所述姓用字的该分句中的姓名候选词；

[0014] 从所述姓名候选词中提取姓名词，得到所获取的中文文本的中文姓名词汇表。

[0015] 第二方面，本发明实施例提供了一种中文姓名的识别装置，包括：

[0016] 文本获取模块，用于获取中文文本；

[0017] 文本处理模块，用于对所述中文文本进行预处理和分词，得到预处理后的中文文本所包含的分句，以及所包含分句的分词结果；

[0018] 姓用字确定模块，用于对于各分句，利用姓用字表，确定该分句的分词结果中的姓用字；

[0019] 第一姓名候选词确定模块，用于对于出现所述姓用字的各分句，基于预设的词汇相关性策略，确定出现所述姓用字的该分句中的姓名候选词；

[0020] 姓名词提取模块，用于从所述姓名候选词中提取姓名词，得到所获取的中文文本的中文姓名词汇表。

[0021] 本发明实施例提供的中文姓名的识别方法和装置，通过对获取到的中文文本进行预处理和分词，得到预处理后的中文文本所包含的分句，以及所包含分句的分词结果；通过确定分句的分词结果中的姓用字，并对出现姓用字的分句进行预设的姓名相关特征词扫描，对于出现预设的姓名相关特征词的分句中的姓用字进行逐位向后扩展，并对每一次向后扩展得到的扩展词组进行处理，一方面，在姓用字向后扩展得到的扩展词组在中文文本中的出现次数不大于设定阈值时，通过将向后扩展操作之前的词组作为姓名候选词，这样，从姓名候选词中提取姓名词，使得中文文本中“有姓无名”的中文姓名能够识别出来；另一方面，由于不仅对姓用字向后扩展的扩展词组进行姓名候选词的判断，而且对姓用字向后扩展得到的扩展词组多次进行向后扩展，每向后扩展一次，从新的扩展词组中提取响应的姓名候选词，从而在避免了中文姓名内部成词以及中文姓名与上下文成词的影响，可以最大程度的提取中文文本中的姓名候选词，这样，再从多次提取到的姓名候选词中提取姓名词，使得中文文本中的中文姓名能够被最大程度的识别，极大提高了中文文本中中文姓名的识别率。

附图说明

[0022] 为了更清楚地说明本发明，下面将对本发明中所需要使用的附图做一简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0023] 图 1a 为本发明实施例一提供的一种中文姓名的识别方法的流程图；

[0024] 图 1b 为本发明实施例一提供的中文姓名的识别方法中基于预设的词汇相关性策略，确定出现姓用字的分句中的姓名候选词的方法流程图；

[0025] 图 2 为本发明实施例二提供的一种中文姓名的识别方法的流程图；

[0026] 图 3 为本发明实施例三提供的一种中文姓名的识别装置的结构示意图。

具体实施方式

[0027] 为使本发明的目的、技术方案和优点更加清楚，下面将结合附图对本发明实施例中的技术方案作进一步详细描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。可以理解的是，此处所描述的具体实施例仅用于解释本发明，而非对本发明的限定，基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。另外还需要说明的是，为了便于描述，附图中仅示出了与本发明相关的部分而非全部内容。

[0028] 在更加详细地讨论示例性实施例之前应当提到的是，一些示例性实施例被描述成作为流程图描绘的处理或方法。虽然流程图将各项操作（或步骤）描述成顺序的处理，但是其中的许多操作可以被并行地、并发地或者同时实施。此外，各项操作的顺序可以被重新安排。当其操作完成时所述处理可以被终止，但是还可以具有未包括在附图中的附加步骤。所述处理可以对应于方法、函数、规程、子例程、子程序等等。

[0029] 实施例一

[0030] 请参阅图 1a，为本发明实施例一提供的一种中文姓名的识别方法的流程图。本发明实施例的方法可以由配置以硬件和 / 或软件实现的中文姓名的识别装置来执行，该识别装置典型的是配置在能够提供中文姓名识别服务的设备中。

[0031] 该方法包括 :步骤 110 ～步骤 150。

[0032] 步骤 110、获取中文文本。

[0033] 步骤 120、对所述中文文本进行预处理和分词，得到预处理后的中文文本所包含的分句，以及所包含分句的分词结果。

[0034] 步骤 130、对于各分句，利用姓用字表，确定该分句的分词结果中的姓用字。

[0035] 中文姓名一般由二字或三字组成，第一字为姓用字（复姓为前两字），姓用字后的一到两个汉字为名用字。统计表明，中文姓名在用字上也有一定规律 :一方面某些字频频出现在姓名中，如在姓用字中，虽然姓氏辞典中列举了几千个姓氏字，但目前实际使用的不过几百个，而张、王、李、赵、以及刘 5 个姓占了 32％。另一方面，某些字又从不被用作姓用字，如最、仅、紧、以、以及且等字。根据这一特性，从语料库中抽取常见的姓用字，形成姓用字表。

[0036] 步骤 140、对于出现所述姓用字的各分句，基于预设的词汇相关性策略，确定出现所述姓用字的该分句中的姓名候选词。

[0037] 请参阅图 1b，本步骤具体包括 :步骤 141 ～步骤 147。

[0038] 步骤 141、对于出现所述姓用字的各分句，判断出现所述姓用字的该分句中是否出现预设的姓名相关特征词 ;若是，执行步骤 142，否则，丢弃出现所述姓用字的该分句中的姓用字。

[0039] 其中，姓名相关特征词是指经常和中文姓名一起出现在分句中的词汇，如“说”、“称”、“指出”、“告诉”、“通知”以及“合作”等，可对大量中文文本进行统计分析，然后人工配置。

[0040] 示例 1

[0041] 示例性地，假设预处理后的中文文本所包含的某分句“王国维和赵朝阳在汉语言

学领域有合作”，该分句中出现两个姓用字，为“王”和“赵”，判断到该分句中出现预设的姓名相关特征词“合作”。

[0042] 示例 2

[0043] 示例性地，假设预处理后的中文文本所包含的某分句“李称杨已离开上海”，该分句中出现两个姓用字，为“李”和“杨”，判断到该分句中出现预设的姓名相关特征词“称”。

[0044] 步骤 142、对所出现的姓用字向后扩展，得到扩展词组。

[0045] 在本步骤中，向后扩展具体是逐位扩展。一次只进行一次向后扩展，每向后扩展一次，继续执行步骤 143。

[0046] 需要说明的是，对于出现姓用字的各分句，在判断到出现所述姓用字的该分句中是否出现预设的姓名相关特征词时，对该分句中所有的姓用字都进行向后扩展。

[0047] 接上述示例 1，对分句“王国维和赵朝阳在汉语言学领域有合作”中的姓用字“王”向后扩展，得到扩展词组“王国”，对姓用字“赵”向后扩展，得到扩展词组“赵朝”。

[0048] 接上述示例 2，对分句“李称杨已离开上海”中的姓用字“李”向后扩展，得到扩展词组“李称”，对姓用字“杨”向后扩展，得到扩展词组“杨已”。

[0049] 步骤 143、统计所述扩展词组在所述预处理后的中文文本中的出现次数。

[0050] 步骤 144、判断所述出现次数是否大于设定阈值，若大于，则执行步骤 145，否则，执行步骤 146，在执行步骤 145 或步骤 146 之后，继续执行步骤 147。

[0051] 步骤 145、将出现次数大于设定阈值的所述扩展词作为姓名候选词。

[0052] 步骤 146、将向后扩展操作之前的词组作为姓名候选词。

[0053] 接上述示例 1，假设该篇中文文本后续分句对“王国维”、“赵朝阳”以及具体的合作进行了详细介绍。判断到扩展词组“王国”和扩展词组“赵朝”在该篇中文文本中的出现次数均大于设定阈值，这样，将扩展词组“王国”和扩展词组“赵朝”均作为姓名候选词。

[0054] 接上述示例 2，假设该篇中文文本后续分句并未对“李”和“杨”进行介绍。判断到扩展词组“李称”和扩展词组“杨已”在该篇中文文本中的出现次数均不大于设定阈值，这样，将向后扩展操作之前的词组“李”和“杨”均作为姓名候选词。

[0055] 步骤 147、对向后扩展得到的扩展词组进行向后扩展，得到新的扩展词组，并返回执行统计操作。

[0056] 接上述示例 1，对分句“王国维和赵朝阳在汉语言学领域有合作”中的扩展词组“王国”进行向后扩展，得到新的扩展词组“王国维”，对扩展词组“赵朝”进行向后扩展，得到新的扩展词组“赵朝阳”。

[0057] 在返回执行统计操作后，由于该篇中文文本后续分句对“王国维”、“赵朝阳”以及具体的合作进行了详细介绍，判断到新的扩展词组“王国维”和新的扩展词组“赵朝阳”在该篇中文文本中的出现次数均大于设定阈值，这样，将新的扩展词组“王国维”和新的扩展词组“赵朝阳”均作为姓名候选词。

[0058] 多次执行后向扩展操作，并返回执行统计操作后，即可得到获取到的中文文本的候选姓名词。

[0059] 步骤 150、从所述姓名候选词中提取姓名词，得到所获取的中文文本的中文姓名词汇表。

[0060] 本步骤中，可以使用支持向量机从姓名候选词中提取姓名词。

[0061] 本实施例的技术方案，通过对获取到的中文文本进行预处理和分词，得到预处理后的中文文本所包含的分句，以及所包含分句的分词结果；通过确定分句的分词结果中的姓用字，并对出现姓用字的分句进行预设的姓名相关特征词扫描，对于出现预设的姓名相关特征词的分句中的姓用字进行逐位向后扩展，并对每一次向后扩展得到的扩展词组进行处理，一方面，在姓用字向后扩展得到的扩展词组在中文文本中的出现次数不大于设定阈值时，通过将向后扩展操作之前的词组作为姓名候选词，这样，从姓名候选词中提取姓名词，使得中文文本中“有姓无名”的中文姓名能够识别出来；另一方面，由于不仅对姓用字向后扩展的扩展词组进行姓名候选词的判断，而且对姓用字向后扩展得到的扩展词组多次进行向后扩展，每向后扩展一次，从新的扩展词组中提取响应的姓名候选词，从而在避免了中文姓名内部成词以及中文姓名与上下文成词的影响，可以最大程度的提取中文文本中的姓名候选词，这样，再从多次提取到的姓名候选词中提取姓名词，使得中文文本中的中文姓名能够被最大程度的识别，极大提高了中文文本中中文姓名的识别率。

[0062] 在上述方案中，在从所述姓名候选词中提取姓名词之前，所述方法还可以包括：

[0063] 对所述姓名候选词进行去重处理。

[0064] 这样处理的好处在于：由于对姓名候选词进行了去重处理，使得从处理后的姓名候选词中提取姓名词的提取效率提高。

[0065] 实施例二

[0066] 请参阅图2，为本发明实施例二提供的一种中文姓名的识别方法的流程图。本实施例在上述实施例的基础上，提供了根据确定的分句的分词结果中的姓用字，确定分句的姓名候选词的技术方案。

[0067] 该方法包括：步骤210～步骤250。

[0068] 步骤210、获取中文文本。

[0069] 步骤220、对所述中文文本进行预处理和分词，得到预处理后的中文文本所包含的分句，以及所包含分句的分词结果。

[0070] 步骤230、对于各分句，利用姓用字表，确定该分句的分词结果中的姓用字。

[0071] 步骤240、对于出现所述姓用字的各分句，基于预设的词汇相关性策略，确定出现所述姓用字的该分句中的姓名候选词。

[0072] 本步骤同样适用于上述实施例中的步骤141～步骤147，不再赘述。

[0073] 步骤250、从所述姓名候选词中提取姓名词，得到所获取的中文文本的中文姓名词汇表。

[0074] 在步骤230之后，在步骤250之前，所述方法还包括：

[0075] 步骤241、利用预先配置的上下文信息表，从出现所述姓用字的各分句中提取姓名候选词。

[0076] 中文姓名在中文文本中不是孤立存在的，其依存的上下文信息具有一定的特点。上下文信息包括：前置信息和后置信息。其中，前置信息是指中文姓名的前端多冠有对人的职业、职务及与说话人的关系的称谓。后置信息是指中文姓名的后端多随有对此人的职业、职务及与说话人的关系的称谓。根据这一特性，从语料库中抽取前置信息词汇和后置信息词汇，形成上下文信息表。

[0077] 本步骤具体可以包括：

[0078] 对于出现所述姓用字的各分句，对所出现的姓用字向后扩展，得到多元扩展词组；

[0079] 利用预先配置的上下文信息表，从所得到的多元扩展词组中提取姓名候选词。

[0080] 需要说明的是，在利用预先配置的上下文信息表中的后置信息词汇时，是将严格位于姓用字和后置信息词汇之间的多元扩展词组作为姓名候选词。

[0081] 示例性地，对于分句“张丽丽老师在学术研讨会上发表了重要讲话”，姓用字为“张”，向后扩展，得到多元扩展词组“张丽”和“张丽丽”，根据预先配置的上下文信息表中的后置信息词汇“老师”，将严格位于姓用字“张”和后置信息词汇“老师”之间的多元扩展词组“张丽丽”作为姓名候选词。

[0082] 在利用预先配置的上下文信息表中的前置信息词汇时，由于中文姓名一般为 2-4 字，姓用字一般为 1-2 字，名用字一般为姓用字之后的 1-3 字，因此通常将前置信息词汇之后姓用字进行设定位数的向后扩展，得到多元扩展词组，直接作为姓名候选词。

[0083] 本实施例对步骤 240 和步骤 241 的执行顺序不进行限制。

[0084] 本实施例的技术方案，对于出现所述姓用字的各分句，基于预设的词汇相关性策略，确定出现所述姓用字的该分句中的姓名候选词，一方面，使得中文文本中“有姓无名”的中文姓名能够识别出来；另一方面，由于不仅对姓用字向后扩展的扩展词组进行姓名候选词的判断，而且对姓用字向后扩展得到的扩展词组多次进行向后扩展，每向后扩展一次，从新的扩展词组中提取响应的姓名候选词，从而在避免了中文姓名内部成词以及中文姓名与上下文成词的影响，可以最大程度的提取中文文本中的姓名候选词，这样，再从多次提取到的姓名候选词中提取姓名词，使得中文文本中的中文姓名能够被最大程度的识别，极大提高了中文文本中中文姓名的识别率；同时，对于出现所述姓用字的各分句，基于预先配置的上下文信息表，提取姓名候选词，并从姓名候选词中提取姓名词，由于丰富了姓名候选词，因此进一步提高了中文文本中中文姓名的识别率。

[0085] 在上述方案中，在从所述姓名候选词中提取姓名词之前，所述方法还可以包括：

[0086] 对所述姓名候选词进行去重处理。

[0087] 由于对姓名候选词进行了去重处理，使得从处理后的姓名候选词中提取姓名词的提取效率提高。

[0088] 实施例三

[0089] 请参阅图 3，为本发明实施例三提供的一种中文姓名的识别装置的结构示意图。该装置包括：文本获取模块 310、文本处理模块 320、姓用字确定模块 330、第一姓名候选词确定模块 340 和姓名词提取模块 350。

[0090] 其中，文本获取模块 310 用于获取中文文本；文本处理模块 320 用于对所述中文文本进行预处理和分词，得到预处理后的中文文本所包含的分句，以及所包含分句的分词结果；姓用字确定模块 330 用于对于各分句，利用姓用字表，确定该分句的分词结果中的姓用字；第一姓名候选词确定模块 340 用于对于出现所述姓用字的各分句，基于预设的词汇相关性策略，确定出现所述姓用字的该分句中的姓名候选词；姓名词提取模块 350 用于从所述姓名候选词中提取姓名词，得到所获取的中文文本的中文姓名词汇表。

[0091] 在上述方案中，第一姓名候选词确定模块 340 可具体用于：

[0092] 对于出现所述姓用字的各分句，判断出现所述姓用字的该分句中是否出现预设的

姓名相关特征词；

[0093] 若是，对所出现的姓用字向后扩展，得到扩展词组；

[0094] 统计所述扩展词组在所述预处理后的中文文本中的出现次数；

[0095] 判断所述出现次数是否大于设定阈值；

[0096] 若大于，则将出现次数大于设定阈值的所述扩展词作为姓名候选词，否则，将向后扩展操作之前的词组作为姓名候选词；

[0097] 对向后扩展得到的扩展词组进行向后扩展，得到新的扩展词组，并返回执行统计操作。

[0098] 在上述方案中，所述装置还可包括：

[0099] 第二姓名候选词确定模块，用于在对于各分句，利用姓用字表，确定该分句的分词结果中的姓用字之后，在从所述姓名候选词中提取姓名词之前，利用预先配置的上下文信息表，从出现所述姓用字的各分句中提取姓名候选词。

[0100] 进一步地，第二姓名候选词确定模块可具体用于：

[0101] 对于出现所述姓用字的各分句，对所出现的姓用字向后扩展，得到多元扩展词组；

[0102] 利用预先配置的上下文信息表，从所得到的多元扩展词组中提取姓名候选词。

[0103] 在上述方案中，所述装置还可包括：

[0104] 姓名候选词过滤模块，用于在从所述姓名候选词中提取姓名词之前，对所述姓名候选词进行去重处理。

[0105] 本发明实施例提供的中文姓名的识别装置可执行本发明任意实施例所提供的中文姓名的识别方法，具备执行方法相应的功能模块和有益效果。

[0106] 最后应说明的是：以上各实施例仅用于说明本发明的技术方案，而非对其进行限制；实施例中优选的实施方式，并非对其进行限制，对于本领域技术人员而言，本发明可以有各种改动和变化。凡在本发明的精神和原理之内所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

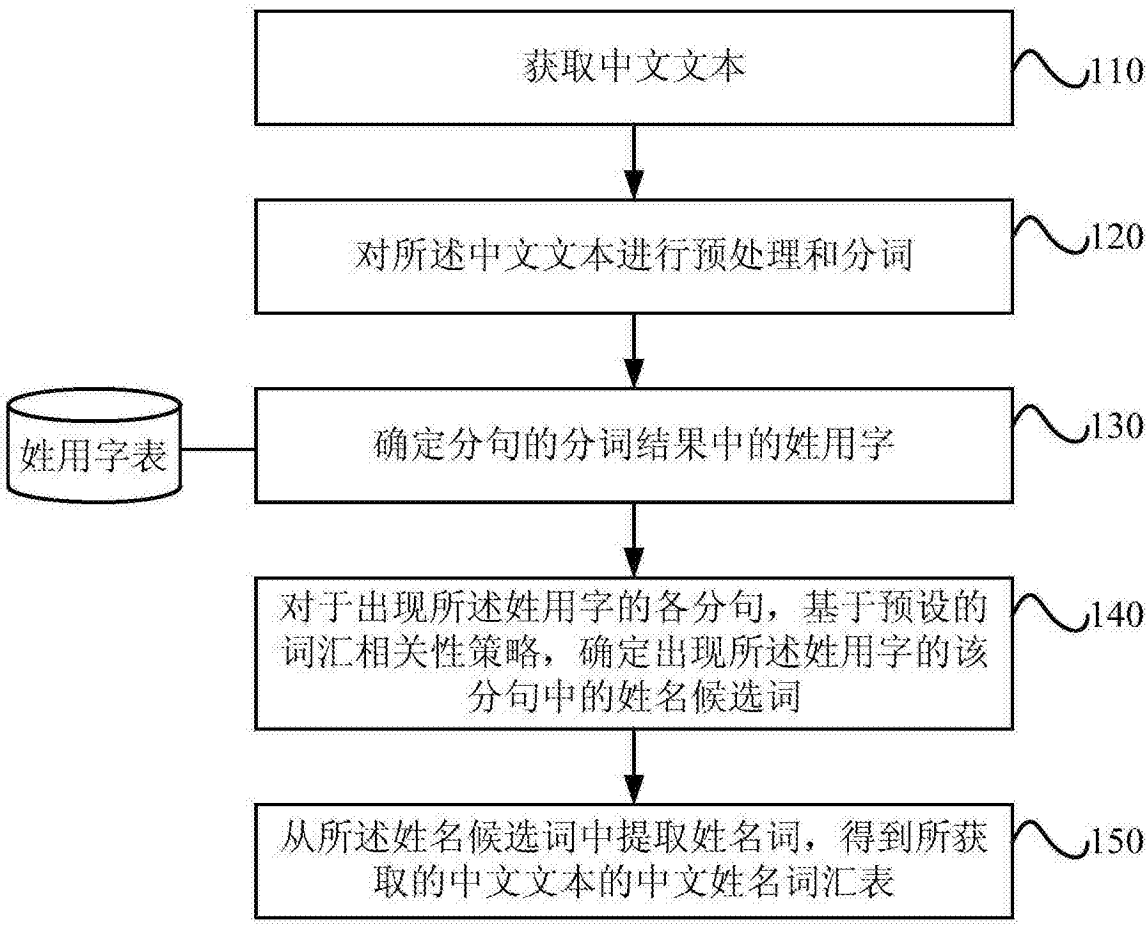

获取中文文本
110
对所述中文文本进行预处理和分词
120
姓用字表
确定分句的分词结果中的姓用字
130
对于出现所述姓用字的各分句，基于预设的词汇相关性策略，确定出现所述姓用字的该分句中的姓名候选词
140
从所述姓名候选词中提取姓名词，得到所获取的中文文本的中文姓名词汇表
150

图 1a

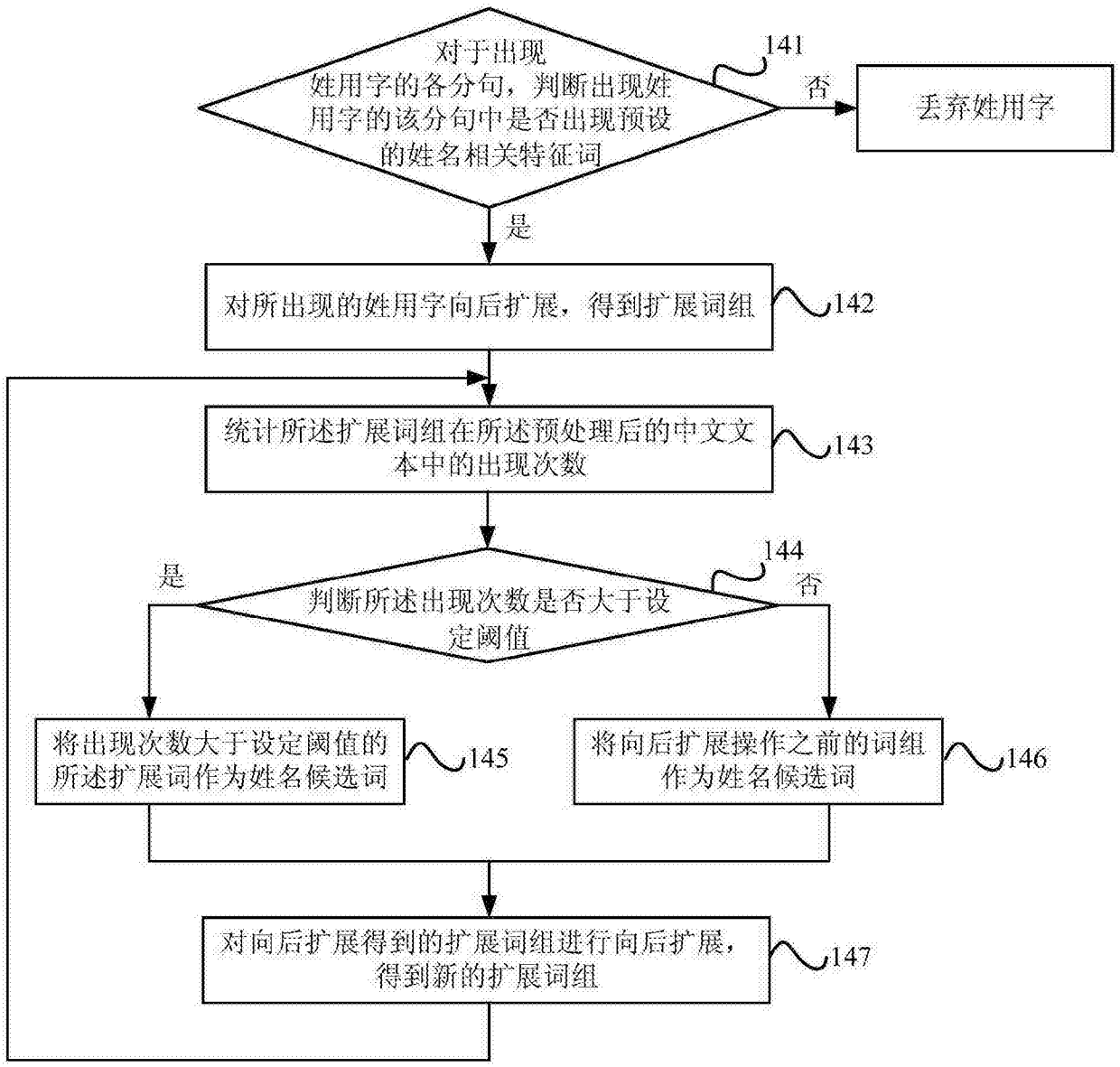

对于出现姓用字的各分句，判断出现姓用字的该分句中是否出现预设的姓名相关特征词
141
否
丢弃姓用字
是
对所出现的姓用字向后扩展，得到扩展词组
142
统计所述扩展词组在所述预处理后的中文文本中的出现次数
143
144
是
否
判断所述出现次数是否大于设定阈值
将出现次数大于设定阈值的所述扩展词作为姓名候选词
145
将向后扩展操作之前的词组作为姓名候选词
146
对向后扩展得到的扩展词组进行向后扩展，得到新的扩展词组
147

图 1b

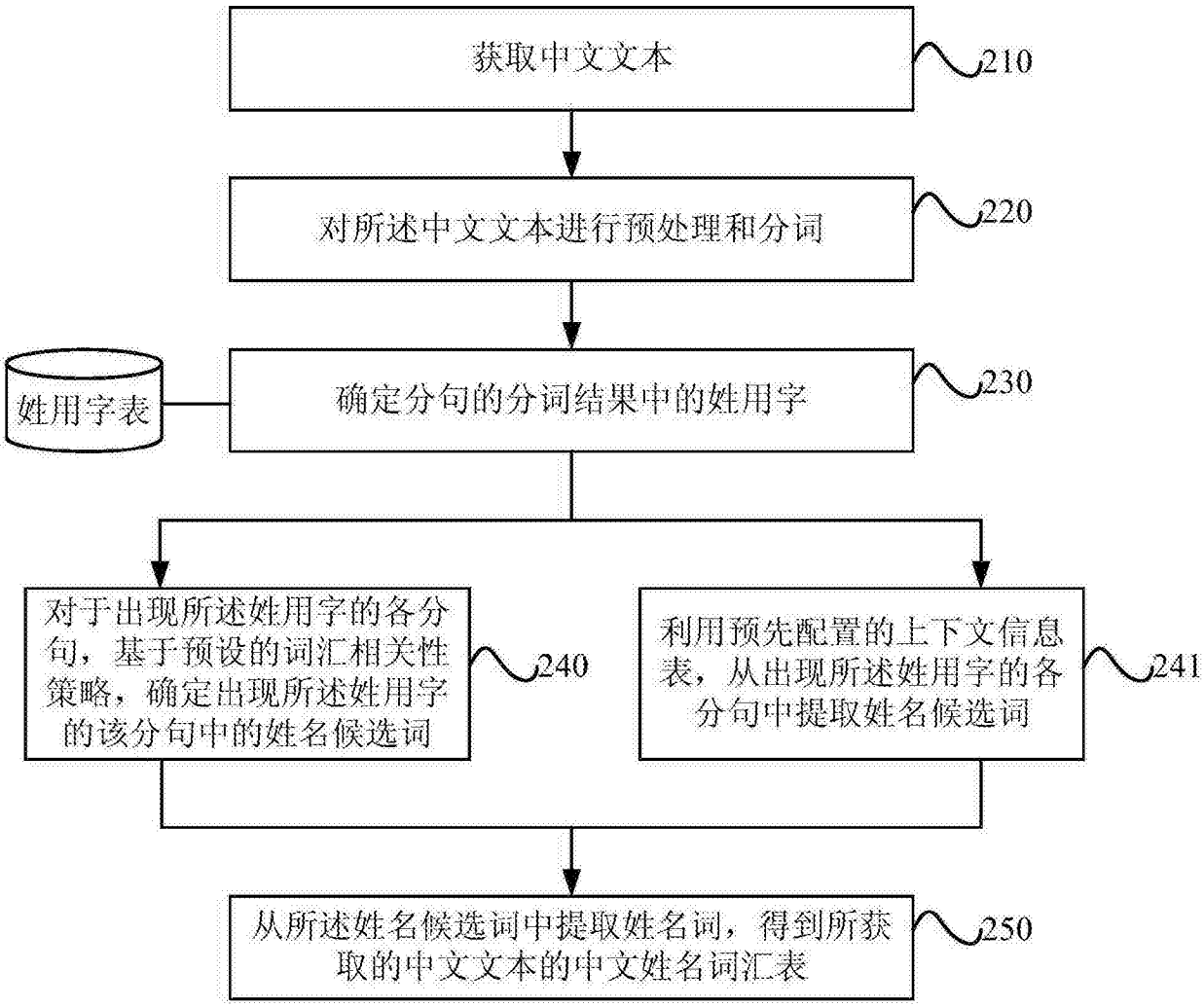

获取中文文本
210
对所述中文文本进行预处理和分词
220
姓用字表
确定分句的分词结果中的姓用字
230
对于出现所述姓用字的各分句，基于预设的词汇相关性策略，确定出现所述姓用字的该分句中的姓名候选词
240
利用预先配置的上下文信息表，从出现所述姓用字的各分句中提取姓名候选词
241
从所述姓名候选词中提取姓名词，得到所获取的中文文本的中文姓名词汇表
250

图 2

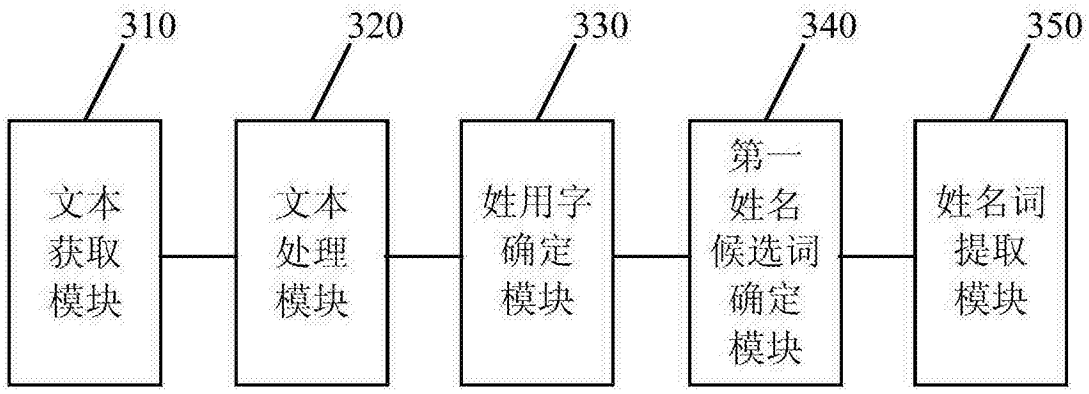

310
文本获取模块
320
文本处理模块
330
姓用字确定模块
340
第一姓名候选词确定模块
350
姓名词提取模块

图 3