(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 106789890 A
(43)申请公布日 2017.05.31

(21)申请号 201611031048.4

(22)申请日 2016.11.17

(71)申请人 北京光年无限科技有限公司
地址 100000 北京市石景山区石景山路3号玉泉大厦四层常青藤青年创业工作室193号

(72)发明人 魏鹏

(74)专利代理机构 北京聿华联合知识产权代理有限公司 11611
代理人 朱绘 张文娟

(51)Int.Cl.
H04L 29/06(2006.01)
G06F 17/30(2006.01)
G06F 21/62(2013.01)

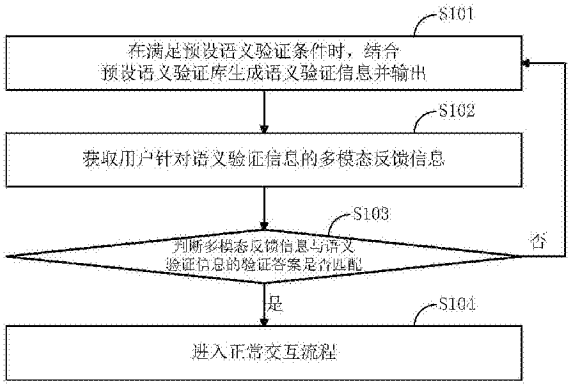
权利要求书1页 说明书6页 附图2页

(54)发明名称

一种用于智能机器人的反数据爬虫的数据处理方法及装置

(57)摘要

一种用于智能机器人的反数据爬虫的数据处理方法及装置，其中，该方法用于智能机器人服务器端的数据库访问，包括：验证信息生成步骤，在满足预设语义验证条件时，结合预设语义验证库生成语义验证信息并输出；语义验证步骤，获取用户针对语义验证信息所输入的多模态反馈信息，并判断多模态反馈信息与语义验证信息的验证答案是否匹配，如果匹配，则进入正常交互流程。该方法通过进行语义验证来判断数据库的访问请求是否是由数据爬虫发起的，由于数据爬虫并不具有语义理解能力，因此数据爬虫也就无法通过上述语义验证过程，这样也就有效保护了服务器端的数据库中的数据，从而达到了反数据爬虫的目的。



CN 106789890 A

1.一种用于智能机器人的反数据爬虫的数据处理方法，其特征在于，所述方法用于智能机器人服务器端的数据库访问，包括：

验证信息生成步骤，在满足预设语义验证条件时，结合预设语义验证库生成语义验证信息并输出；

语义验证步骤，获取用户针对所述语义验证信息所输入的多模态反馈信息，并判断所述多模态反馈信息与所述语义验证信息的验证答案是否匹配，如果匹配，则进入正常交互流程。

2.如权利要求1所述的方法，其特征在于，在所述验证信息生成步骤中，间隔预设时长来生成所述语义验证信息并输出。

3.如权利要求2所述的方法，其特征在于，所述方法还包括：

用户评分步骤，根据语义验证结果对所述用户进行评分，得到所述用户的信用评分；

验证条件调整步骤，根据所述用户的信用评分来调整所述预设时长的长度。

4.如权利要求1～3中任一项所述的方法，其特征在于，在所述预设语义验证库中，每个语义验证信息仅对应一个验证答案。

5.如权利要求1～4中任一项所述的方法，其特征在于，所述智能机器人服务器端的数据库包括以下所列项中的任一项或几项：

对话数据库、图片数据库和情感数据库。

6.一种用于智能机器人的反数据爬虫的数据处理装置，其特征在于，所述装置用于智能机器人服务器端的数据库访问，包括：

验证信息生成模块，其用于在满足预设语义验证条件时，结合预设语义验证库生成语义验证信息并输出；

语义验证模块，其用于获取用户针对所述语义验证信息所输入的多模态反馈信息，并判断所述多模态反馈信息与所述语义验证信息的验证答案是否匹配，如果匹配，则进入正常交互流程。

7.如权利要求6所述的装置，其特征在于，所述验证信息生成模块配置为间隔预设时长来生成所述语义验证信息并输出。

8.如权利要求7所述的装置，其特征在于，所述装置还包括：

用户评分模块，其用于根据语义验证结果对所述用户进行评分，得到所述用户的信用评分；

验证条件调整模块，其用于根据所述用户的信用评分来调整所述预设时长的长度。

9.如权利要求6～8中任一项所述的装置，其特征在于，在所述预设语义验证库中，每个语义验证信息仅对应一个验证答案。

10.如权利要求6～9中任一项所述的装置，其特征在于，所述智能机器人服务器端的数据库包括以下所列项中的任一项或几项：

对话数据库、图片数据库和情感数据库。

一种用于智能机器人的反数据爬虫的数据处理方法及装置

技术领域

[0001] 本发明涉及机器人技术领域，具体地说，涉及一种用于智能机器人的反数据爬虫的数据处理方法及装置

背景技术

[0002] 随着科学技术的不断发展，信息技术、计算机技术以及人工智能技术的引入，机器人的研究已经逐步走出工业领域，逐渐扩展到了医疗、保健、家庭、娱乐以及服务行业等领域。而人们对于机器人的要求也从简单重复的机械动作提升为具有拟人问答、自主性及与其他机器人进行交互的智能机器人，人机交互也就成为决定智能机器人发展的重要因素。

[0003] 对于人机交互系统来说，其所使用的数据库是整个系统的关键，这就使得一些公司会采用数据爬虫来获取竞争对手的数据库内容。因此，如何确保自身数据库的内容无法被竞争对手所获取成为亟需解决的技术问题。

发明内容

[0004] 为解决上述问题，本发明提供了一种用于智能机器人的反数据爬虫的数据处理方法，所述方法用于智能机器人服务器端的数据库访问，其包括：

[0005] 验证信息生成步骤，在满足预设语义验证条件时，结合预设语义验证库生成语义验证信息并输出；

[0006] 语义验证步骤，获取用户针对所述语义验证信息所输入的多模态反馈信息，并判断所述多模态反馈信息与所述语义验证信息的验证答案是否匹配，如果匹配，则进入正常交互流程。

[0007] 根据本发明的一个实施例，在所述验证信息生成步骤中，间隔预设时长来生成所述语义验证信息并输出。

[0008] 根据本发明的一个实施例，所述方法还包括：

[0009] 用户评分步骤，根据语义验证结果对所述用户进行评分，得到所述用户的信用评分；

[0010] 验证条件调整步骤，根据所述用户的信用评分来调整所述预设时长的长度。

[0011] 根据本发明的一个实施例，在所述预设语义验证库中，每个语义验证信息仅对应一个验证答案。

[0012] 根据本发明的一个实施例，所述智能机器人服务器端的数据库包括以下所列项中的任一项或几项：

[0013] 对话数据库、图片数据库和情感数据库。

[0014] 本发明还提供了一种用于智能机器人的反数据爬虫的数据处理装置，所述装置用于智能机器人服务器端的数据库访问，其包括：

[0015] 验证信息生成模块，其用于在满足预设语义验证条件时，结合预设语义验证库生成语义验证信息并输出；

[0016] 语义验证模块，其用于获取用户针对所述语义验证信息所输入的多模态反馈信息，并判断所述多模态反馈信息与所述语义验证信息的验证答案是否匹配，如果匹配，则进入正常交互流程。

[0017] 根据本发明的一个实施例，所述验证信息生成模块配置为间隔预设时长来生成所述语义验证信息并输出。

[0018] 根据本发明的一个实施例，所述装置还包括：

[0019] 用户评分模块，其用于根据语义验证结果对所述用户进行评分，得到所述用户的信用评分；

[0020] 验证条件调整模块，其用于根据所述用户的信用评分来调整所述预设时长的长度。

[0021] 根据本发明的一个实施例，在所述预设语义验证库中，每个语义验证信息仅对应一个验证答案。

[0022] 根据本发明的一个实施例，所述智能机器人服务器端的数据库包括以下所列项中的任一项或几项：

[0023] 对话数据库、图片数据库和情感数据库。

[0024] 本发明所提供的用于智能机器人的反数据爬虫的数据处理方法通过进行语义验证来判断数据库的访问请求是否是由数据爬虫发起的，由于数据爬虫并不具有语义理解能力，因此数据爬虫也就无法通过上述语义验证过程，这样也就有效保护了服务器端的数据库中的数据，从而达到了反数据爬虫的目的。

[0025] 此外，本发明所提供的方法还会利用语义验证结果来调节用户的信用评分，并利用用户的信用评分来对语义验证的频率进行调节。语义验证频率的调节能够进一步防止数据爬虫对数据库进行数据爬取从而保证数据库的安全性。

[0026] 本发明的其它特征和优点将在随后的说明书中阐述，并且，部分地从说明书中变得显而易见，或者通过实施本发明而了解。本发明的目的和其他优点可通过在说明书、权利要求书以及附图中所特别指出的结构来实现和获得。

附图说明

[0027] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要的附图做简单的介绍：

[0028] 图1是根据本发明一个实施例的用于智能机器人的反数据爬虫的数据处理方法的实现流程示意图；

[0029] 图2是根据本发明另一个实施例的用于智能机器人的反数据爬虫的数据处理方法的实现流程示意图；

[0030] 图3是根据本发明一个实施例的用于智能机器人的反数据爬虫的数据处理装置的结构示意图。

具体实施方式

[0031] 以下将结合附图及实施例来详细说明本发明的实施方式，借此对本发明如何应用技术手段来解决技术问题，并达成技术效果的实现过程能充分理解并据以实施。需要说明

的是，只要不构成冲突，本发明中的各个实施例以及各实施例中的各个特征可以相互结合，所形成的技术方案均在本发明的保护范围之内。

[0032] 同时，在以下说明中，出于解释的目的而阐述了许多具体细节，以提供对本发明实施例的彻底理解。然而，对本领域的技术人员来说显而易见的是，本发明可以不用这里的具体细节或者所描述的特定方式来实施。

[0033] 另外，在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行，并且，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

[0034] 网络爬虫(又被称为网页蜘蛛、网络机器人或是网页追逐者)，是一种按照一定的规则，自动的抓取万维网信息的程序或者脚本，已被广泛应用于互联网领域。搜索引擎能够使用网络爬虫抓取Web网页、文档甚至图片、音频、视频等资源，通过相应的索引技术组织这些信息，提供给搜索用户进行查询。

[0035] 为了与用户之间能够进行有效、顺畅以及个性化的人机交互，人机交互系统通常会使用数据库(知识库)来存储数据并通过检索、组合知识数据来生成输出数据。因此，对于人机交互系统来说，数据库中所存储的数据是十分重要的。而一些人机交互系统的提供商为了获取竞争对手的数据库内容，往往会采用数据爬虫来爬取竞争对手的数据库中的数据。

[0036] 针对上述问题，本发明提供了一种新的用于智能机器人的反数据爬虫的数据处理方法，该方法用于智能机器人服务器端的数据库访问。通过对数据爬虫进行分析，发现数据爬虫并不具备语义理解能力，因此本发明所提供的方法采用语义验证的方式来防止数据爬虫来爬取智能机器人服务器端的数据库中的相关数据。

[0037] 需要指出的是，本发明所提及的智能机器人服务器端的数据库既可以指对话数据库，也可以指图片数据库，还可以指情感数据库等其他数据库，抑或是上述两种或多种数据库的组合等，本发明不限于此。

[0038] 为了更加清楚地阐述本发明所提供的用于智能机器人的反数据爬虫的数据处理方法的实现原理、实现流程以及优点，以下分别结合不同的实施例来对该数据处理方法作进一步地说明。

[0039] 实施例一：

[0040] 图1示出了本实施例所提供的用于智能机器人的反数据爬虫的数据处理方法的实现流程示意图。

[0041] 如图1所示，本实施例所提供的数据处理方法首先在步骤S101中在满足预设语义验证条件时，结合预设语义验证库生成语义验证信息并输出。具体地，本实施例中，该方法在步骤S101中间隔预设时长来结合预设语义验证库生成语义验证信息并输出，即如果距离前一次生成并输出语义验证信息的时长达到预设时长，那么该方法也就会判定此时满足预设语义验证条件。

[0042] 需要指出的是，在本发明的不同实施例中，根据实际需要，上述预设时长可以配置为不同的合理值，本发明不限于此。

[0043] 本实施例中，如果判定当前满足预设语义验证条件，该方法会在步骤S101中结合预设语义验证库来生成并输出相应的语义验证信息。优选地，本实施例中，上述预设语义验

证库中包含有词条包，各个词条包中均包含有验证问题和验证答案，该方法会在步骤S101中将所选取的词条包中的验证问题来生成语义验证信息并输出给用户。

[0044] 需要指出的是，为了避免“一问多答”的情况，本实施例中，每一个语义验证信息仅对应一个验证答案，即各个验证问题仅对应一个验证答案。这种对应方式有助于提高验证答案的精确性，并避免由于“一问多答”的情况的存在而给语义判断带来的影响，从而避免数据爬虫对智能机器人服务器端的数据库中数据的爬取。

[0045] 例如，在预设语义验证库中存储有词条包“爸爸的妈妈＝奶奶”，其中，“爸爸的妈妈”即为该词条包的验证问题，而“奶奶”则为对应的验证答案。在满足预设语义验证条件时，该方法在步骤S101中根据“爸爸的妈妈”这一验证问题便可以生成并输出诸如“这个问题我一会再回答你，请告诉我爸爸的妈妈叫什么”的语义验证信息。

[0046] 同时还需要指出的是，本实施例中，该方法在步骤S101中优选地从上述预设语义验证库中随机抽取词条包来进行后续语义验证，然而，在本发明的其他实施例中，该方法还可以采用其他合理方式来从预设语义验证库中抽取用于进行后续语义验证的词条包，本发明不限于此。

[0047] 如图1所示，当输出语义验证信息后，该方法会在步骤S102中获取用户针对上述语义验证信息所输入的多模态反馈信息，并在步骤S103中判断步骤S102中所获取到的多模态反馈信息与上述语义验证信息的验证答案是否匹配。

[0048] 本实施例中，该方法在步骤S103中优选地判断步骤S102中所获取到的多模态反馈信息中的关键词与上述语义验证信息的验证答案是否相同。如果多模态反馈信息中的关键词与上述语义验证信息的验证答案相同，该方法则会判定此时上述多模态反馈信息与语义验证信息的验证答案匹配，从而在步骤S104中进入正常交互流程。而如果多模态反馈信息中的关键词与上述语义验证信息的验证答案不相同，该方法则会判定此时上述多模态反馈信息中的关键词与语义验证信息的验证答案不匹配，此时该方法将返回步骤S101以重新进行语义验证直至语义验证成功为止。

[0049] 当然，在本发明的其他实施例中，如果该方法在步骤S103中判定出步骤S102中所获取到的多模态反馈信息中的关键词与语义验证信息的验证答案不匹配，其可以通过重复返回步骤S101以重新进行语义验证来确定此次数据库的访问发起方是否为数据爬虫。

[0050] 其中，如果上述语义验证重复次数达到预设次数(根据实际需要，该预设次数可以设置为1次或多次)，那么该方法也就可以判定此次数据库的访问发起方为数据爬虫，从而拒绝此次数据库的反问请求，这样也就实现了反数据爬虫的效果。

[0051] 从上述描述中可以看出，本实施例所提供的用于智能机器人的反数据爬虫的数据处理方法通过进行语义验证来判断数据库的访问请求是否是由数据爬虫发起的，由于数据爬虫并不具有语义理解能力，因此数据爬虫也就无法通过上述语义验证过程，这样也就有效保护了服务器端的数据库中的数据，从而达到了反数据爬虫的目的。

[0052] 实施例二：

[0053] 图2示出了本实施例所提供的用于智能机器人的反数据爬虫的数据处理方法的实现流程示意图。

[0054] 如图2所示，本实施例所提供的数据处理方法首先在步骤S201中在满足预设语义验证条件时，结合预设语义验证库生成验证信息并输出。当输出语义验证信息后，该方法会

在步骤S202中获取用户针对上述语义验证信息所输入的多模态反馈信息，并在步骤S203中判断步骤S202中所获取到的多模态反馈信息与上述语义验证信息的验证答案是否匹配，并得到语义验证结果。其中，如果步骤S202中所获取到的多模态反馈信息与语义验证信息的验证答案匹配，那么该方法也就会在步骤S204中根据上述语义验证结果来进入正常交互流程。

[0055] 需要指出的是，本实施例中步骤S201至步骤S204实现各自功能的原理以及流程与上述实施例一中步骤S101至步骤S104所涉及的内容类似，故在此不再对步骤S201至步骤S204的相关内容进行赘述。

[0056] 如图2所示，在得到上述验证结果后，本实施例所提供的方法还会在步骤S205中根据该验证结果对当前用户进行评分，从而得到该用户的信用评分。需要指出的是，本实施例中所涉及的“用户”既可能是具有数据库访问权限的合作用户，也可能是数据爬虫，本发明不限于此。

[0057] 本实施例中，该方法优选地在步骤S205中根据多模态反馈信息与语义验证信息的验证答案是否匹配来对当前用户进行评分。其中，如果多模态反馈信息与语义验证信息的验证答案匹配，该方法则会上调该用户的信用评分。而如果多模态反馈信息与语义验证信息的验证答案不匹配，该方法则会下调该用户的信用评分。

[0058] 在得到用户的信用评分后，该方法会在步骤S206中根据步骤S205中所得到的用户的信用评分来调整上述预设时长的长度，即调整生成并输出语义验证信息的时间间隔，也可以视作调整进行语义验证的频率。其中，如果用户的信用评分较高，那么该方法也就会在步骤S206中降低进行语义验证的频率；而如果用户的信用评分较低，那么为了防止数据爬虫对数据库进行数据爬取从而保证数据库的安全性，该方法则会在步骤S206中提高进行语义验证的频率。

[0059] 从上述描述中可以看出，本实施例所提供的方法在实施例一所提供的方法的基础上，还会利用语义验证结果来调节用户的信用评分，并利用用户的信用评分来对语义验证的频率进行调节。语义验证频率的调节能够进一步防止数据爬虫对数据库进行数据爬取从而保证数据库的安全性。

[0060] 本发明还提供了一种用于智能机器人的反数据爬虫的数据处理装置，该数据处理装置用于智能机器人服务器端的数据库访问。需要指出的是，本发明所提及的智能机器人服务器端的数据库既可以指对话数据库，也可以指图片数据库，还可以指情感数据库等其他数据库，抑或是上述两种或多种数据库的组合等，本发明不限于此。

[0061] 图3示出了本实施例中该数据处理装置的结构示意图。如图3所示，本实施例中，该数据处理装置优选地包括：验证信息生成模块301以及语义验证模块302。其中，验证信息生成模块301用于在满足预设语义验证条件时，结合预设语义验证库来生成相应的语义验证信息并输出。具体地，验证信息生成模块301优选地间隔预设时长来结合语义验证库生成语义验证信息并输出。

[0062] 当验证信息生成模块301输出上述语义验证信息后，语义验证模块302会获取用户针对上述语义验证信息所输入的多模态反馈信息。随后，语义验证模块302会判断该多模态反馈信息与语音验证信息的验证答案是否匹配。如果二者匹配，那么语义验证模块302将会判定当前用户并非数据爬虫，从而进入正常交互流程。而如果二者不匹配，那么语义验证模

块302则会生成相应的指示信息并将该指示信息发送给验证信息生成模块301，以由验证信息生成模块301重新生成语义验证信息。

[0063] 需要指出的是，本实施例中，验证信息生成模块301以及语义验证模块302实现其各自功能的具体原理以及流程与上述实施例一中步骤S101至步骤S104所涉及的内容类似，故在此不再对验证信息生成模块301和语义验证模块302的相关内容进行赘述。

[0064] 如图3所示，本实施例所提供的数据处理方法还包括：用户评分模块303以及验证条件调整模块304。其中，用户评分模块303与语义验证模块302连接，其能够根据语义验证模块302所生成的验证结果来对当前用户进行评分，从而得到当前用户的信用评分。

[0065] 在生成用户的信用评分后，用户评分模块303会将上述用户的信用评分发送至验证条件调整模块304。验证条件调整模块304在接收到上述用户的信用评分后，会根据该信用评分来调整上述预设时长的长度，即调整生成并输出语义验证信息的时间间隔，也可以视作调整进行语义验证的频率。

[0066] 其中，如果用户的信用评分较高，那么验证条件调整模块304会降低进行语义验证的频率；而如果用户的信用评分较低，那么为了防止数据爬虫对数据库进行数据爬取从而保证数据库的安全性，验证条件调整模块304则会提高进行语义验证的频率。

[0067] 需要指出的是，本实施例中，用户评分模块303以及验证条件调整模块304实现其各自功能的具体原理以及过程与上述实施例二中步骤S205和步骤S206所涉及的内容类似，故在此不再对用户评分模块303以及验证条件调整模块304的相关内容进行赘述。

[0068] 应该理解的是，本发明所公开的实施例不限于这里所公开的特定结构或处理步骤，而应当延伸到相关领域的普通技术人员所理解的这些特征的等同替代。还应当理解的是，在此使用的术语仅用于描述特定实施例的目的，而并不意味着限制。

[0069] 说明书中提到的“一个实施例”或“实施例”意指结合实施例描述的特定特征、结构或特性包括在本发明的至少一个实施例中。因此，说明书通篇各个地方出现的短语“一个实施例”或“实施例”并不一定均指同一个实施例。

[0070] 虽然上述示例用于说明本发明在一个或多个应用中的原理，但对于本领域的技术人员来说，在不背离本发明的原理和思想的情况下，明显可以在形式上、用法及实施的细节上作各种修改而不用付出创造性劳动。因此，本发明由所附的权利要求书来限定。

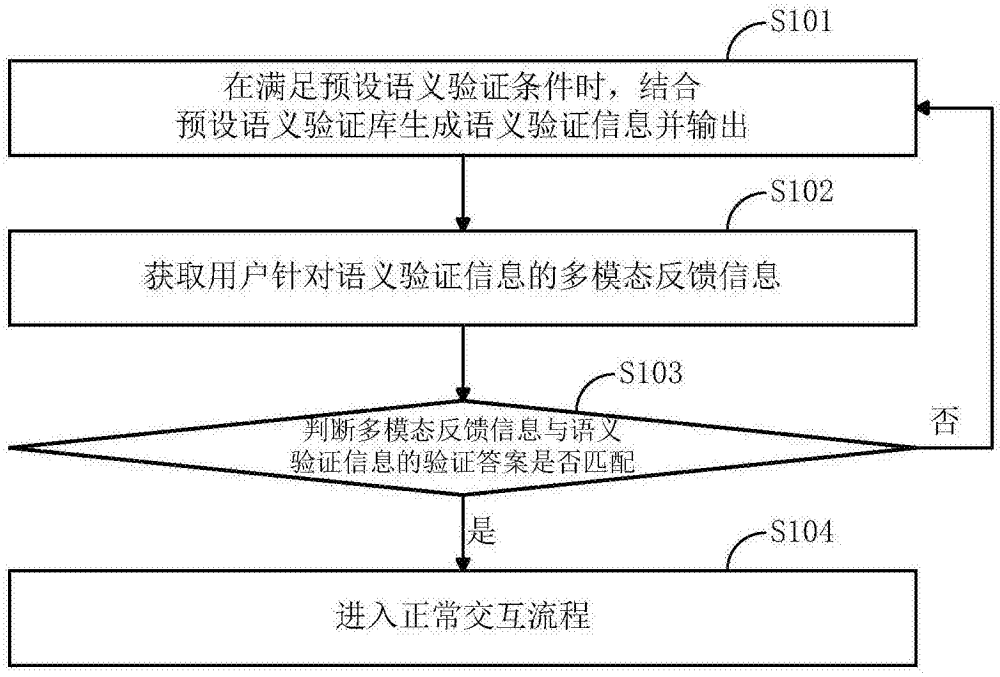

S101
在满足预设语义验证条件时，结合
预设语义验证库生成语义验证信息并输出
S102
获取用户针对语义验证信息的多模态反馈信息
S103
判断多模态反馈信息与语义
验证信息的验证答案是否匹配
否
是
S104
进入正常交互流程

图1

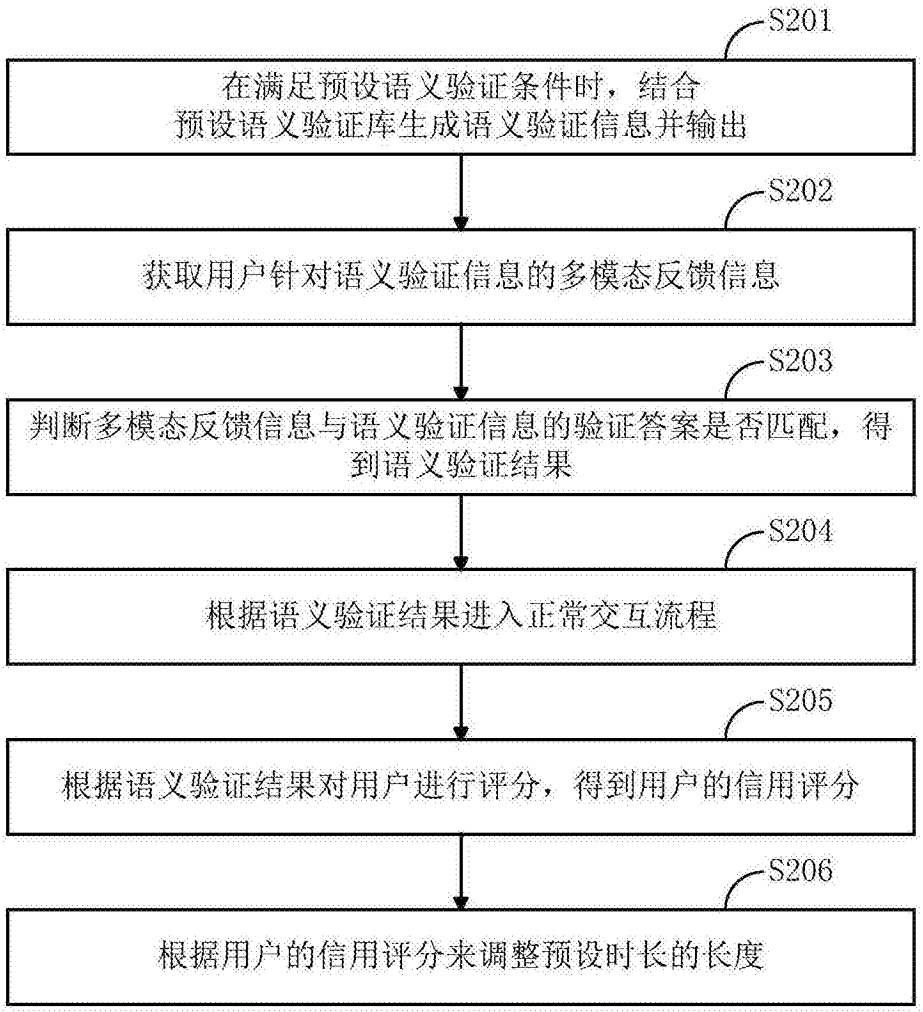

S201
在满足预设语义验证条件时，结合
预设语义验证库生成语义验证信息并输出
S202
获取用户针对语义验证信息的多模态反馈信息
S203
判断多模态反馈信息与语义验证信息的验证答案是否匹配，得到语义验证结果
S204
根据语义验证结果进入正常交互流程
S205
根据语义验证结果对用户进行评分，得到用户的信用评分
S206
根据用户的信用评分来调整预设时长的长度

图2

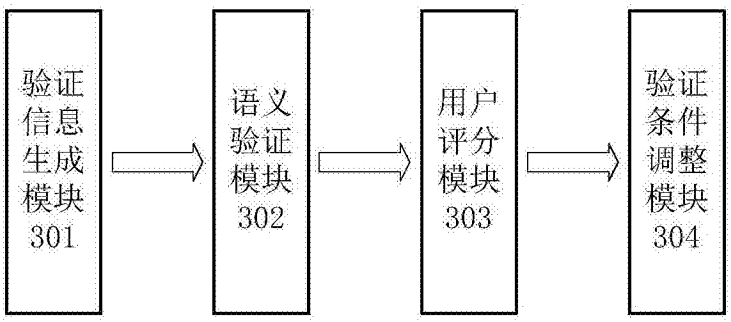

验证信息生成模块301
语义验证模块302
用户评分模块303
验证条件调整模块304

图3