[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06N 5/02 (2006.01)
G06F 17/30 (2006.01)

[12] 发明专利申请公布说明书

[21] 申请号 200610059919.3

[43] 公开日 2007 年 9 月 5 日

[11] 公开号 CN 101030267A

[22] 申请日 2006.2.28
[21] 申请号 200610059919.3
[71] 申请人 腾讯科技（深圳）有限公司
地址 518044 广东省深圳市福田区振兴路赛格科技园 2 栋东 410 室
[72] 发明人 杨海松 邓大付 余祥鑫

[74] 专利代理机构 北京集佳知识产权代理有限公司
代理人 张 良 逯长明

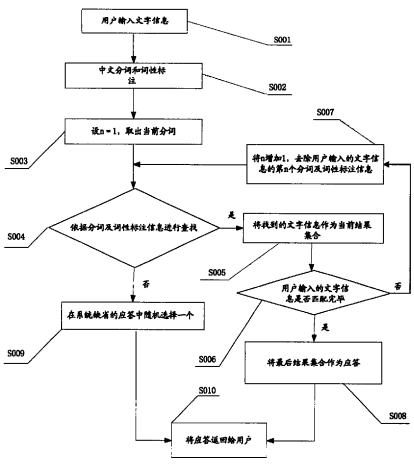
权利要求书 3 页 说明书 11 页 附图 2 页

[54] 发明名称

自动问答方法及系统

[57] 摘要

本发明公开了一种自动问答方法，包括：将输入的文字信息进行切分；根据切分的结果进行查找；用相匹配的查找结果刷新结果集合；判断输入的文字信息匹配是否完成；当输入的文字信息匹配完成，依据结果集合选择应答。 本发明还公开了一种自动问答系统。 本发明能够应用于不同的语言，特别是针对中文用词灵活、句法复杂多变的特点，在自动问答系统中利用中文词性通配符可以提高中文知识库的内容覆盖面，减少建库的工作量，同时显著的提高基于模式匹配的中文自动问答系统的准确率和招回率。



1、一种自动问答方法，包括:

（1） 将输入的文字信息进行切分;

（2） 根据切分的结果进行查找;

（3） 用相匹配的查找结果刷新结果集合;

（4） 判断输入的文字信息匹配是否完成;

（5） 当输入的文字信息匹配完成，依据结果集合选择应答。

2、根据权利要求 1 所述的自动问答方法，其特征在于，推理知识库中存储有所述库存文字信息。

3、根据权利要求 2 所述的自动问答方法，其特征在于，所述步骤（2）中，库存文字信息经过分词和词性标注处理。

4、根据权利要求 1 所述的自动问答方法，其特征在于，所述步骤（1）具体为，中文分词和词性标注模块对输入的文字信息进行切分处理，输出文字信息的分词和词性标注信息。

5、根据权利要求 1 所述的自动问答方法，其特征在于，步骤（2）具体为，根据输入文字信息的分词和词性标注信息，在结果集合内查找具有相同分词的库存文字信息；步骤（4）具体为，当输入文字信息匹配没有完成，提取输入文字信息的下一个分词作为查找依据，并将结果集合作为查找目标，执行步骤（2）。

6、根据权利要求 1 所述的自动问答方法，其特征在于，步骤（2）具体为:

根据输入文字信息的分词和词性标注信息，在结果集合内查找具有指定词性通配符的库存文字信息；步骤（4）具体为，当输入文字信息匹配没有完成，提取输入文字信息的下一个分词作为查找依据，并将结果集合作为查找

目标，执行步骤（2）。

7、根据权利要求 1 所述的自动问答方法，其特征在于，步骤（2）具体为：

根据输入文字信息的分词和词性标注信息，在结果集合内查找指定任意词通配符的库存文字信息；步骤（4）具体为，当输入文字信息匹配没有完成，提取输入文字信息的下一个分词作为查找依据，并将结果集合作为查找目标，执行步骤（2）。

8、根据权利要求 1 所述的自动问答方法，其特征在于，步骤（2）具体为：

根据输入文字信息的分词和词性标注信息，查找具有相同分词的库存文字信息，并按照设定的分值积分；同时，根据输入文字信息的分词和词性标注信息，在结果集合内中查找指定词性通配符的库存文字信息，并按照设定的分值积分；步骤（4）具体为，当输入文字信息匹配没有完成，提取文字信息的下一个分词作为查找依据，并将结果集合作为查找目标，执行步骤（2）。

9、根据权利要求 8 所述的自动问答方法，其特征在于，步骤（2）还包括：

根据输入文字信息的分词和词性标注信息，在结果集合内查找指定任意词通配符的库存文字信息，并按照设定的分值积分；步骤（4）具体为，当输入文字信息匹配没有完成，提取文字信息的下一个分词作为查找依据，并将结果集合作为查找目标，执行步骤（2）。

10、根据权利要求 9 所述的自动问答方法，其特征在于，步骤（2）进一步包括：

当没有查找到相匹配的库存文字信息时，在推理知识库中的缺省应答中随机选择一个作为应答，发送并结束。

11、一种自动问答系统，其特征在于,包括::

网络接口模块，接收输入的文字信息， 将应答发送;

分词和词性标注模块，对输入的文字信息进行分词和词性标注，将切分出来的分词及其词性标注信息发送;

推理模块，根据分词及其词性标注信息查找对应的应答，将应答发送到所述的网络接口模块。

12、根据权利要求 11 所述的自动问答系统，其特征在于，所述分词和词性标注模块调用计算语言知识库中的分词和词性标注方法。

13、根据权利要求 12 所述的自动问答系统，其特征在于，所述计算语言知识库中还存储有词语使用频率的统计数据和词库。

14、根据权利要求 11 所述的自动问答系统，其特征在于，所述推理模块调用推理知识库中与文字信息相对应的应答。

15、根据权利要求 14 所述的自动问答系统，其特征在于，所述推理知识库存储有库存文字信息，所述库存文字信息经过分词或者词性标注处理。

16、根据权利要求 15 所述的自动问答系统，其特征在于，所述库存文字信息包含分词、任意词通配符或者词性通配符。

自动问答方法及系统

技术领域

本发明涉及一种计算机应用系统及方法，具体说，利用语言匹配技术的自动问答方法及语言的自动问答系统。

背景技术

现有的语言自动问答系统中，大多是采用简单的模式匹配技术实现的，其方法是从句子的第一个词开始，对用户输入的句子和知识库中的句子进行匹配，如果两个词相同就继续下一个词的匹配，中间可能利用任意词通配符来忽略掉用户输入的句子中存在的一些不太关键的词，重复这一过程直到整个用户输入的句子匹配完毕，如果匹配成功就将知识库中的句子所对应的应答返回给用户。但是相对于外文而言，中文具有用词灵活、句法复杂多变的特点，并不适合简单的模式匹配技术。现有的中文自动问答系统是参考了国外一些英文的自动问答系统，采用简单的模式匹配技术实现的，这导致中文自动问答系统普遍存在中文知识库的覆盖面窄、系统的准确率和招回率都很低的问题，对用户体验造成了伤害。

自动问答系统又称QA（automatic Question Answering）系统，它采用自然语言处理技术，一方面完成对用户问题的分析处理，另一方面完成正确答案的生成。自动问答系统以自然语言理解技术为核心，涉及到计算语言学、信息科学和人工智能等多门学科，是计算机应用研究的热点之一。

自然语言理解是人工智能领域中的一个重要研究方向，它使计算机能够理解和运用人类的自然语言，可以实现人与计算机之间基于自然语言的有效通信。

知识库是自动问答系统的关键组成部分，通常以问答语句对的形式存储了大量的信息。当用户输入的自然语言句子与知识库中的某一个句子匹配成功的时候，其对应的应答就会被返回给用户。

中文分词和词性标注词是最小的能够独立活动的有意义的语言成分。在中文中，词与词之间不存在分隔符，词本身也缺乏明显的形态标记，因此，中文信息处理的特有问题就是如何将中文的字串分割为合理的词语序列，即中文分词。中文分词是句法分析等深层处理的基础，也是机器翻译、信息检索和信息抽取等应用的重要环节。而词性标注就是根据句子上下文中的信息给句中的每个词一个正确的词性标记。

自动问答系统的准确率为自动问答系统做出正确应答的次数除以总共的应答次数。例如用户向机器人输入了 100 个句子，机器人做出了 100 次应答，其中有 20 次是正确的，那么这个机器人系统的准确率就是 20％。

自动问答系统的召回率为自动问答系统做出正确应答的次数除以知识库中存在正确应答的次数。例如用户向机器人输入了 100 个句子，机器人做出了 100 次应答，其中有 20 次是正确的，但是用户输入的 100 个句子中，知识库中只有其中 25 个句子的正确应答存在，那么这个机器人系统的招回率就是 80％。

下面举例说明采用简单的模式匹配技术实现的中文自动问答系统的缺点。

假设自动问答系统的知识库中存在以下两组问答语句对，每组都包括一个用户输入的自然语言句子（以下简称用户句子）和系统应答。

第一组：

用户句子：你出生在深圳吗？

系统应答：是啊，你怎么知道的？

第二组：

用户句子：你出生在北京吗？

系统应答：不对，我出生在深圳。

当用户输入“你出生在深圳吗？”或是“你出生在北京吗？”的时候，应答都是正确的。但是当用户输入“你出生在上海吗？”，自动问答系统就无法找到匹配的用户句子，从而返回了错误的应答（可能是系统缺省的应答）。但是实际上，第二组中的系统应答才是用户输入的正确应答。

因为可以替换“上海”的词非常多，所以上述问题也无法通过增加更多的问答语句对来解决。另外，将“北京”替换为任意词通配符也不可行，因为用户可能会输入“你出生在76年吗？”，同样会匹配成功，导致应答出错。

综上所述，简单的模式匹配技术并不适合中文自动问答系统，导致中文知识库的覆盖面窄，系统的准确率和招回率都很低，会对用户体验造成伤害。

发明内容

本发明所解决的技术问题是提供一种自动问答系统，能够提高中文知识库的内容覆盖面，同时显著的提高模式匹配的准确率和招回率。

本发明的技术方案如下：

一种自动问答方法，包括：

（1） 将输入的文字信息进行切分；

（2） 根据切分的结果进行查找；

（3） 用相匹配的查找结果刷新结果集合；

（4） 判断输入的文字信息匹配是否完成；

（5） 当输入的文字信息匹配完成，依据结果集合选择应答。

优选的，推理知识库中存储有所述库存文字信息。

优选的，所述步骤（2）中，库存文字信息经过分词和词性标注处理。

优选的，所述步骤（1）具体为，中文分词和词性标注模块对输入的文字信息进行切分处理，输出文字信息的分词和词性标注信息。

优选的，步骤（2）具体为，根据输入文字信息的分词和词性标注信息，在结果集合内查找具有相同分词的库存文字信息；步骤（4）具体为，当输入文字信息匹配没有完成，提取输入文字信息的下一个分词作为查找依据，并将结果集合作为查找目标，执行步骤（2）。

优选的，步骤（2）具体为：

根据输入文字信息的分词和词性标注信息，在结果集合内查找具有指定词性通配符的库存文字信息；步骤（4）具体为，当输入文字信息匹配没有完成，提取输入文字信息的下一个分词作为查找依据，并将结果集合作为查找目标，执行步骤（2）。

优选的，步骤（2）具体为：

根据输入文字信息的分词和词性标注信息，在结果集合内查找指定任意词通配符的库存文字信息；步骤（4）具体为，当输入文字信息匹配没有完成，提取输入文字信息的下一个分词作为查找依据，并将结果集合作为查找目标，执行步骤（2）。

优选的，步骤（2）具体为：

根据输入文字信息的分词和词性标注信息，查找具有相同分词的

库存文字信息，并按照设定的分值积分；同时，根据输入文字信息的分词和词性标注信息，在结果集合内中查找指定词性通配符的库存文字信息，并按照设定的分值积分；步骤（4）具体为，当输入文字信息匹配没有完成，提取文字信息的下一个分词作为查找依据，并将结果集合作为查找目标，执行步骤（2）。

优选的，步骤（2）还包括:

根据输入文字信息的分词和词性标注信息，在结果集合内查找指定任意词通配符的库存文字信息，并按照设定的分值积分；步骤（4）具体为，当输入文字信息匹配没有完成，提取文字信息的下一个分词作为查找依据，并将结果集合作为查找目标，执行步骤（2）。

优选的，步骤（2）进一步包括:

当没有查找到相匹配的库存文字信息时，在推理知识库中的缺省应答中随机选择一个作为应答，发送并结束。

本发明的另一个技术方案如下:

一种自动问答系统，包括::

网络接口模块，接收输入的文字信息， 将应答发送;

分词和词性标注模块，对输入的文字信息进行分词和词性标注，将切分出来的分词及其词性标注信息发送;

推理模块，根据分词及其词性标注信息查找对应的应答，将应答发送到所述的网络接口模块。

优选的，所述分词和词性标注模块调用计算语言知识库中的分词和词性标注方法。

优选的，所述计算语言知识库中还存储有词语使用频率的统计数据和词库。

优选的，所述推理模块调用推理知识库中与文字信息相对应的应答。

优选的，所述推理知识库存储有库存文字信息，所述库存文字信息经过分词或者词性标注处理。

优选的，所述库存文字信息包含分词、任意词通配符或者词性通配符。

本发明能够应用于不同的语言，特别是针对中文用词灵活、句法复杂多变的特点，在自动问答系统中利用中文词性通配符可以提高中文知识库的内容覆盖面，减少建库的工作量，同时显著的提高基于模式匹配的中文自动问答系统的准确率和招回率,从而提升用户的体验。

附图说明

图 1 是自动问答方法的操作流程图;

图 2 是自动问答系统的结构示意图。

具体实施方式

下面参照图 1，对中文的自动问答方法作详细描述。

本技术方案只给出了一个具体的实施例，实际应用时可以选择不同的模式匹配方法来使用词性通配符。

步骤 S001：接收端收到用户输入的文字信息。本优选实施例中，自动问答系统 100 通过网络接口模块 101 接收用户输入的文字信息。

步骤 S002：对接收到的文字信息进行切分处理，输出一系列的词和词性标注信息。本优选实施例中，中文分词和词性标注模块 102 调用计算语言知识库中的分词和词性标注方法，对接收到的文字信息进行切分处理，输出文字信息的分词和词性标注信息。

步骤 S003：从这些分词和词性信息中，取出当前分词及词性标注

信息作为查找依据。本优选实施例中，取出第一个分词和词性标注信息作为查找依据。

步骤 S004：依据当前分词和词性标注信息进行查找。

本优选实施例中，推理模块 104 从第一个词开始，依据第一个分词和词性标注信息，在推理知识库 105 内进行查找，并将找到的结果作为结果结合。推理知识库 105 内存储有应答和经过分词处理的库存文字信息，该库存文字信息包含分词、任意词通配符或者词性通配符，并且每个分词可以对应多个应答。

查找的目标是找到以下三类特征的库存文字信息，以及与该库存文字信息相对应的应答：

第一、 推理知识库 105 内的库存文字信息在当前位置的分词与用户输入的文字信息的第一个分词相同。给选中的库存文字信息记分，每选中一次，将此类库存文字信息的分值增加 1（初始值为 0）。

第二、 库存文字信息在当前位置出现了词性通配符，而且该词性通配符所指定的词性与用户输入的文字信息的当前分词的词性相同。给选中的库存文字信息记分，每选中一次，将此类用户句子的分值增加 0.5。

第三、 库存文字信息在当前位置出现了任意词通配符。给选中的库存文字信息记分，每选中一次，将此类用户句子的分值增加 0.2。

上述三类匹配模式可以任意选取其一，也可以选取几个进行组合，作为匹配模式。本优选实施例中，上述三类匹配模式同时选用，并将依照三类匹配模式选取的库存文字信息都放入结果集合。

本发明中，在推理知识库 105 内对用户输入的文字信息（例如句子）增加了词性通配符，表示所有具有指定词性的词。自动问答系统在收到用户输入的句子后首先进行分词和词性标注，然后再转交给推理模块 104。当推理模块 104 对用户输入的句子和推理知识库 105 内的用户句子进行模式匹配的时候，词性通配符可以和具有指定词性的任意词匹配成功，但是，如果用户输入的句子中的词和知识库中其他用户句子的词完全匹配，则词性通配符的优先级低于完全匹配的优先级。通过本方法可以显著提高基于模式匹配的中文自动问答系统的准确率和招回率

步骤 S005：如果找到与用户输入的文字信息相匹配的库存文字信息和应答，则将这些库存文字信息和应答作为当前的结果集合。由于后续找到的库存文字信会不断刷新上一个结果集合，所以结果集合能够及时得到更新。本优选实施例中，随着匹配的进行和积分的累加，该结果集合的库存文字信息的数量是在不断缩小，因此应答的正确率在不断地提高。

推理知识库 105 内还存储有缺省应答，如果上述查找都失败，则推理模块 104 认为推理知识库 105 内没有与用户输入的文字信息相符的应答，系统会从推理知识库 105 内调用缺省应答，随机选择一个（步骤 S009），返回给用户（步骤 S010）。

步骤 S006：判断用户输入的文字信息是否已经匹配完毕。本优选实施例中，该步骤由推理模块执行，以便于及时判断匹配是否完成。

步骤 S007：如果用户输入的文字信息没有匹配完毕，则提取下一个分词和词性标注信息作为查找依据，执行步骤 S004，继续上述查找过程，直到全部匹配成功，或者中途匹配失败。本优选实施例中，推

理模块 104 认为没有匹配完毕，则进行 n + 1 操作，将下一个分词作为查找的依据。

步骤 S008：如果用户输入的文字信息已经匹配完成，则从结果集合中相匹配的应答中随机选取一个，返回给用户（步骤 S010）。

本优选实施例中，推理模块 104 判断已经匹配完成，从结果集合中选择积分值最高的应答发送到网络接口模块 101，通过网络接口模块 101 发送给用户。

步骤 S009：如果没有在推理知识库 105 中找到匹配的库存文字信息，则推理模块 104 将从推理知识库 105 中的缺省应答中随机选取一个，作为应答。

步骤 S010：将接收到的应答发送给用户。

本优选实施例中，网络接口模块 101 接收推理模块 104 发送的应答，并将该应答发送给用户。

本发明中，利用中文的词性通配符提高了推理知识库 105 的内容覆盖面，减少了建库的工作量，同时能够显著的提高基于模式匹配的自动问答方法的准确率和招回率，从而提升用户的体验，是一项非常有意义的创新。

参考背景技术中的例子，在支持中文词性通配符的本发明的推理知识库 105 中，构造了以下两组问答语句对：

第一组：

用户输入的文字信息：你出生在深圳吗？

系统响应的应答：是啊，你怎么知道的？

第二组：

用户输入的文字信息：你出生在 POSnsPOS 吗？

系统响应的应答：不对，我出生在深圳。

其中 POSnsPOS 是本实施例中采用词性通配符表示的方式，其中 POS 是词性信息的起止标记，而 ns 是表示方位的名词词性。

当用户输入“你出生在深圳吗？”的时候，与第一组的用户句子匹配成功，系统向用户响应“是啊，你怎么知道的？”；

当用户输入“你出生在北京吗？”或“你出生在上海吗？”的时候，都与第二组中用户句子匹配成功，系统向用户响应“不对，我出生在深圳。”

实际上，只要用户输入的是类似北京和上海的、任何具备 ns 词性的词，都可以与第二组问答语句对匹配成功；但是类似“76 年”这种词不具备 ns 词性，所以不会被误匹配为第二组问答语句对。

本发明中，还可以选择不同的模式匹配方法来使用词性通配符，用于提高基于模式匹配的准确率和招回率，例如，不对用户输入的句子进行逐词的匹配，而是打乱词的顺序直接匹配。

下面参照图 2 对本发明的优选实施例作详细描述。

不同的语言有不同的语法，使得词之间有不同的匹配模式。本优选实施例中，系统选用中文作为识别目标。

选用中文的自动问答系统 100 包括网络接口模块 101、中文分词和词性标注模块 102、推理模块 104，以及计算语言知识库 103 和推理知识库 105。

网络接口模块 101 负责接收用户输入的句子，并发送给中文分词和词性标注模块 102。

中文分词和词性标注模块 102 调用计算语言知识库 103 中的分词和词性标注方法，对用户输入的文字信息进行中文分词和词性标注，

然后将所有切分出来的词及其词性标注信息提交给推理模块 104。

推理模块 104 根据分词和词性标注模块 104 输出的词及其词性标注信息在推理知识库 105 内查找对应的应答，当存储在推理知识库 105 内的库存文字信息包含词性通配符的时候，该词性通配符可以和用户输入的句子中具有指定词性的任意词匹配成功，从而继续后面的匹配。

本优选实施例中，计算语言知识库 103 内存储的是中文分词和词性标注所必需的信息，还包括词典以及词频等各种统计数据，该计算语言知识库 103 可以根据实际需要进行升级，及时将新的分词和词性标注方法补入。

推理知识库 105 内存储的是库存文字信息，该库存文字信息为用户可能输入的文字信息。推理知识库 105 内还存储有对应这些库存文字信息的应答，其中每个库存文字信息都经过分词处理，可以对应一个或多个应答。推理知识库 105 由推理模块 104 在系统启动的时候读入内存，并在收到中文分词和词性标注的命令和信息后与之进行匹配。存储在推理知识库 105 中的库存文字信息除了可以包括具体的词和任意词通配符之外，还可以包括词性通配符，用来表示所有具有指定词性的词，另外，推理知识库 105 中还存储有缺省应答。

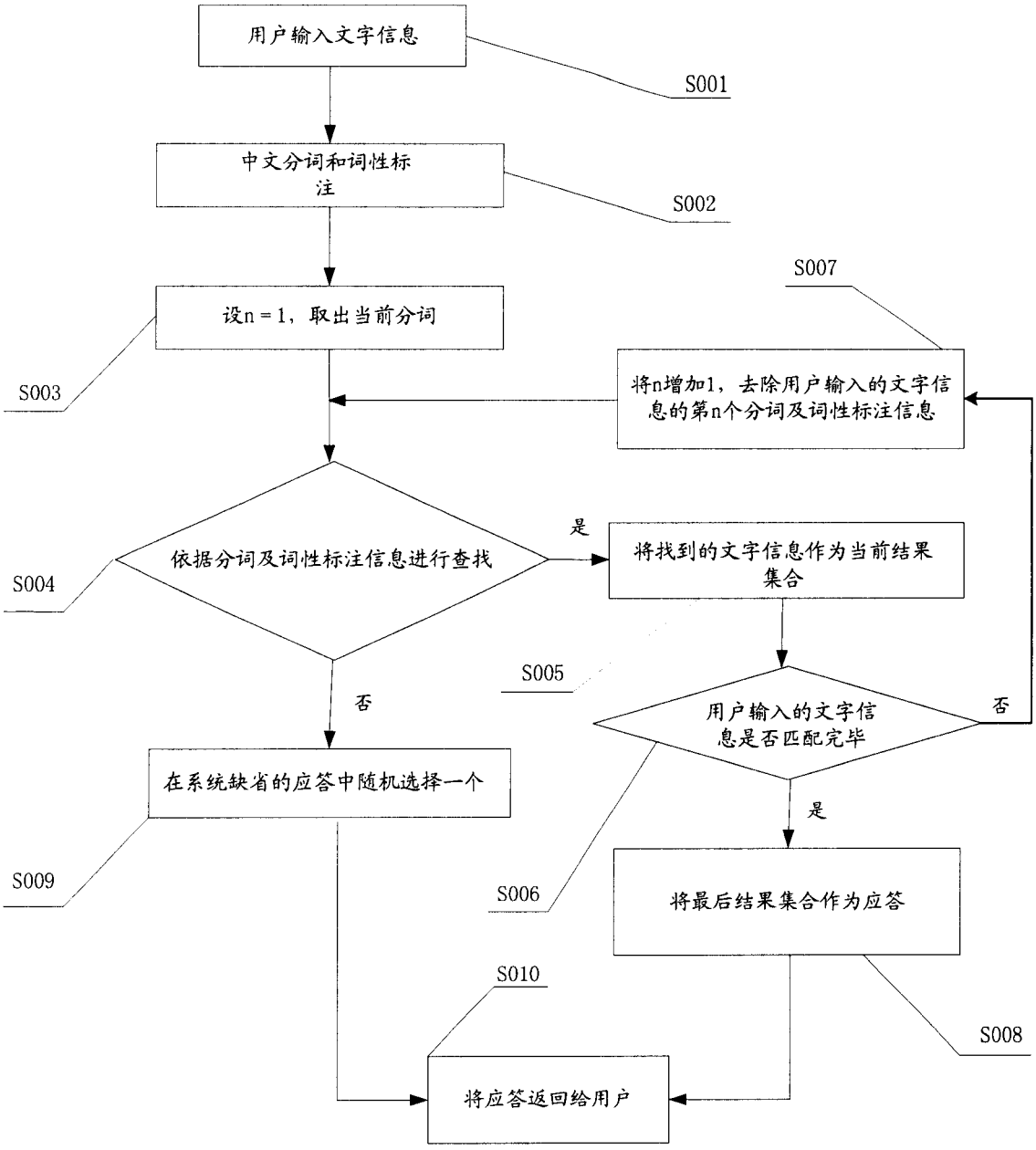

用户输入文字信息
S001
中文分词和词性标注
S002
设n = 1，取出当前分词
S003
S007
将n增加1，去除用户输入的文字信息的第n个分词及词性标注信息
依据分词及词性标注信息进行查找
S004
是
将找到的文字信息作为当前结果集合
S005
用户输入的文字信息是否匹配完毕
否
是
否
在系统缺省的应答中随机选择一个
S009
S006
将最后结果集合作为应答
S010
S008
将应答返回给用户

图 1

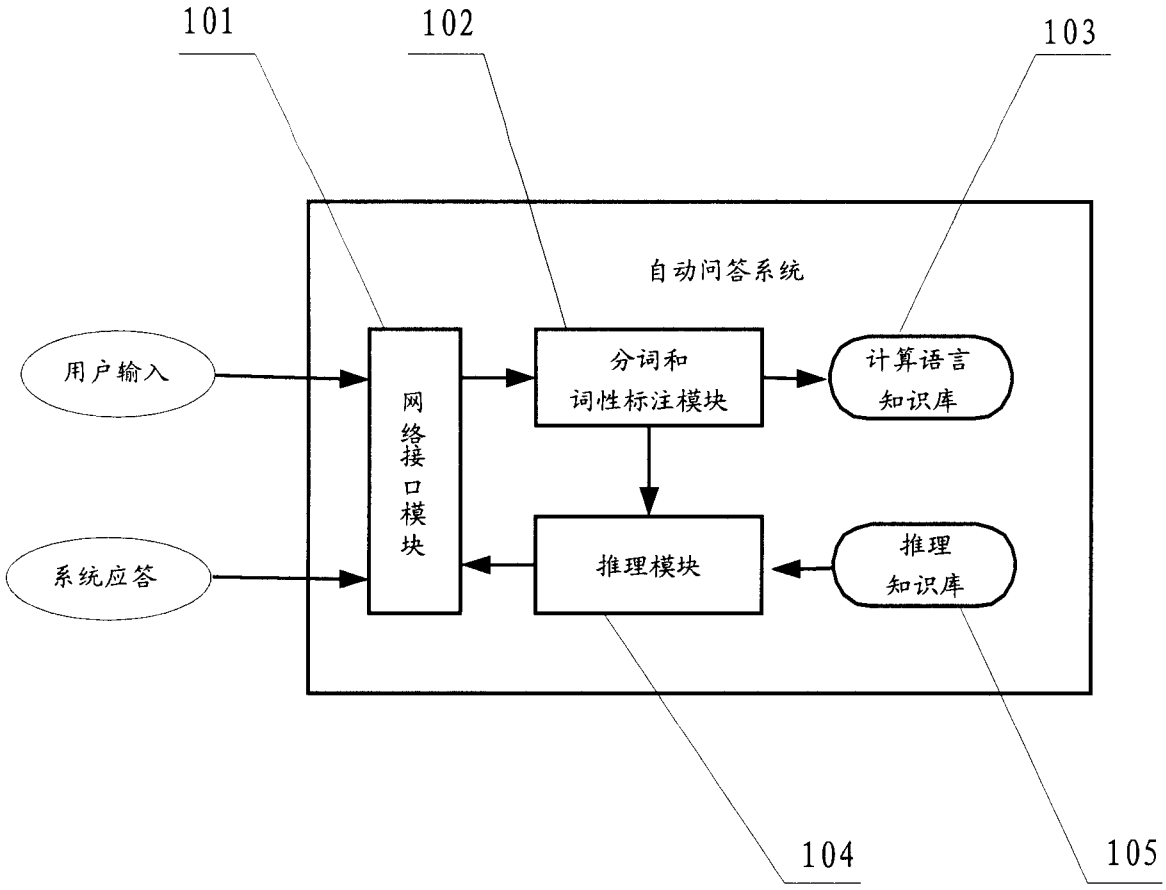

101
102
103
自动问答系统
用户输入
网络接口模块
分词和词性标注模块
计算语言知识库
系统应答
推理模块
推理知识库
104
105

图 2