[19] 中华人民共和国国家知识产权局

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

[12] 发明专利申请公布说明书

[21] 申请号 200810226947.9

[43] 公开日 2009 年 4 月 8 日

[11] 公开号 CN 101404035A

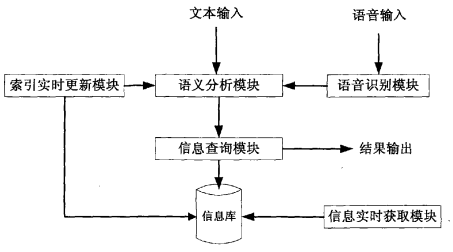
[22] 申请日 2008.11.21
[21] 申请号 200810226947.9
[71] 申请人 北京得意音通技术有限责任公司
地址 100085 北京市海淀区上地信息路 2 号 D 栋 505 室
共同申请人 清华大学
[72] 发明人 邬晓钧 郑 方 潘胜逖 苏保飞
[74] 专利代理机构 北京清亦华知识产权代理事务所
代理人 罗文群
权利要求书 1 页 说明书 3 页 附图 1 页

[54] 发明名称

一种基于文本或语音的信息搜索方法

[57] 摘要

本发明涉及一种基于文本或语音的信息搜索方法，属于网络信息查询技术领域。 首先，用户将搜索目标以输入至语义分析模块，得到搜索目标；对搜索目标进行语义处理后得到用户搜索关键词，并发送至信息查询模块及索引实时更新模块；根据搜索关键词，对信息库进行搜索，得到用户所需信息；信息实时获取模块通过网络获取最新信息，并将其发送至信息库中；索引实时更新模块根据最新信息，对语义分析模块中的信息进行实时更新。 本发明方法的优点是可以让用户享受更为人性化、更快捷的方式进行信息搜索。



1、一种基于文本或语音的信息搜索方法，其特征在于该方法包括以下步骤：

（1）用户将搜索目标输入至语义分析模块，或将搜索目标以语音方式输入至语音识别模块，语音识别模块对搜索目标进行识别后得到正确的搜索目标，并将正确的搜索目标发送至语义分析模块；

（2）语义分析模块对上述搜索目标进行语义处理后得到用户搜索关键词，并将用户搜索关键词发送至信息查询模块及索引实时更新模块；

（3）信息查询模块根据上述搜索关键词，对信息库进行搜索，得到用户所需信息；

（4）信息实时获取模块通过网络获取最新信息，并将最新信息发送至信息库中；

（5）索引实时更新模块根据信息库的上述最新信息，对语义分析模块中的信息进行实时更新。

一种基于文本或语音的信息搜索方法

技术领域

本发明涉及一种基于文本或语音的信息搜索方法，属于网络信息查询技术领域。

背景技术

伴随互联网的深入发展，搜索引擎及搜索服务的出现和演进深刻影响了人们的生活方式，但目前，无论是互联网还是移动互联网搜索技术，都仅仅是基于文本的方式进行信息查询，其缺点是：不能理解用户输入的自然语言，只能做简单的关键词匹配或理解。另一方面，由于互联网的普及，人们被信息爆炸、信息垃圾所困扰。虽然通过搜索引擎、目录、人工编辑的社区等工具，人们可以获得一定的帮助，但是这些工具的准确性和方便性仍很不足够，急迫需要智能化、精确化、专业化、个性化的，以用户为中心的智能信息服务。

发明内容

本发明的目的是提出一种基于文本或语音的信息搜索方法，用多种方法输入查询要求，给用户提供方便、快捷而多样化的查询方式，并通过语义分析正确理解用户的真正需求；同时，本发明从网络实时获取信息，保证用户查询信息的实时性、有效性及准确性。

本发明提出的基于文本或语音的信息搜索方法，包括以下步骤：

（1）用户将搜索目标输入至语义分析模块，或将搜索目标以语音方式输入至语音识别模块，语音识别模块对搜索目标进行识别后得到正确的搜索目标，并将正确的搜索目标发送至语义分析模块；

（2）语义分析模块对上述搜索目标进行语义处理后得到用户搜索关键词，并将用户搜索关键词发送至信息查询模块及索引实时更新模块；

（3）信息查询模块根据上述搜索关键词，对信息库进行搜索，得到用户所需信息；

（4）信息实时获取模块通过网络获取最新信息，并将最新信息发送至信息库中；

（5）索引实时更新模块根据信息库的上述最新信息，对语义分析模块中的信息进行实时更新。

本发明提出的基于文本或语音的信息搜索方法，其优点是：

1、本发明提出的基于文本或语音的信息搜索方法，为用户提供了多种的信息查询方式，用户既能通过互联网输入文字进行信息查询，也可以通过语音输入的方式了解最新信息，例如将本发明信息查询方法用于农业市场行情时，如价格参数、产地、交易市场等，用户只要说一句“今天大白菜多少钱一斤”，系统就将会实时反馈用户有关该问题的详细

答案，包括大白菜的最高价、最低价、均价等信息，这样可以让用户享受更为人性化，更快捷的方式进行农业市场信息搜索。

2、本发明的信息搜索方法，用自然语言与计算机进行交流，获取合适的信息，使人们能够方便、快捷地获取所需信息，即使用户的信息输入中带有错别字，系统也能自动识别并纠正，例如将本发明信息查询方法用于餐饮领域时，用户输入“我想在上地附近吃大渣蟹”（应为“大闸蟹”），系统也能完全准备的理解用户的输入并查找出正确的信息。

3、本发明的信息获取方法，采用从互联网实时抓取网页信息，并且系统能 7*24 小时运行不间断获取信息，同时本发明中系统抓取信息的一个重要特点是，即使网站页面信息是动态生成、需要复杂的脚本验证或登录才能看到的信息，本系统程序也能自动模拟实际用户访问网站的方式，最终将对方网站的信息抓取下来并保存在信息库中。

附图说明

图 1 是本发明方法的流程示意图。

具体实施方式

本发明提出的基于文本或语音的信息搜索方法，如图 1 所示，首先用户将搜索目标输入至语义分析模块，或将搜索目标以语音方式输入至语音识别模块，语音识别模块对搜索目标进行识别后得到正确的搜索目标，并将正确的搜索目标发送至语义分析模块；语义分析模块对上述搜索目标进行语义分析后得到用户搜索关键词，并将用户搜索关键词发送至信息查询模块及索引实时更新模块；信息查询模块根据上述搜索关键词，对信息库进行搜索，得到用户所需信息；信息实时获取模块通过网络获取最新信息，并将最新信息发送至信息库中；索引实时更新模块根据信息库的上述最新信息，对语义分析模块中的信息进行实时更新。

下面结合附图，对本发明内容作详细的阐述。图 1 示出了一种基于文本或语音的信息搜索方法示意图，主要涉语音识别模块、语义分析模块、信息查询模块、实时索引更新模块和信息实时获取模块。

语义分析模块：用户将搜索目标提交到本发明系统中的语义分析模块或语音识别模块，经过语义处理后，得到用户搜索关键词，并且语义分析模块具有一定的自动纠错功能，如“大扎蟹”能够识别为“大闸蟹”；该模块的实现方法包括以下步骤：

1）从语法配置文件中读入基于语义类的上下文无关增强文法；

2）对用户输入的句子进行分词；

3）对分词结果进行句法分析；

4）取最优的句法分析结果进行语义分析，得到用户最终的搜索关键词信息。

上述基于语义类的上下文无关增强文法，其具体实现过程包括以下步骤：

a）根据领域任务定义文法中所有的终结符、非终结符和规则

b) 终结符为按语义分类的关键词，关键词可包含阿拉伯数字和英文字母，每个关键词都有相应的拼音；

c) 每一条规则都被赋以一个优先级别；

d) 一个优先级的规则集合可以是词法分析的或非词法分析的；

e) 所说的规则与语义直接关联，每一条规则都对应一个语义分析函数。

语音识别模块：对用户的语音信息进行识别并转换成文本信息，然后将文本信息提交至语义分析模块；该模块的实现方法包括以下步骤：

1）系统初始化加载声音模型及中文词库文件；

2）接收用户语音信息，提取用户声音特征；

3）建立用户声音模型，用代表语义信息的声音模型给用户声音模型信息打分；

4）检出语义单元中最大的输出信息；

5）最后，从最大的语义单元输出信息中获取文本信息。

信息查询模块：接收从语义分析模块传送的搜索关键词，以此为作为查询信息库的具体条件，得到用户所需要的信息。

实时索引更新模块：实时索引更新模块自动将信息库中的最新信息实时更新到语义分析模块的关键词列表中，从而保证用户查询数据的有效性、准确性。

信息实时获取模块：抓取互联网特定领域（例如餐饮、租房）网站的网页数据，并自动分类整理到信息库中。

以下对本方法的具体过程说明如下：通过各类硬件终端将文本（例如能连接到互联网的电脑或手机输入）或语音信息（例如电话）输入到语义分析模块。用户可以通过各类硬件终端将文本（例如能连接到互联网的电脑或手机输入）提交至本发明系统中的语义分析模块。用户也可以首先将语音信息（例如电话）提交至本发明系统中的语音识别模块，语音识别模块将语音信息转换为文本信息后，再将文本信息发送至语义分析模块。

语义分析模块接收来自上述两种情形的信息，得到的搜索关键词，并将搜索关键词信息提交至信息查询模块。

信息查询模块通过关键词信息从信息库中找出结果，并组织成友好的数据呈现反馈给用户。

本发明方法中的信息实时获取模块 7*24 小时不间断运行抓取互联网网页信息，经自动整理、分类后更新至本系统的信息库中，同时索引更新模块实时将信息库中新增的数据更新到语义分析模块的关键词列表中，以保证用户搜索的准确与实时。

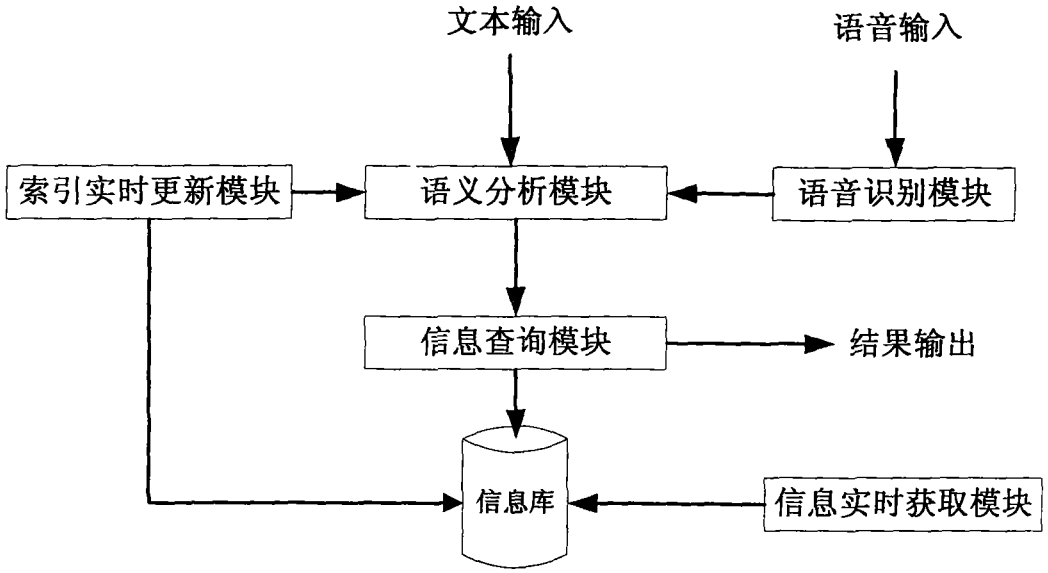


图1