(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 102915380 A
(43)申请公布日 2013.02.06

(21)申请号 201210469129.8

(22)申请日 2012.11.19

(71)申请人 北京奇虎科技有限公司
地址 100088 北京市西城区新街口外大街28号D座112室(德胜园区)
申请人 奇智软件(北京)有限公司

(72)发明人 李天华

(74)专利代理机构 北京市浩天知识产权代理事务所 11276
代理人 靳春鹰 刘云贵

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

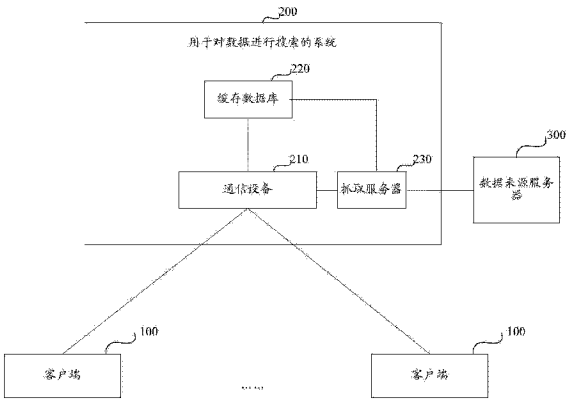
权利要求书 2 页 说明书 10 页 附图 2 页

(54)发明名称

用于对数据进行搜索的方法和系统

(57)摘要

本发明公开了一种用于对数据进行搜索的方法和系统，该系统包括：通信设备、缓存数据库、抓取服务器以及搜索服务器，其中，当在所述缓存数据库中按照预设的匹配规则查找到的与所述搜索词相匹配的关键词及其对应的查询结果的数量少于预设数量时，将获取的搜索服务器的查询结果发送给所述客户端，其中，所述搜索服务器的查询结果用于作为所述缓存数据库的查询结果的补充。本发明的用于对数据进行搜索的方法和系统可以解决现有技术中同时设置信息数据库和索引数据库两个数据库时需要用复杂的算法才能完成数据匹配过程，导致用户等待时间过长的问题，能够取得根据预设的缓存数据库和匹配规则迅速查找到匹配的数据的有益效果。



CN 102915380 A

1. 一种用于对数据进行搜索的方法，包括：

预先提取关键词列表，通过访问外部的数据来源服务器获取所述关键词列表中每一关键词对应的查询结果，将每一关键词及其对应的查询结果关联存储在缓存数据库中；

获取客户端发送的包含搜索词的搜索请求，将所述搜索请求分发到所述缓存数据库中，在所述缓存数据库中按照预设的匹配规则查找与所述搜索词相匹配的关键词及其对应的查询结果；

将所述关键词对应的查询结果发送给所述客户端；

其中，所述获取客户端发送的包含搜索词的搜索请求的步骤之后，进一步包括：

将所述搜索请求分发到搜索服务器，获取所述搜索服务器从外部的数据来源服务器查找到的所述搜索词对应的查询结果；

当在所述缓存数据库中按照预设的匹配规则查找到的与所述搜索词相匹配的关键词及其对应的查询结果的数量少于预设数量时，该方法进一步包括：

将获取的搜索服务器的查询结果发送给所述客户端，其中，所述搜索服务器的查询结果用于作为所述缓存数据库的查询结果的补充。

2. 如权利要求 1 所述的方法，所述预设的匹配规则包括：自然语言处理分析规则，和 / 或正则表达式规则。

3. 如权利要求 1 或 2 所述的方法，所述缓存数据库中的关键词及其对应的查询结果以键值对的方式存储。

4. 如权利要求 3 所述的方法，其中，所述缓存数据库中的关键词按照预设的分类存储，则所述将每一关键词及其对应的查询结果关联存储在缓存数据库中进一步包括：

确定每一关键词所属的分类；

针对每一关键词，对该关键词及其所属的分类进行加密运算，将得到的加密结果作为键，将该关键词对应的查询结果作为所述键对应的值。

5. 如权利要求 1-4 中任一个所述的方法，当所述缓存数据库中的关键词按照预设的分类存储时，所述搜索请求中进一步包括搜索词所属的分类；

则查找与所述搜索词相匹配的关键词时，在分类与所述搜索词所属分类相同的关键词中查找。

6. 如权利要求 1-5 中任一个所述的方法，所述关键词对应的查询结果是包含所述关键词的网页对应的数据快照，所述数据快照用于存储网页的裸数据或 html 数据。

7. 如权利要求 1-6 中任一个所述的方法，当所述关键词的查询结果与地域相关时，所述缓存数据库中存储的所述关键词的查询结果进一步包括与各个地域相对应的查询结果，

则在缓存数据库中查找所述关键词对应的查询结果进一步包括：根据所述搜索请求中携带的 IP 地址确定所述客户端所处的地域，在缓存数据库中查找与所述地域相对应的查询结果。

8. 一种用于对数据进行搜索的系统，包括：通信设备、缓存数据库以及抓取服务器，其中，

抓取服务器，适于预先提取关键词列表，通过访问外部的数据来源服务器获取所述关键词列表中每一关键词对应的查询结果，将每一关键词及其对应的查询结果关联存储在所述缓存数据库中；

通信设备，适于获取客户端发送的包含搜索词的搜索请求，将所述搜索请求分发到所述缓存数据库中，在所述缓存数据库中按照预设的匹配规则查找与所述搜索词相匹配的关键词及其对应的查询结果，还适于将所述查询结果发送给所述客户端；

搜索服务器，适于从外部的数据来源服务器查找搜索词对应的查询结果；

则所述通信设备进一步适于将所述搜索请求分发到所述搜索服务器，获取所述搜索服务器查找到的所述搜索词对应的查询结果；以及

当在所述缓存数据库中按照预设的匹配规则查找到的与所述搜索词相匹配的关键词及其对应的查询结果的数量少于预设数量时，将获取的搜索服务器的查询结果发送给所述客户端，其中，所述搜索服务器的查询结果用于作为所述缓存数据库的查询结果的补充。

9. 如权利要求 8 所述的系统，所述预设的匹配规则包括：自然语言处理分析规则，和 / 或正则表达式规则。

10. 如权利要求 8 或 9 所述的系统，所述缓存数据库适于将所述关键词及其对应的查询结果以键值对的方式存储。

11. 如权利要求 10 中任一个所述的系统，所述缓存数据库中的关键词按照预设的分类存储，则所述缓存数据库进一步适于：

确定每一关键词所属的分类；

针对每一关键词，对该关键词及其所属的分类进行加密运算，将得到的加密结果作为键，将该关键词对应的查询结果作为所述键对应的值。

12. 如权利要求 8-11 中任一个所述的系统，所述关键词对应的查询结果是包含所述关键词的网页对应的数据快照，所述数据快照用于存储网页的裸数据或 html 数据。

13. 如权利要求 8-12 中任一个所述的系统，当所述关键词的查询结果与地域相关时，所述缓存数据库中存储的所述关键词的查询结果进一步包括与各个地域相对应的查询结果，

则所述查找模块进一步适于：根据所述搜索请求中携带的 IP 地址确定所述客户端所处的地域，在预先设置的缓存数据库中查找与所述地域相对应的查询结果。

14. 如权利要求 8-13 中任一个所述的系统，所述抓取服务器按照预设的频率对所述关键词列表中的关键词和 / 或所述关键词对应的查询结果进行更新。

用于对数据进行搜索的方法和系统

技术领域

[0001] 本发明涉及搜索领域，具体涉及一种用于对数据进行搜索的方法和系统。

背景技术

[0002] 目前，随着计算机技术的发展和互联网用户规模的不断扩大，越来越多的互联网用户使用个人计算机通过互联网获得各种各样所需的信息。同时，为互联网用户提供信息服务的网站也越来越多，互联网网页的数量每天都在以惊人的速度增长，互联网信息呈现出爆发式的增长。因此，对于用户来说，经常需要通过一定的手段(比如，通过搜索引擎服务)，才能在浩如烟海的互联网信息中迅速定位最适合自己需求的网站或者需要的信息。

[0003] 搜索引擎的服务器通常需要根据用户输入的搜索词去数据来源服务器搜索对应的结果，并将结果提供给用户。这里提到的数据来源服务器是指第三方服务器，用于存储原始的网页资源。

[0004] 采用上述的搜索引擎服务，虽然可以满足用户搜索数据的需求，但是，由于每次都需要去数据来源服务器查询，因此，延长了搜索引擎搜索时耗费的时间，导致用户等待时间较长。

发明内容

[0005] 鉴于上述问题，提出了本发明以便提供一种克服上述问题或者至少部分地解决上述问题的用于对数据进行搜索的方法和系统。

[0006] 依据本发明的一个方面，提供了一种用于对数据进行搜索的方法，包括以下步骤：预先提取关键词列表，通过访问外部的数据来源服务器获取关键词列表中每一关键词对应的查询结果，将每一关键词及其对应的查询结果关联存储在缓存数据库中；获取客户端发送的包含搜索词的搜索请求，将搜索请求分发到缓存数据库中，在缓存数据库中按照预设的匹配规则查找与搜索词相匹配的关键词及其对应的查询结果；将关键词对应的查询结果发送给客户端；其中，所述获取客户端发送的包含搜索词的搜索请求的步骤之后，进一步包括：将所述搜索请求分发到搜索服务器，获取所述搜索服务器从外部的数据来源服务器查找到的所述搜索词对应的查询结果；当在所述缓存数据库中按照预设的匹配规则查找到的与所述搜索词相匹配的关键词及其对应的查询结果的数量少于预设数量时，该方法进一步包括：将获取的搜索服务器的查询结果发送给所述客户端，其中，所述搜索服务器的查询结果用于作为所述缓存数据库的查询结果的补充。

[0007] 依据本发明的另一方面，提供了一种用于对数据进行搜索的系统，包括：通信设备、缓存数据库以及抓取服务器，其中，抓取服务器，适于预先提取关键词列表，通过访问外部的数据来源服务器获取关键词列表中每一关键词对应的查询结果，将每一关键词及其对应的查询结果关联存储在缓存数据库中；通信设备，适于接收获取客户端发送的包含搜索词的搜索请求，将搜索请求分发到缓存数据库中，在缓存数据库中按照预设的匹配规则查找与搜索词相匹配的关键词及其对应的查询结果在缓存数据库中按照预设的匹配规则查

找与搜索词相匹配的关键词及其对应的查询结果，还适于将查询结果发送给客户端；搜索服务器，适于从外部的数据来源服务器查找搜索词对应的查询结果；则所述通信设备进一步适于将所述搜索请求分发到所述搜索服务器，获取所述搜索服务器查找到的所述搜索词对应的查询结果；以及当在所述缓存数据库中按照预设的匹配规则查找到的与所述搜索词相匹配的关键词及其对应的查询结果的数量少于预设数量时，将获取的搜索服务器的查询结果发送给所述客户端，其中，所述搜索服务器的查询结果用于作为所述缓存数据库的查询结果的补充。

[0008] 根据本发明的用于对数据进行搜索的方法和系统，可以预先设置缓存数据库以及匹配规则，并预先在缓存数据库中存储所有关键词以及每一关键词对应的查询结果，具体搜索时只需去缓存数据库中即可查找到对应的结果，无需访问数据来源服务器，由此解决了现有技术中搜索耗时过多，导致用户等待时间过长的问题，取得了直接查询缓存数据库即可迅速查找到匹配的数据的有益效果。

[0009] 上述说明仅是本发明技术方案的概述，为了能够更清楚了解本发明的技术手段，而可依照说明书的内容予以实施，并且为了让本发明的上述和其它目的、特征和优点能够更明显易懂，以下特举本发明的具体实施方式。

附图说明

[0010] 通过阅读下文优选实施方式的详细描述，各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的，而并不认为是对本发明的限制。而且在整个附图中，用相同的参考符号表示相同的部件。在附图中：

[0011] 图 1 示出了根据本发明一个实施例的用于对数据进行搜索的方法的流程图；

[0012] 图 2 示出了根据本发明一个实施例的用于对数据进行搜索的系统的结构图；以及

[0013] 图 3 示出了根据本发明一个实施例的查询结果的示意图。

具体实施方式

[0014] 下面将参照附图更详细地描述本公开的示例性实施例。虽然附图中显示了本公开的示例性实施例，然而应当理解，可以以各种形式实现本公开而不应被这里阐述的实施例所限制。相反，提供这些实施例是为了能够更透彻地理解本公开，并且能够将本公开的范围完整的传达给本领域的技术人员。

[0015] 图 1 示出了本发明实施例提供的用于对数据进行搜索的方法的流程图，如图 1 所示，该方法包括以下步骤：

[0016] 步骤 S110：预先提取关键词列表，通过访问外部的数据来源服务器获取关键词列表中每一关键词对应的查询结果，将每一关键词及其对应的查询结果关联存储在缓存数据库中。

[0017] 步骤 S120：获取客户端发送的包含搜索词的搜索请求，将该搜索请求分发到缓存数据库中，在缓存数据库中按照预设的匹配规则（例如自然语言处理分析规则，和 / 或正则表达式规则）查找与搜索词相匹配的关键词及其对应的查询结果。

[0018] 可选地，为了便于查找，缓存数据库中存储的关键词以及每一关键词对应的查询结果以键值对的方式存储，且关键词对应的查询结果可以是包含该关键词的网页对应的数

据快照，该数据快照用于存储网页的裸数据或 html 数据。

[0019] 另外，缓存数据库中的所有关键词还可以进一步按照预设的分类进行存储，则客户端发送的搜索请求中进一步包括搜索词所属的分类。相应地，在查找与搜索词相匹配的关键词时，只需在分类类别与搜索词所属类别相同的关键词中查找，从而进一步简化了查找时的工作量，节约了查找时间。

[0020] 而且，当关键词的查询结果与地域相关时，缓存数据库中存储的关键词的查询结果还可以进一步包括与各个地域相对应的查询结果，这样，在预先设置的缓存数据库中查找关键词对应的查询结果时进一步包括：根据客户端发送的搜索请求中携带的 IP 地址来确定该客户端所处的地域，并在缓存数据库中查找与该地域相对应的查询结果，从而可以为客户端发送与其所处的地域相符合的查询结果。

[0021] 步骤 S130：将步骤 S120 中查找到的关键词对应的查询结果发送给该客户端。

[0022] 通过本发明的用于对数据进行搜索的方法，可以预先设置缓存数据库以及匹配规则，并在缓存数据库中存储所有关键词以及每一关键词对应的查询结果，因此，根据预设的缓存数据库和匹配规则可以迅速查找到匹配的数据。

[0023] 下面以一个优选实施例详细描述一下本发明提供的用于对数据进行搜索的方法。

[0024] 可选地，为了提高数据搜索的精准性，缩短搜索时间，在本优选实施例中，预先将用户可能会搜索的关键词按照一定的分类规则进行分类，相应地，在提供给用户的搜索界面中，可以针对每一类别，分别为用户提供一个搜索框。例如，可以预先将搜索词分为以下类别：生活服务、投资理财以及娱乐资讯等，这样，在搜索界面中可以进一步包括生活服务对应的搜索框、投资理财对应的搜索框，以及娱乐资讯对应的搜索框。这样，当用户需要输入搜索词进行搜索时，会先判断该搜索词属于哪一类别，然后，在该类别对应的搜索框中输入搜索词。例如，当用户要查询股票信息时，会选择投资理财对应的搜索框进行搜索，这样，由于在搜索时限定了搜索词所属的分类，搜索时仅对同一分类中的关键词进行查找，因此，既提高了查找速度，又使得查找结果更加准确，不易出现偏差。另外，还可以按照其他的分类方式进行分类，例如，按照视频、文本、图片等方式进行分类。而且，还可以进一步对一个大的分类中的数据进行细小的分类，例如，“生活服务”分类又可以进一步细分为“天气预报”、“车票预定”等，甚至“车票预定”又可以进一步细分为“飞机票预定”、“火车票预定”等，从而进一步方便查找。

[0025] 下面以生活服务这一分类为例详细描述一下本优选实施例中的用于对数据进行搜索的方法。该方法主要包括以下步骤：

[0026] 步骤一、预先提取“生活服务”这一分类中的关键词，组成关键词列表，针对该关键词列表中的每一关键词，将包含该关键词的网页所对应的 URL 与该关键词一起关联存储在该关键词列表里。

[0027] 具体地，在提取“生活服务”这一分类中的关键词时，可以根据用户的搜索频率来确定要提取的关键词，例如，将预定时段内（例如，上一星期之内）用户搜索的频率较高的搜索词筛选出来作为关键词。具体实现时，可以设定一个搜索阈值，将预定时段内的搜索次数大于该搜索阈值的搜索词筛选出来作为关键词。然后，针对每一关键词，获取包含该关键词的网页所对应的URL信息，并将该URL信息与该关键词关联存储。其中，对于每一关键词，包含该关键词的网页的数量可能是一个，也可能是多个，当网页数量为多个时，还可以进一步

判定多个网页中的内容是否重复，当多个网页中的内容重复时，只要挑选其中的一个网页的 URL 进行存储即可，这样，既可以避免因存储的数据量过大而占用存储空间过多的问题，也可以在用户搜索时缩短查询时间。

[0028] 步骤二、根据步骤一中生成的关键词列表，访问外部的数据来源服务器，获取该数据来源服务器中存储的与 URL 对应的网页数据，并根据获取的网页数据生成该网页对应的数据快照，将该数据快照作为与 URL 对应的关键词的查询结果，每一关键词及其对应的查询结果关联存储在缓存数据库中。

[0029] 具体地，网络爬虫根据关键词列表中存储的与关键词对应的 URL，到数据来源服务器中抓取与 URL 对应的网页数据，抓取后会对网页数据进行分析并拍照，形成该网页对应的数据快照。该数据快照中包含该 URL 对应的关键词，因此，将该数据快照作为该关键词对应的查询结果，与该关键词一起关联存储在缓存数据库中。其中，数据快照具体用来存储网页的裸数据或 html 数据，采用数据快照进行存储的方式具有访问速度快、便于显示的优点。

[0030] 具体存储时，为了方便查找，可以通过键值对(key-value)的方式存储，即，将关键词作为 key，将该关键词对应的查询结果(即数据快照)作为 value。或者，也可以对关键词及该关键词所属的分类进行加密运算，将得到的加密结果作为 key，将该关键词对应的查询结果作为 value。例如，假设关键词为“枫叶”，其所属的分类为图片，加密运算为 md5 运算，则只需对“枫叶”和“图片”进行 md5 运算，将得到的运算结果作为 key 即可。键值对其实是指一种数据存储方式，该数据存储方式能够通过 key-value 的模式实现直接映射，具体实现时，按照 redis 结构将键值对存储在内存中即可。通过键值对的方式进行存储的存储速度快，且读取效率高。

[0031] 步骤三、获取用户通过客户端发送的包含搜索词的搜索请求，将搜索请求分发到上述的缓存数据库中，并在上述的缓存数据库中按照预设的匹配规则查找与输入的搜索词相匹配的关键词，以及该关键词对应的查询结果。

[0032] 具体地，在接收到包含搜索词的搜索请求后，需要在缓存数据库中查找与该搜索词相匹配的关键词。本实施例中在判断搜索词与关键词是否匹配时，是根据预设的匹配规则进行判断的。

[0033] 其中，该预设的匹配规则可以是自然语言处理分析规则(简称 NLP)，或者，也可以是正则表达式规则，或者，也可以是二者的结合。其中，自然语言处理分析规则大致分为两个层面，一个是浅层分析，如分词，词性标注，通常只需对句子的局部范围进行分析处理；另一个层面是对语言进行深层的处理，需要对句子进行全局分析，在分析时通常对句法、语义以及语用这三个层次进行分析。正则表达式规则一般是通过一些具有特定含义的字符来表示匹配规则的，例如，字符“^”匹配一个输入或一行的开头，如“^a”匹配“an A”，而不匹配“An a”；字符“$”匹配一个输入或一行的结尾，如“a$”匹配“An a”，而不匹配“an A”；字符“*”匹配前面元字符 0 次或多次，如“ba*”将匹配“b”，“ba”，“baa”以及“baaa”等。通常情况下，自然语言处理分析规则主要用来解决同义词的问题，正则表达式规则主要用来处理长尾词。另外，还可以自定义一些匹配规则。例如，在本实施例中，可以预先定义“手机卫士”以及“手机卫士”都对应“360 手机卫士”。通过匹配规则的设置，可以准确地确定与用户输入的搜索词相匹配的关键词，而且，当用户输入搜索词时有少许偏差，例如，搜索词

中有一个错别字或丢掉了一个字，这时，根据自然语言处理分析规则，仍然可以确定出用户实际想要的关键词。

[0034] 通俗地说，这种按照预设的匹配规则在缓存数据库中查找与该搜索词相匹配的关键词的实现方式，就相当于预先在缓存数据库中建立了一个“词池”（即步骤二中以键值对方式存储的关键词的集合），该“词池”中预先存储了所有热门的关键词，这些关键词可以按照 redis 结构分类存储。当获取到搜索请求中的搜索词之后，按照一定的模式识别方式（例如正则表达式匹配）在这个“词池”中查找与该搜索词匹配的关键词，并获取该关键词对应的查询结果。

[0035] 通过上述匹配规则确定出与输入的搜索词相匹配的关键词之后，进一步在缓存数据库中查找该关键词的查询结果。

[0036] 步骤四、将查找到的与输入的搜索词相匹配的关键词以及该关键词的查询结果发送给该客户端。

[0037] 客户端接收到该关键词以及该关键词的查询结果后，将查询结果显示给用户。

[0038] 通过上面的步骤就实现了本发明提供的用于对数据进行搜索的方法。可选地，由于某些类型的关键词的查询结果是与地域相关的，例如，对于“天气预报”这一关键词来说，北京的天气与深圳的天气通常是不同的，因此，“天气预报”这一关键词的查询结果就是与地域相关的，对于这样的关键词，在缓存数据库中存储对应的查询结果时，需要分别存储与各个地域相对应的查询结果，即：需要同时存储北京、深圳甚至其他地区的天气情况。相应地，当用户输入的搜索词与地域相关时，例如，当用户输入“天气”时，本实施例中的方法进一步包括：根据包含“天气”这一搜索词的搜索请求中携带的 IP 地址来确定发送搜索请求的客户端所处的地域，然后，在缓存数据库中查找与该地域相对应的查询结果。例如，如果发送搜索请求的客户端的 IP 地址显示为北京，则向该客户端返回的查询结果默认为北京的天气情况。通过判断客户端的 IP 地址，并提供与该 IP 地址相对应的查询结果，可以使查询结果更加符合用户的需求。

[0039] 另外，本发明实施例提供的用于对数据进行搜索的方法还可以进一步为用户提供补全搜索词的服务，即，当用户输入的搜索词仅为一部分时，可以自动地根据存储的关键词将搜索词补全并提示给用户。例如，当用户在生活服务类别的搜索框中输入“火车”时，可以自动为用户提示“火车票”以供用户选择，或者，也可以进一步向用户推荐多个与“火车”相关的词汇供用户选择。

[0040] 另外，为了进一步确保查询结果的全面性，本发明实施例中提供的用于对数据进行搜索的方法在获取到客户端发送的包含搜索词的搜索请求的步骤之后，进一步包括步骤：将搜索请求分发到搜索服务器，获取搜索服务器从外部的数据来源服务器查找到的搜索词对应的查询结果。相应地，当在缓存数据库中按照预设的匹配规则查找到的与搜索词相匹配的关键词及其对应的查询结果的数量少于预设数量时，该方法进一步包括：将获取的搜索服务器的查询结果发送给客户端，其中，搜索服务器的查询结果用于作为缓存数据库的查询结果的补充。具体地，每当获取到搜索请求后，同时将该搜索请求分发给搜索服务器，由该搜索服务器直接访问外部的数据来源服务器，得到查询结果，然后，对从缓存数据库中获取的查询结果以及搜索服务器中获取的查询结果进行合并，并根据需要选择是否采用自然搜索服务器的查询结果作为对缓存数据库中的查询结果的补充。例如，当从缓存数

据库中获取的查询结果的数量少于预设数量时，将获取的搜索服务器的查询结果发送给客户端作为补充。举例来说，假设客户端的结果显示页面中通常在一页上显示 10 条查询结果，这样，如果从缓存数据库中获取的查询结果不足十个（例如查询结果小于 10 个，甚至查询结果为 0），则需要从搜索服务器获取的查询结果中挑选一定数量的查询结果进行补充，具体挑选时，可以根据查询结果的相关度或热门度确定挑选顺序。通过这样的方式，由于搜索服务器可以从外部的数据来源服务器进行更广泛地搜索，因而既可以在通常情况下（即：缓存数据库缓存了用户要查找的词汇）为用户提供更加高效快捷的服务，又可以在特殊情况下（即：缓存数据库没有缓存用户要查找的词汇或缓存内容的数量不够丰富），实现更加全面地搜索，以满足用户多样化的搜索需求。

[0041] 图 2 示出了本发明实施例提供的用于对数据进行搜索的系统的结构示意图。如图 2 所示，该用于对数据进行搜索的系统 200 包括通信设备 210、缓存数据库 220 以及抓取服务器 230。其中，抓取服务器 230 预先提取关键词列表，通过访问外部的数据来源服务器 300 获取关键词列表中每一关键词对应的查询结果，将每一关键词及其对应的查询结果关联存储在缓存数据库中。通信设备 210 获取客户端 100 发送的包含搜索词的搜索请求，将搜索请求分发到缓存数据库中，在缓存数据库中按照预设的匹配规则查找与搜索词相匹配的关键词及其对应的查询结果，将查询结果发送给客户端 100。

[0042] 可选地，为了便于查找，缓存数据库中存储的关键词以及每一关键词对应的查询结果以键值对的方式存储，且关键词对应的查询结果可以是包含该关键词的网页对应的数据快照。

[0043] 而且，当关键词的查询结果与地域相关时，缓存数据库 230 中存储的关键词的查询结果还可以进一步包括与各个地域相对应的查询结果，这样，查找模块 220 在预先设置的缓存数据库 230 中查找关键词对应的查询结果时，进一步根据客户端 100 发送的搜索请求中携带的 IP 地址来确定该客户端 100 所处的地域，并在缓存数据库 230 中查找与该地域相对应的查询结果，从而可以为客户端 100 发送与其所处的地域相符合的查询结果。

[0044] 下面详细描述一下本发明提供的用于对数据进行搜索的系统。

[0045] 可选地，为了提高数据搜索的精准性，缩短搜索时间，在本实施例中，预先将用户可能会搜索的关键词按照一定的分类规则进行分类，相应地，在提供给用户的搜索界面中，针对每一类别，分别为用户提供一个搜索框。例如，可以预先将搜索词分为以下类别：生活服务、投资理财以及娱乐资讯等，这样，在搜索界面中可以进一步包括生活服务对应的搜索框、投资理财对应的搜索框，以及娱乐资讯对应的搜索框。这样，当用户需要输入搜索词进行搜索时，会先判断该搜索词属于哪一类别，然后，在该类别对应的搜索框中输入搜索词。例如，当用户要查询股票信息时，会选择投资理财对应的搜索框进行搜索，这样，由于在搜索时限定了搜索词所属的分类，搜索时仅对同一分类中的关键词进行查找，因此，既提高了查找速度，又使得查找结果更加准确，不易出现偏差。另外，还可以按照其他的分类方式进行分类，例如，按照视频、文本、图片等方式进行分类。

[0046] 下面以生活服务这一分类为例详细描述一下本实施例中的用于对数据进行搜索的系统的工作原理。

[0047] 首先，需要由抓取服务器 230 预先提取“生活服务”这一分类中的关键词，组成关键词列表，针对该关键词列表中的每一关键词，将包含该关键词的网页所对应的 URL 与该

关键词一起关联存储在该关键词列表里。

[0048] 具体地，在提取“生活服务”这一分类中的关键词时，抓取服务器 230 可以根据用户的搜索频率来确定要提取的关键词，例如，将预定时段内（例如，上一星期之内）用户搜索的频率较高的搜索词筛选出来作为关键词，其中，可以通过通信设备来完成对搜索词的搜索频率的统计。具体实现时，可以设定一个搜索阈值，将预定时段内的搜索次数大于该搜索阈值的搜索词筛选出来作为关键词。然后，针对每一关键词，由抓取服务器 230 获取包含该关键词的网页所对应的 URL 信息，并将该 URL 信息与该关键词关联存储。其中，对于每一关键词，包含该关键词的网页的数量可能是一个，也可能是多个，当网页数量为多个时，还可以进一步判定多个网页中的内容是否重复，当多个网页中的内容重复时，只要挑选其中的一个网页的 URL 进行存储即可，这样，既可以避免因存储的数据量过大而占用存储空间过多的问题，也可以在用户搜索时缩短查询时间。

[0049] 然后，抓取服务器 230 根据生成的关键词列表，访问外部的数据来源服务器 300，获取该数据来源服务器 300 中存储的与 URL 对应的网页数据，并根据获取的网页数据生成该网页对应的数据快照，将该数据快照与 URL 对应的关键词关联存储在缓存数据库 220 中。

[0050] 具体地，网络爬虫根据关键词列表中存储的与关键词对应的 URL，到数据来源服务器 300 中抓取与 URL 对应的网页数据，抓取后会对网页数据进行分析并拍照，形成该网页对应的数据快照。该数据快照中包含该 URL 对应的关键词，因此，将该数据快照作为该关键词对应的查询结果，与该关键词一起关联存储在缓存数据库中。具体存储时，为了方便查找，可以在缓存数据库 230 中通过键值对（key-value）的方式存储，即，将关键词作为 key，将该关键词对应的查询结果（即数据快照）作为 value。

[0051] 通过上面的方式，该用于对数据进行搜索的系统就建立起了缓存数据库 220，上面只是以“生活服务”这一个类别为例进行说明的，实际上，对于其他类别的关键词以及查询结果的获取，也是通过类似的方式实现的。

[0052] 缓存数据库 220 建立好之后，该系统就可以通过通信设备 210 获取用户通过客户端 100 发送的包含搜索词的搜索请求，将搜索请求分发到缓存数据库 220 中，在上述的缓存数据库 220 中按照预设的匹配规则查找与输入的搜索词相匹配的关键词，以及该关键词对应的查询结果。

[0053] 具体地，在通信设备 210 接收到包含搜索词的搜索请求后，需要在缓存数据库 220 中查找与该搜索词相匹配的关键词。本实施例中在判断搜索词与关键词是否匹配时，是根据预设的匹配规则进行判断的。

[0054] 其中，该预设的匹配规则可以是自然语言处理分析规则（简称 NLP），或者，也可以是正则表达式规则，或者，也可以是二者的结合。其中，自然语言处理分析规则大致分为两个层面，一个是浅层分析，如分词，词性标注，通常只需对句子的局部范围进行分析处理；另一个层面是对语言进行深层的处理，需要对句子进行全局分析，在分析时通常对句法、语义以及语用这三个层次进行分析。正则表达式规则一般是通过一些具有特定含义的字符来表示匹配规则的，例如，字符“^”匹配一个输入或一行的开头，如“^a”匹配“an A”，而不匹配“An a”；字符“$”匹配一个输入或一行的结尾，如“a$”匹配“An a”，而不匹配“an A”；字符“*”匹配前面元字符 0 次或多次，如“ba*”将匹配“b”，“ba”，“baa”以及“baaa”等。另外，还可以自定义一些匹配规则。例如，在本实施例中，可以预先定义“手机卫士”以及“手

机卫士”都对应“360 手机卫士”。通过匹配规则的设置，可以准确地确定与用户输入的搜索词相匹配的关键词，而且，当用户输入搜索词时有少许偏差，例如，搜索词中有一个错别字或丢掉了一个字，这时，根据自然语言处理分析规则，仍然可以确定出用户实际想要的关键词。

[0055] 通信设备 210 通过上述匹配规则确定出与输入的搜索词相匹配的关键词之后，进一步在缓存数据库 230 中查找该关键词的查询结果，然后，通信设备 210 将查找到的与输入的搜索词相匹配的关键词以及该关键词的查询结果发送给该客户端 100。客户端 100 接收到该关键词以及该关键词的查询结果后，将查询结果显示给用户。

[0056] 图 3 示出了当客户端发送的搜索请求中包含的搜索词为“蜘蛛侠”时显示的查询结果的示意图。通过图 3 可以看出，当用户输入“蜘蛛侠”时，本发明提供的用于对数据进行搜索的方法和系统会为用户提供图 3 中的四个包含蜘蛛侠的视频内容。这四个视频的共同特点是在内容简介部分都包含“蜘蛛侠”三个字，与搜索词匹配，因此，作为查询结果提供给用户。

[0057] 在上面描述的用于对数据进行搜索的系统中，抓取服务器 230 还可以进一步按照预设的频率对关键词列表中的关键词和 / 或关键词对应的查询结果进行更新。例如，可以设置每天或每星期进行一次更新，具体实现时，可以从如下两方面进行更新 :第一个方面为，每隔一段时间后，将近期用户搜索频率较高的搜索词添加到关键词列表中，并获取新添加的关键词的查询结果，也就是对关键词列表中的关键词数量进行更新，以确保及时加入近期较热门的搜索词 ;第二个方面为，每隔一段时间后，针对关键词列表中现有的关键词，重新从数据来源服务器上获取每一关键词对应的查询结果，也就是对关键词列表中每一关键词的查询结果进行更新，以确保所有关键词的查询结果都是比较新的。

[0058] 而且，在上面描述的用于对数据进行搜索的系统中，缓存数据库中还可以进一步包括排序模块，用于对缓存数据库中的关键词进行排序。具体排序时，可以根据一定的时间段内(例如一天、一月等)用户的点击频次来确定关键词的排列顺序。或者，也可以为每个关键词设置一个权重，根据权重的大小来确定关键词的排列顺序。具体地，在确定每个关键词的权重时，可以结合多方面的因素来确定，例如，结合关键词的搜索频率、关键词的重要性和 / 或一定时间段内用户的点击频次来确定。通过对缓存数据库中的关键词进行排序，可以使用户优选找到最符合需求的关键词，能够提高查找效率。

[0059] 另外，为了进一步确保查询结果的全面性，本发明实施例中提供的用于对数据进行搜索的系统还可以进一步包括搜索服务器(图中未示出)。该搜索服务器一端与通信设备 210 相连，另一端与外部的数据来源服务器相连，用于从外部的数据来源服务器查找搜索词对应的查询结果。具体地，每当通信设备 210 接收到搜索请求后，同时将该搜索请求分发给该搜索服务器，由该搜索服务器直接访问外部的数据来源服务器，得到查询结果，并将该查询结果提供给通信设备 210，由通信设备 210 对从缓存数据库中获取的查询结果以及搜索服务器中获取的查询结果进行合并，并根据需要选择是否采用自然搜索服务器的查询结果作为对缓存数据库中的查询结果的补充。也就是说，通信设备 210 具有分发合并的功能。例如，当通信设备 210 从缓存数据库中获取的查询结果的数量少于预设数量时，将获取的搜索服务器的查询结果发送给客户端作为补充。举例来说，假设客户端的结果显示页面中通常在一页上显示 10 条查询结果，这样，如果通信设备 210 从缓存数据库中获取的查询结果

不足十个(例如查询结果小于 10 个,甚至查询结果为 0),则需要从搜索服务器获取的查询结果中挑选一定数量的查询结果进行补充,具体挑选时,可以根据查询结果的相关度或热门度确定挑选顺序。通过这样的方式,可以实现更加全面地搜索,从而为用户提供更多的搜索结果。

[0060] 本发明实施例提供的用于对数据进行搜索的方法和系统,在搜索之前,可以预先对所有的关键词进行分类,然后,在缓存数据库中将关键词按照类别进行存储,这样,用户在输入搜索词时,可以在该搜索词所属分类对应的搜索框中进行搜索,这样,本发明中的用于对数据进行搜索的方法和系统则只对该分类中的关键词进行查询,这一方式也被称为垂直领域搜索。采用这种方式,一方面,由于只查询一个分类中的关键词,无需检索全部的关键词,因此,提高了查询的速度。另一方面,由于确定了搜索词所属的分类,不会错误地将其他类别的查询结果误当作用户输入的搜索词的查询结果,因此,还提高了查询的精准度,关于这一点,当搜索词有可能同时属于多个类别时尤为重要。

[0061] 而且,本发明实施例提供的用于对数据进行搜索的方法和系统,在缓存数据库中通过键值对的方式存储关键词和对应的查询结果,这种存储方式简单明了,占用存储空间小,且算法简单、检索速度快,从而进一步提高了查询的速度。

[0062] 另外,本实施例提供的用于对数据进行搜索的方法和系统,预先将关键词及其对应的查询结果以数据快照的方式存储在了本地的缓存数据库中,因此,向用户提供服务时,无需再访问数据来源服务器,只需访问本地的缓存数据库即可,由此降低了合作数据服务(即数据来源服务器)的压力。而且,由于有了缓存数据库,网络爬虫只需在向缓存数据库中存储关键词的阶段去数据来源服务器上抓取数据即可,而在后续处理用户搜索请求时,该系统只要根据缓存数据库上已经存储的数据就可以为用户提供查询服务,不必像常规的搜索方式那样,需要每次在处理用户搜索请求时都由网络爬虫去数据来源服务器上抓取数据,从而也减轻了网络爬虫的爬取压力。而且,由于本发明中的缓存数据库中的关键词可以按照分类进行存储,因此还进一步减轻了网络爬虫爬取垂直数据(同一分类下的数据)的压力。通过上述方式,有利于提高查询速度。

[0063] 另外,本发明实施例提供的用于对数据进行搜索的方法和系统,在确定与搜索词相匹配的关键词时,预先定义了匹配规则,例如,自然语言处理分析规则或正则表达式规则,这样,在匹配时即使用户输入的搜索词有少许误差,也可以精准地匹配到合适的关键词,从而提高了查询的精准度。

[0064] 综上所述,本发明实施例提供的用于对数据进行搜索的方法和系统,提高了查询速度以及查询的精准度。

[0065] 在此提供的算法和显示不与任何特定计算机、虚拟系统或者其它设备固有相关。各种通用系统也可以与基于在此的示教一起使用。根据上面的描述,构造这类系统所要求的结构是显而易见的。此外,本发明也不针对任何特定编程语言。应当明白,可以利用各种编程语言实现在此描述的本发明的内容,并且上面对特定语言所做的描述是为了披露本发明的最佳实施方式。

[0066] 在此处所提供的说明书中,说明了大量具体细节。然而,能够理解,本发明的实施例可以在没有这些具体细节的情况下实践。在一些实例中,并未详细示出公知的方法、结构和技术,以便不模糊对本说明书的理解。

[0067] 类似地，应当理解，为了精简本公开并帮助理解各个发明方面中的一个或多个，在上面对本发明的示例性实施例的描述中，本发明的各个特征有时被一起分组到单个实施例、图、或者对其的描述中。然而，并不应将该公开的方法解释成反映如下意图：即所要求保护的本发明要求比在每个权利要求中所明确记载的特征更多的特征。更确切地说，如下面的权利要求书所反映的那样，发明方面在于少于前面公开的单个实施例的所有特征。因此，遵循具体实施方式的权利要求书由此明确地并入该具体实施方式，其中每个权利要求本身都作为本发明的单独实施例。

[0068] 本领域那些技术人员可以理解，可以对实施例中的设备中的模块进行自适应性地改变并且把它们设置在与该实施例不同的一个或多个设备中。可以把实施例中的模块或单元或组件组合成一个模块或单元或组件，以及此外可以把它们分成多个子模块或子单元或子组件。除了这样的特征和 / 或过程或者单元中的至少一些是相互排斥之外，可以采用任何组合对本说明书（包括伴随的权利要求、摘要和附图）中公开的所有特征以及如此公开的任何方法或者设备的所有过程或单元进行组合。除非另外明确陈述，本说明书（包括伴随的权利要求、摘要和附图）中公开的每个特征可以由提供相同、等同或相似目的的替代特征来代替。

[0069] 此外，本领域的技术人员能够理解，尽管在此所述的一些实施例包括其它实施例中所包括的某些特征而不是其它特征，但是不同实施例的特征的组合意味着处于本发明的范围之内并且形成不同的实施例。例如，在下面的权利要求书中，所要求保护的实施例的任意之一都可以以任意的组合方式来使用。

[0070] 本发明的各个部件实施例可以以硬件实现，或者以在一个或者多个处理器上运行的软件模块实现，或者以它们的组合实现。本领域的技术人员应当理解，可以在实践中使用微处理器或者数字信号处理器（DSP）来实现根据本发明实施例的用于对数据进行搜索的系统中的一些或者全部部件的一些或者全部功能。本发明还可以实现为用于执行这里所描述的方法的一部分或者全部的设备或者装置程序（例如，计算机程序和计算机程序产品）。这样的实现本发明的程序可以存储在计算机可读介质上，或者可以具有一个或者多个信号的形式。这样的信号可以从因特网网站上下载得到，或者在载体信号上提供，或者以任何其他形式提供。

[0071] 应该注意的是上述实施例对本发明进行说明而不是对本发明进行限制，并且本领域技术人员在不脱离所附权利要求的范围的情况下可设计出替换实施例。在权利要求中，不应将位于括号之间的任何参考符号构造成对权利要求的限制。单词“包含”不排除存在未列在权利要求中的元件或步骤。位于元件之前的单词“一”或“一个”不排除存在多个这样的元件。本发明可以借助于包括有若干不同元件的硬件以及借助于适当编程的计算机来实现。在列举了若干装置的单元权利要求中，这些装置中的若干个可以是通过同一个硬件项来具体体现。单词第一、第二、以及第三等的使用不表示任何顺序。可将这些单词解释为名称。

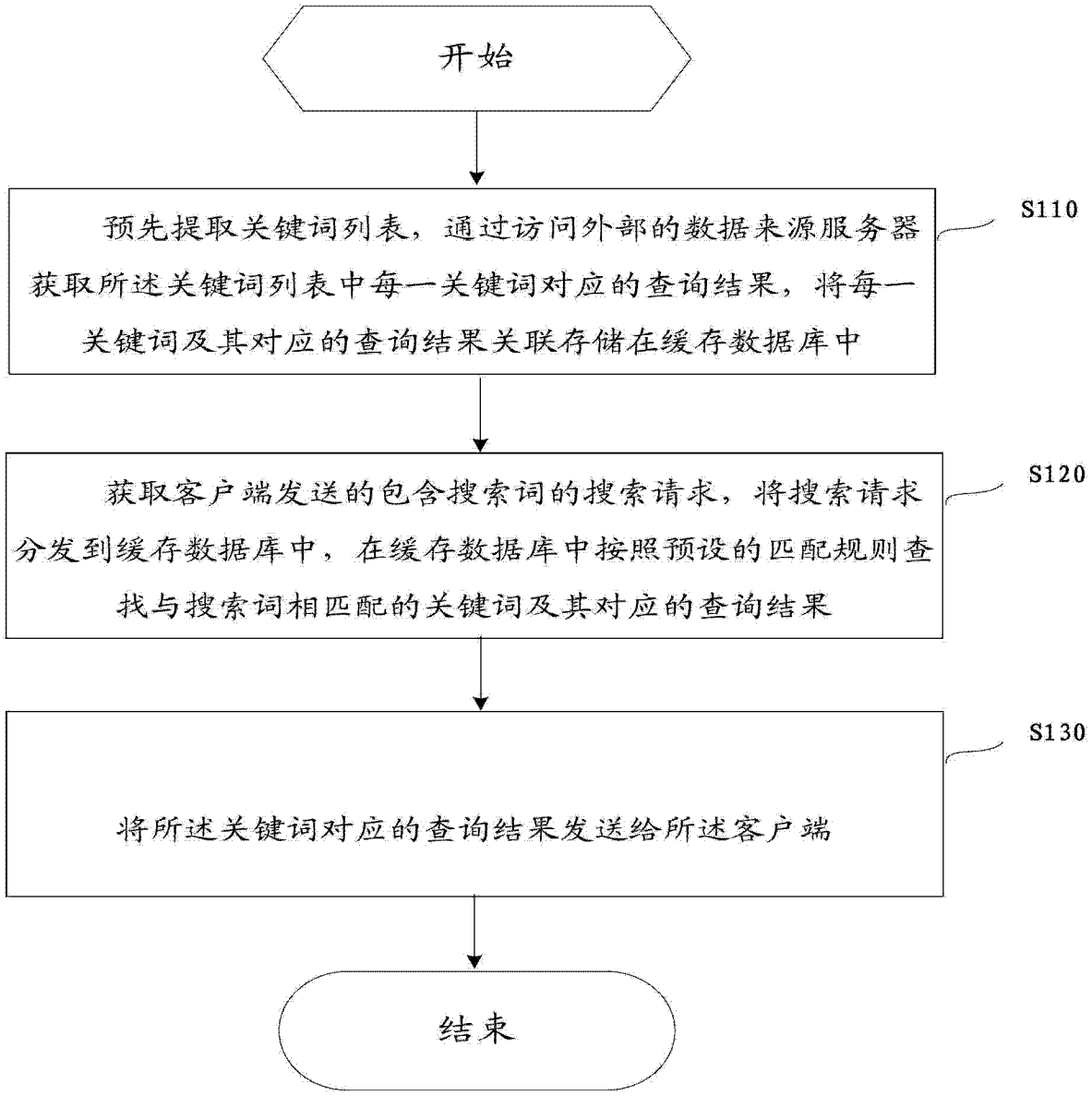

开始
预先提取关键词列表，通过访问外部的数据来源服务器获取所述关键词列表中每一关键词对应的查询结果，将每一关键词及其对应的查询结果关联存储在缓存数据库中
S110
获取客户端发送的包含搜索词的搜索请求，将搜索请求分发到缓存数据库中，在缓存数据库中按照预设的匹配规则查找与搜索词相匹配的关键词及其对应的查询结果
S120
将所述关键词对应的查询结果发送给所述客户端
S130
结束

图 1

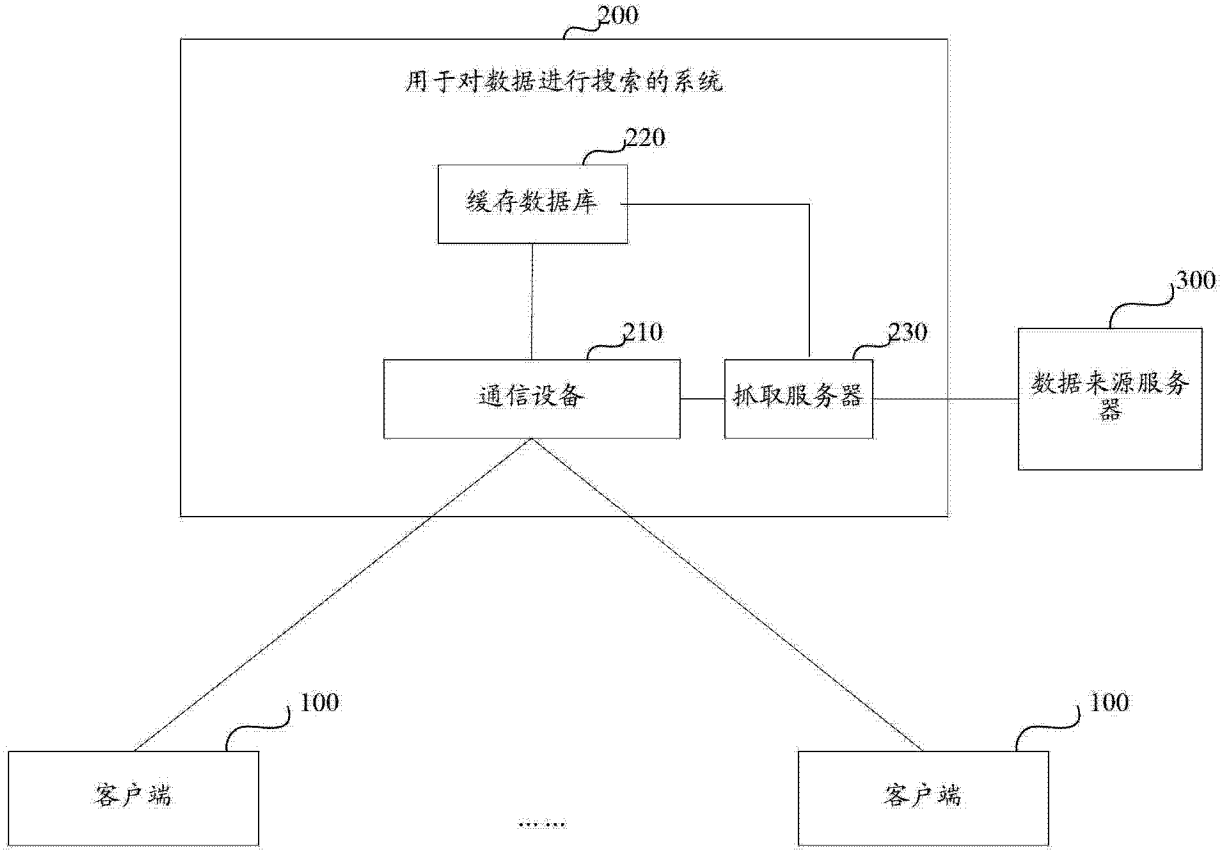

200
用于对数据进行搜索的系统
220
缓存数据库
210
通信设备
230
抓取服务器
300
数据来源服务器
100
客户端
……
100
客户端

图 2


蜘蛛侠4在线观看_360视频
03:50
134:05
01:34
01:57
【拍客】广州 蜘蛛...
蜘蛛侠4【超凡蜘蛛...
男子为开门自拍蜘蛛...
超凡跑酷侠来袭!刺...
来源：优酷
来源：搜狐
来源：搜狐
来源：优酷
so.v.360.cn/mini.php?kw=%E8%9C%98%E8%9B%... 2012-09-11

图 3