[19] 中华人民共和国国家知识产权局

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

[12] 发明专利说明书

专利号 ZL 200710100309.8

[45] 授权公告日 2009 年 4 月 8 日

[11] 授权公告号 CN 100476830C

[22] 申请日 2007.6.7

[21] 申请号 200710100309.8

[73] 专利权人 北京金山软件有限公司

地址 100083 北京市海淀区北四环中路 238 号柏彦大厦 20 层

共同专利权人 北京金山数字娱乐科技有限公司

哈尔滨工业大学

[72] 发明人 周连强 贾建坤 高立琦 刘 挺

[56] 参考文献

US6691104B1 2004.2.10

US2006/0282416A1 2006.12.14

CN1808426A 2006.7.26

CN1809827A 2006.7.26

审查员 刘 曼

[74] 专利代理机构 北京集佳知识产权代理有限公司

代理人 逯长明

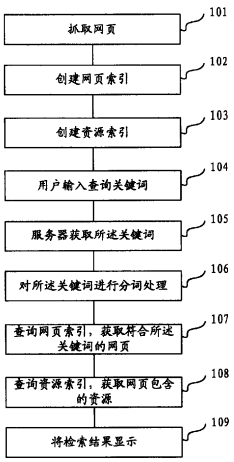
权利要求书 1 页 说明书 8 页 附图 3 页

[54] 发明名称

一种网络资源检索方法及系统

[57] 摘要

本发明公开了一种网络资源检索方法及系统，以解决现有的网页信息检索，耗费用户时间和精力而无法快速、准确地获取资源的问题。 所述方法包括：创建网页索引，并对网页中包含的资源，创建对应每个网页的资源索引；接收用户输入的检索关键词，并在网页索引中查询符合所述关键词的网页；在资源索引中查询所述符合关键词的网页包含的资源；将包含所述符合关键词的网页信息和相应资源信息的检索结果显示。 本发明在页面的一侧(例如左侧)显示网页正文摘要，另一侧(例如右侧)显示对应的资源信息(如资源名称，资源链接)，用户可以直观地获知每个网页中都包含了哪些可下载的资源，通过直接下载可快速地获取自己想要的各种资源。



1、一种网络资源检索方法，其特征在于，包括:

创建网页索引，并对网页中包含的资源，创建对应每个网页的资源索引;

接收用户输入的检索关键词，并在网页索引中查询符合所述关键词的网页;

在资源索引中查询所述符合关键词的网页包含的资源;

将包含所述符合关键词的网页信息和相应资源信息的检索结果显示。

2、根据权利要求 1 所述的方法，其特征在于：在页面的一侧显示网页信息，另一侧显示相应的资源信息。

3、根据权利要求 1 所述的方法，其特征在于：按照资源与所述关键词的相关性高低,将网页包含的所有资源排序,并将排名靠前的部分资源信息显示。

4、根据权利要求 1 所述的方法，其特征在于：以资源所在网页的 URL 为索引建立资源索引。

5、根据权利要求 1 所述的方法，其特征在于，还包括：根据用户的不同侧重点,按照侧重网页内容或者侧重资源内容,对检索到的网页信息进行排序。

6、一种网络资源检索系统，其特征在于，包括:

索引单元，用于创建网页索引，并对网页中包含的资源，创建对应每个网页的资源索引;

检索单元，用于在网页索引中查询符合检索关键词的网页，并在资源索引中查询所述符合关键词的网页包含的资源;

查询代理单元，用于接收用户输入的检索关键词，并通过所述检索单元的检索，将包含符合关键词的网页信息和相应资源信息的检索结果显示给用户。

7、根据权利要求 6 所述的系统，其特征在于：所述查询代理单元在页面的一侧显示网页信息，另一侧显示相应的资源信息。

8、根据权利要求 6 所述的系统，其特征在于，还包括：排序单元，用于根据用户的不同侧重点，按照侧重网页内容或者侧重资源内容，对检索到的网页信息进行排序。

9、根据权利要求 8 所述的系统，其特征在于：所述排序单元还按照资源与所述关键词的相关性高低，将网页包含的所有资源排序，并将排名靠前的部分资源信息通过所述查询代理单元显示。

10、根据权利要求 6 所述的系统，其特征在于：所述索引单元以资源所在网页的 URL 为索引建立资源索引。

一种网络资源检索方法及系统

技术领域

本发明涉及搜索引擎技术，特别是涉及一种网络资源检索方法及系统。

背景技术

随着网络技术的快速发展，网页所承载的信息内容越来越多，例如MP3、应用软件、学习课程等。因此在很多情况下，用户在进行Web信息检索时，不仅仅关心页面上的内容，同时也关心页面上所含有的各种资源链接，如音频文件、视频文件等。

现有的网页信息检索，例如百度、google 等，假如用户输入关键词检索某个视频资源，在搜索结果页面中返回了包含该关键词的网页链接及页面内容的简要介绍；用户需要点击所选页面链接，通过浏览该页面，才能确定该页面中是否包含需要的资源或所关心的其他内容，进一步进行下载或获取。

按照上述方法，用户可以通过查找网页获取所关心的信息或者资源。但是，由于在检索结果的页面中，用户无法得知每个网页中都包含了哪些可下载的资源，因此需要用户耗费时间和精力进一步进行筛选，而无法快速地获取到自己想要的资源。而且，大部分网页中的资源名称都用了简单的标识，用户通过关键字检索网页时，经常无法获得准确的结果。

例如，一个网页内容中包含了“大学听力第一册”关键词，该网页中提供了“part1.mp3”，“part2.mp3”，“part3.mp3”等资源，用户需要检索到该页面并进行资源下载。用户在以“大学听力第一册”为关键词进行搜索网页时，可能会返回一系列与“大学听力第一册”相关的网页内容，但不一定每个网页中都包含以上资源的下载，用户需要进一步浏览网页进行筛选；若用户以“part1.mp3”为关键词进行搜索，经常搜索出的网页内容除包含大学听力第一册外，可能还包括其他不相关的资源，例如某个电影的下载片断也叫part1.mp3，用户同样需要进一步进行筛选。

总之，虽然现有的搜索网站提供了特定资源的直接下载，例如百度提供的mp3的检索，但是不能满足用户对各种资源下载的需求。

发明内容

本发明所要解决的技术问题是提供一种网络资源检索方法及系统，以解决

现有的网页信息检索，需要用户耗费时间和精力进一步进行筛选，而无法快速、准确地获取资源的问题。

为解决上述技术问题，根据本发明提供的具体实施例，本发明公开了以下技术方案：

一种网络资源检索方法，包括：

创建网页索引，并对网页中包含的资源，创建对应每个网页的资源索引；

接收用户输入的检索关键词，并在网页索引中查询符合所述关键词的网页；

在资源索引中查询所述符合关键词的网页包含的资源；

将包含所述符合关键词的网页信息和相应资源信息的检索结果显示。

优选的，在页面的一侧显示网页信息，另一侧显示相应的资源信息。

优选的，按照资源与所述关键词的相关性高低，将网页包含的所有资源排序，并将排名靠前的部分资源信息显示。

其中，以资源所在网页的URL为索引建立资源索引。

所述方法还包括：根据用户的不同侧重点，按照侧重网页内容或者侧重资源内容，对检索到的网页信息进行排序。

一种网络资源检索系统，包括：

索引单元，用于创建网页索引，并对网页中包含的资源，创建对应每个网页的资源索引；

检索单元，用于在网页索引中查询符合检索关键词的网页，并在资源索引中查询所述符合关键词的网页包含的资源；

查询代理单元，用于接收用户输入的检索关键词，并通过所述检索单元的检索，将包含符合关键词的网页信息和相应资源信息的检索结果显示给用户。

优选的，所述查询代理单元在页面的一侧显示网页信息，另一侧显示相应的资源信息。

所述系统还包括：排序单元，用于根据用户的不同侧重点，按照侧重网页内容或者侧重资源内容，对检索到的网页信息进行排序。

其中，所述排序单元还按照资源与所述关键词的相关性高低，将网页包含的所有资源排序，并将排名靠前的部分资源信息通过所述查询代理单元显示。

其中，所述索引单元以资源所在网页的URL为索引建立资源索引。

根据本发明提供的具体实施例，本发明公开了以下技术效果：

首先，通过建立网页索引和对应网页的资源索引，能够将符合用户检索关键词的网页信息和资源信息同时显示。所述将资源信息直接展示，用户可以直观地获知每个网页中都包含了哪些可下载的资源，而无需进入资源所在页面，用户通过在检索结果页面直接下载，即可快速地获取自己想要的各种资源。

而且，所述显示界面新颖，在页面的一侧（例如左侧）显示网页正文摘要，另一侧（例如右侧）显示对应的资源信息（如资源名称，资源链接），突破了传统搜索引擎的显示方式。

其次，结果页面中网页的摘要介绍，对相应网页中的资源提供了一个辅助性的说明，用户可以根据资源所在页面的摘要信息判断该资源是否为所需。因此，资源所在页面的摘要信息作为用户判断该资源的依据，增加了用户判断资源内容的准确性，从而提高了用户获取资源的准确性。

再次，在进行检索结果排序时，考虑用户的侧重方向（侧重网页内容或侧重资源内容），将网页中的资源的锚也作为指标进行权重的计算。根据用户的侧重点返回的检索结果顺序不同，可以更好地满足用户的需求。

附图说明

图1是本发明实施例所述快检索网页所含资源的步骤流程图；

图2是本发明实施例中网页正文索引与资源索引之间的关系示意图；

图3是本发明实施例中检索结果的页面显示效果图；

图4是本发明实施例所述快检索网页所含资源的系统结构图。

具体实施方式

为使本发明的上述目的、特征和优点能够更加明显易懂，下面结合附图和具体实施方式对本发明作进一步详细的说明。

针对在检索结果的页面中，用户无法得知每个网页中都包含了哪些可下载的资源，以及由于资源名称简单，用户无法获得准确的检索结果的问题，本发明实施例提供了一种可快速检索网页中包含的资源的方法。通过创建网页索引，并创建以资源所在网页的URL为索引的资源索引，可以在检索网页时，将网页中的资源一同检索出来，并同时显示在检索结果页面中，便于用户直接下

载，快速地获取自己想要的各种资源。

参照图 1，是本发明实施例所述快速检索网页所含资源的步骤流程图。下面将以 Web 搜索中的资源获取为例进行说明。

步骤 101，利用网页抓取工具，从互联网获取网页。

步骤 102，对获取的网页建立索引。具体过程是：提取网页正文，并根据网页的编码对网页正文进行相应的编码转换；然后对正文进行分词处理，去掉“的、啊、哦”等等停用词；再对剩下的正文关键词，以所述正文关键词为索引，建立倒排索引。建立倒排索引的示例如下：

文本 1 的正文关键词是：aaa bbb ccc ddd;

文本 2 的正文关键词是：bbb ddd yyy;

以关键词建立倒排索引后：aaa 1

bbb 1,2

ccc 1

ddd 1,2

yyy 2

如果需要查找哪些文本中含有关键词 bbb 时，只需取出该关键词所对应的文本号 1,2 即可。

步骤 103，分析网页中可能含有的资源链接，创建一个独立的资源索引。创建步骤如下：

首先，获取网页中以{名称}标签标识的链接以及锚文本。通常情况下，{名称}为 HTML 语言，用以定义一个链接，其中“名称”即为显示在网页中的文字，称为锚文本。例如，在个人网站上把中央电视台（www.cctv.com）作为新闻频道的链接，访问者通过点击网站上的“新闻频道”就能进入 http://www.cctv.com 网站，那么“新闻频道”就是中央电视台网站首页的锚文本。

其次，判断获取的链接是否为资源。如果链接以“.mp3”、“.exe”之类的字符串结尾，则显然是可以下载的资源；如果链接中含有“?”、“&”等信息，则该链接可能为重定向链接，需要进一步确认其是否对应一个资源。关于如何判断一个链接是否为资源链接，可采用本领域技术人员所熟知的各种方法实

现，在此不作详细说明。

再次，经判断后，如果是资源链接，则对每个包含资源的网页，创建一个独立的资源索引。

本步骤中以资源所在网页的URL为索引，如图2所示，是所述实施例中网页正文索引与资源索引之间的关系示意图。图中，“特征项”即为建立网页正文索引的索引关键词，每个“特征项”都对应着一系列的网页URL，其中每个包含资源的网页URL又对应着一系列该网页所包含的所有资源。

当然，在建立资源索引时，也可以选取其他索引词，例如每个网页在网页索引中的位置编号等。

步骤104，用户在搜索框中输入查询关键词，并触发查询事件。

步骤105，服务器收到所述查询事件后，获取用户输入的查询关键词。

步骤106，对获得的查询关键词进行分词处理。所述分词处理是为了获取关键词中最常用的词根，例如关键词为“中国政府推出知识产权新举措”，分词结果可能为“中国”、“政府”、“知识产权”、“举措”，或者是“中国政府”、“知识产权举措”等等，能有效的排除不是常用组合的搭配，例如“国政”，这样可以减少搜索的词根。

步骤107，在网页索引中进行查询，获取符合所述关键词的网页。例如图2所示中，在以“特征项”为索引词的网页索引中，查找出“特征项”是所述关键词的索引，该“特征项”对应的所有网页即为符合所述关键词的网页。

步骤108，查找每个网页URL对应的资源索引，在对应的资源索引中找到该网页包含的所有资源。本发明与传统的信息检索不同，在检索与用户关键词符合的网页信息时，一同将网页中包含的资源信息也检索出来。

步骤109，对检索返回的结果进行显示，检索结果包括网页信息和网页中包含的资源信息。显示方式是在结果页面的一侧显示检索到的网页信息，另一侧显示对应的资源信息。

参照图3所示，是本发明实施例中检索结果的页面显示效果图。本例中，在页面的左侧显示检索到的网页的正文摘要及链接等信息，在页面的右侧显示对应该网页的资源名称及资源链接等信息。

本发明实施例优选的，在显示所述检索结果之前，先对检索结果进行排序

处理，然后按照排序结果显示。其中对于网页信息的排序，排序规则分为侧重网页内容的排序和侧重资源内容的排序。

通常，服务器对检索出的相关网页，采用一定的策略进行排序，例如对网页进行打分，然后按照分数高低决定返回顺序；而打分的方法是参考几个指标，然后按照网页与所述指标的相关性进行打分，比如关键词出现的频率或区分度等，最后对所述指标进行加权做和得到网页的最后得分。本发明所述实施例中，由于引入了资源索引，所以在对网页打分时，将资源的锚的关键词出现的频率值也作为一个排序的指标，如果用户侧重资源，则调高这个指标所占的权重，而如果用户侧重网页，则调低这个指标的权重，然后提高其他指标的权重。

因此，根据不同侧重点，对网页的排序结果也不同。通过设置用户选项，如果用户在搜索时选择侧重网页内容，则在网页索引中检索出的网页内容所占的权值高；如果用户选择侧重资源内容，则在资源索引中检索出的资源的锚文本所占的权值高。

在显示网页所包含的资源信息时，由于显示空间所限，如果网页中的资源信息较多，通常选取部分显示。选取方法有多种，例如按照资源在网页中出现的先后顺序选取前几个，或者按照资源名称选取，等等。本发明所述实施例中，为给用户带来更好的使用体验，便于用户直观获取自己想要的资源，在选取要显示的资源时先对资源进行了排序。按照资源与检索关键词的相关性，将相关性高的资源显示在页面。

上述实施例提供了一种新颖的搜索结果展示界面，将网页包含的资源信息直接展示，用户可以直观地获知每个网页中都包含了哪些可下载的资源，无需进入资源所在页面即可直接下载；并且，用户可以根据资源所在页面的正文摘要信息，判断资源是否是自己需要的，进一步增加了资源获取的准确性。

而且，用户在搜索网页的时候，由于右侧显示出了对应的资源，用户可能会在无意中发现需要的资源，然后对资源进行下载，这样就激发了用户的潜在需求。如果用户觉得这个网站比较有新意、比较实用，然后会更多的访问，从而提高了网站的粘性。

本发明实施例还提供了一种资源获取系统，仍以 Web 搜索中的资源获取为例，参照图 4，是本发明实施例所述快检索网页所含资源的系统结构图。所述

系统主要包括索引单元401、检索单元402和查询代理单元403。

索引单元401用于建立网页索引和资源索引。建立网页索引时，索引单元401先提取网页正文，并根据网页的编码对网页正文进行相应的编码转换，然后对正文进行分词处理，以分词后的正文关键词为索引建立网页倒排索引。

对应每个包含资源链接的网页，索引单元401还建立了单独的资源索引，以资源所在网页的URL为索引词，可参见图2所示，通过查找网页的URL，即可找到网页包含的所有资源。索引单元401首先需要分析网页获取其中的链接及锚文本，然后判断所述链接是否为资源链接，如果是资源链接，则为该网页中存在的所有资源建立一个资源链接。

检索单元402用于根据索引单元401建立的网页索引和资源索引，查询与检索关键词符合的网页信息和资源信息。首先，检索单元402对检索关键词进行分词处理，排除不是常用组合的搭配；然后，根据检索关键词查询网页索引，获取符合所述关键词的网页；再根据网页URL查找到网页包含的所有资源。这样，检索单元402在检索网页信息时，一同将网页包含的资源信息也检索出来。

查询代理单元403用于接收用户输入的检索关键词，并传给检索单元402处理；当检索单元402返回检索结果时，将所述检索结果显示给用户。本发明提出了一种新颖的结果展示方式，在结果页面的一侧（例如左侧）显示检索到的网页信息，如网页的正文摘要及链接等信息，另一侧（例如右侧）显示对应的资源信息，如资源名称及资源链接等信息。

优选的，还提供了用户选项，根据用户选择侧重网页内容还是侧重资源内容，检索单元402先对检索到的网页信息和资源信息分别进行排序处理，再返回给查询代理单元403。在对网页进行排序时，将资源的锚的关键词出现的频率值也作为一个排序的指标，如果用户侧重资源，则调高这个指标所占的权重，而如果用户侧重网页，则调低这个指标的权重，然后提高其他指标的权重。在对资源进行排序时，按照资源与检索关键词的相关性，将相关性高的资源排在前面。如果网页中存在的资源较多，查询代理单元403显示部分资源信息。

所述系统的整体处理流程是：首先利用网页抓取工具404从互联网获取网页，并存入数据库405中；然后索引单元401从数据库405中提取网页正文，创建网页索引和资源索引；当查询代理单元403接收用户输入的检索关键词

后，由检索单元402实现信息检索，检索单元402通过查询网页索引和资源索引，将与所述检索关键词符合的网页信息和对应的资源信息进行排序处理后，返回给查询代理单元403；查询代理单元403在页面的左侧显示网页的正文摘要及链接等信息，右侧显示对应的资源名称及资源链接等信息。因此，用户可以直接在搜索结果页面下载自己需要的资源，提高了资源获取的速度和准确性。

图4所示系统中未详述的部分可以参见图1所示方法的相关部分，为了篇幅考虑，在此不再详述。

以上对本发明所提供的一种网络资源检索方法及系统，进行了详细介绍，本文中应用了具体个例对本发明的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本发明的方法及其核心思想；同时，对于本领域的一般技术人员，依据本发明的思想，在具体实施方式及应用范围上均会有改变之处。综上所述，本说明书内容不应理解为对本发明的限制。

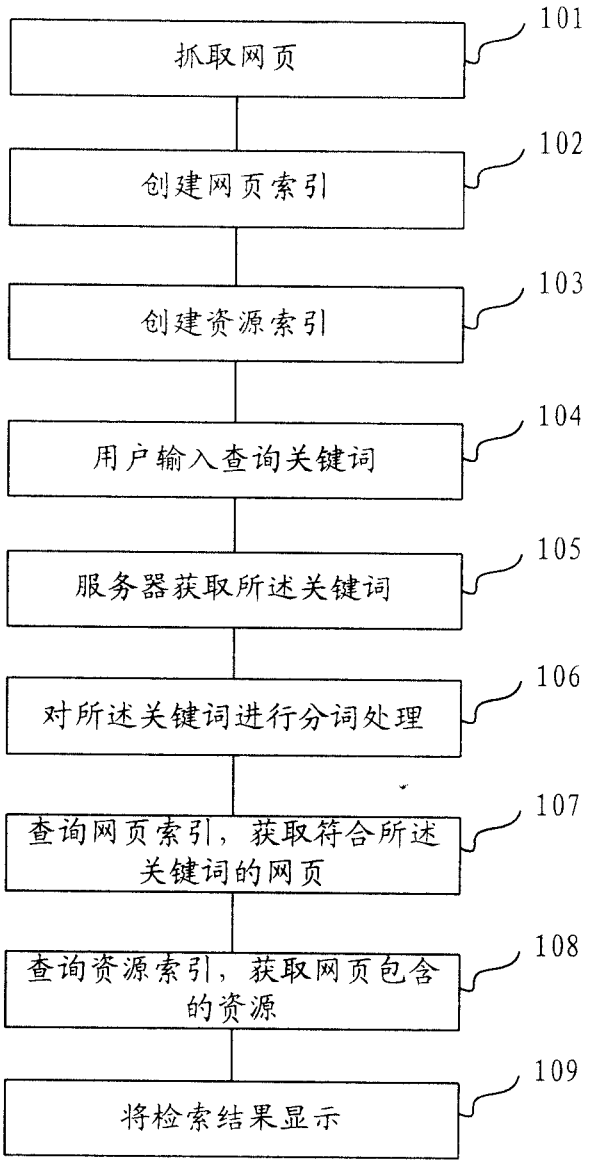

抓取网页
101
创建网页索引
102
创建资源索引
103
用户输入查询关键词
104
服务器获取所述关键词
105
对所述关键词进行分词处理
106
查询网页索引，获取符合所述关键词的网页
107
查询资源索引，获取网页包含的资源
108
将检索结果显示
109

图 1

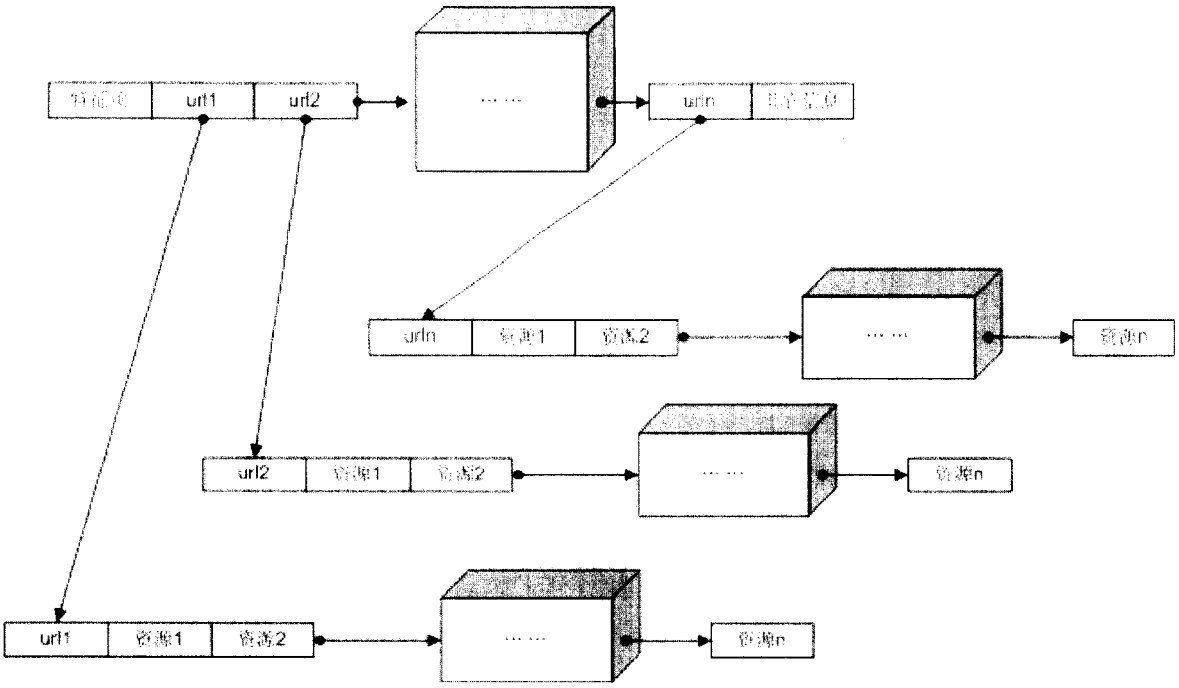


图 2

专题 1996-2005年大学英语四六级听力原文及...

... 考研 | GRE | TOEFL | IELTS | GMAT | 英语游戏 | 英文歌曲 | 留学移民 | 听力口语 四六级听力在线试听 2005年12月英语四级听力：原文下载 2005年06月英语四级听力：原文下载 2005年01月英语四级听力：原文下载 2005年12月英语六级听力：原文下载 2005年06月英语六级听力：原文下载 2005年01月英语六级听力：原文下载 闭

http://www.jxue.com/zt/06zt/cetlisten/ - 问天快照

更多结果...

本页与资源相关的链接 约53个

- 《美文与野兽》剧本下载
- 新概念英语2-4册mp3下载
- 决战gre：《红宝书》下载

更多资源

四六级历年真题mp3听力下载 点点英语 中国大...

... 听力MP3 1998年6月四级听力MP3 1999年1月四级听力MP3 1999年6月四级听力MP3 2000年1月四级听力MP3 2000年6月四级听力MP3 2001年1月四级听力MP3 2001年6月四级听力MP3 2002年1月四级听力MP3 2002年6月四级听力MP3 2003年1月四级听力MP3 2003年6月四级听力MP3 2004年1月四级

http://www.diandian.net/Exam/1003/1807236654.htm - 问天快照

更多结果...

本页与资源相关的链接 约1个

- 资料下载

图 3

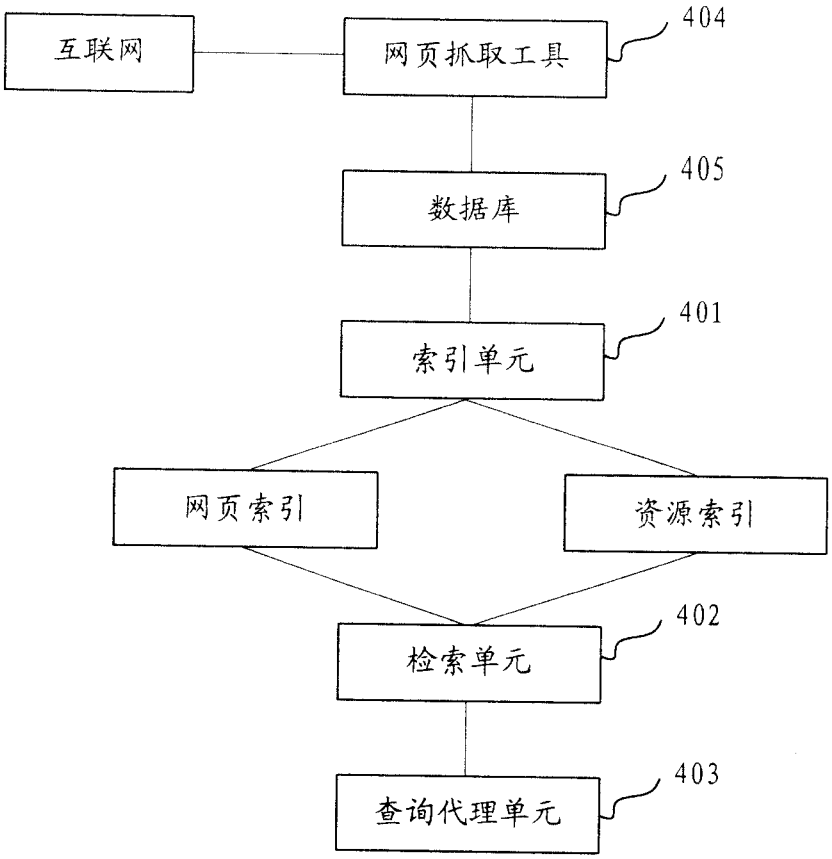


图 4