[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 200610090568.2

[45] 授权公告日 2009 年 1 月 28 日

[11] 授权公告号 CN 100456296C

[22] 申请日 2006.6.28

[21] 申请号 200610090568.2

[73] 专利权人 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 403 室

[72] 发明人 余祥鑫 文 杰 熊 应 刘致远

[74] 专利代理机构 北京德琦知识产权代理有限公司

代理人 罗正云 宋志强

[56] 参考文献

US2003/0208482A1 2003.11.6

CN1755678A 2006.4.5

CN1710560A 2005.12.21

US2002/0152267A1 2002.10.17

US6285999B1 2001.9.4

审查员 王 洵

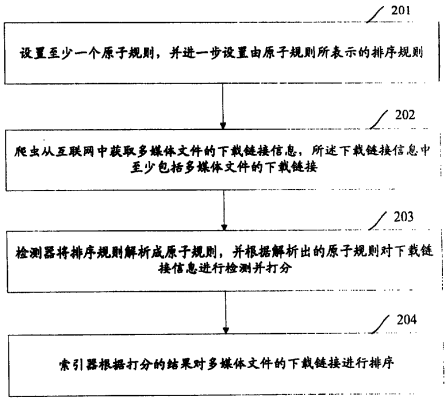
权利要求书 2 页 说明书 10 页 附图 2 页

[54] 发明名称

一种多媒体文件搜索引擎的排序方法

[57] 摘要

本发明公开了一种多媒体文件搜索引擎的排序方法，预先设置至少一个原子规则，并进一步设置由原子规则所表示的排序规则，该方法还包括以下步骤：A、爬虫从互联网中获取多媒体文件的下载链接信息，所述下载链接信息中至少包括多媒体文件的下载链接；B、检测器将排序规则解析成原子规则，并根据解析出的原子规则对所述下载链接信息进行检测并打分；C、索引器根据所述打分的结果对多媒体文件的下载链接进行排序。 应用本发明以后，针对各种欺骗能够快速地制定响应规则，从而动态地降低甚至克服搜索过程中的欺骗，并且能够对排序结果进行优化。



1、一种多媒体文件搜索引擎的排序方法，其特征在于，预先设置至少一个原子规则，并进一步设置由原子规则所表示的排序规则，该方法还包括以下步骤：

A、爬虫从互联网中获取多媒体文件的下载链接信息，所述下载链接信息中至少包括多媒体文件的下载链接；

B、检测器将排序规则解析成原子规则，并根据解析出的原子规则对所述下载链接信息进行检测并打分；

C、索引器根据所述打分的结果对多媒体文件的下载链接进行排序。

2、根据权利要求 1 所述的多媒体文件搜索引擎的排序方法，其特征在于，所述原子规则包括以下逻辑规则中的任一个或其中至少一个的任意组合：

信息百分比大于预先设定值、信息百分比包含预先设定值、信息百分比不等于预先设定值、信息百分比小于预先设定值、信息百分比等于预先设定值、丢弃信息、不丢弃信息，所述的信息百分比是指所述的下载链接信息在所有网站信息中的比例。

3、根据权利要求 1 所述的多媒体文件搜索引擎的排序方法，其特征在于，所述设置由原子规则所表示的排序规则为：将排序规则设置为原子规则的正则表达式；

步骤 B 所述检测器将排序规则解析成原子规则为：检测器分析所述正则表达式，以将所述排序规则解析成原子规则。

4、根据权利要求 1 所述的多媒体文件搜索引擎的排序方法，其特征在于，所述排序规则保存在文本文件中，

步骤 B 中检测器在初始化时从所述文本文件中调入排序规则，以对排序规则进行解析。

5、根据权利要求 4 所述的多媒体文件搜索引擎的排序方法，其特征在于，所述文本文件为可扩展标记语言 XML 文件。

6、根据权利要求1所述的多媒体文件搜索引擎的排序方法，其特征在于，所述下载链接信息中进一步包括Tag信息，所述排序规则包括:

如果所述Tag信息里面包含有链接，则不从Tag信息中获取内容并将该多媒体文件的下载链接打分降低。

7、根据权利要求1所述的多媒体文件搜索引擎的排序方法，其特征在于，所述下载链接信息中进一步包括锚文本，所述排序规则包括:

如果同一个网站的相同锚文本超过预定比例，则不从该网站的锚文本获取内容。

8、根据权利要求1所述的多媒体文件搜索引擎的排序方法，其特征在于，所述下载链接信息进一步包括锚文本和/或Tag信息。

9、根据权利要求1所述的多媒体文件搜索引擎的排序方法，其特征在于，所述多媒体文件包括音乐文件、视频文件或图像文件。

10、根据权利要求9所述的多媒体文件搜索引擎的排序方法，其特征在于，所述音乐文件包括：MP3文件、WMA文件或RM文件。

一种多媒体文件搜索引擎的排序方法

技术领域

本发明涉及搜索引擎技术领域，更具体地说，本发明涉及一种多媒体文件搜索引擎的排序方法。

背景技术

搜索引擎技术是近几年非常热门的技术，以其为核心基础的网页搜索、新闻搜索、多媒体文件搜索、地图搜索等都具有很大的实用价值和商业价值。目前，各种搜索引擎技术层出不穷，与其相关的各种搜索应用也在飞速发展当中。

通常而言，多媒体文件搜索一般包括音乐文件搜索、视频文件搜索和图片文件搜索等。音乐文件搜索引擎通常又叫 Mp3 搜索引擎，它以搜索技术为基础，检索和提供 Mp3 及其它各种格式音乐文件的信息搜索和下载统一资源描述符（URL）。同样，视频文件搜索引擎以搜索技术为基础，检索和提供 RM、WMV 及其它各种格式视频文件的信息搜索和下载 URL；图片文件搜索引擎以搜索技术为基础，检索和提供联合图像专家组（JPEG）及其它各种格式图像文件的信息搜索和 URL。

随着搜索技术的不断成熟，以及互联网用户对多媒体文件下载服务的需求不断增大，近年来多媒体文件搜索的竞争越来越激烈，技术发展也越来越快。因此，除了需要从数量上提高搜索结果（比如增加多媒体文件链接的数量、减少死链接等）以外，还必须对搜索质量进行提高，以提供给用户尽可能好的体验。

在文件搜索中需要对搜索结果进行排序，而搜索结果的排序是搜索体验中最为关键的部分之一。对于多媒体文件搜索来说，除了需要由搜索引擎搜

索出多媒体文件的URL之外，通常还需要提供一些额外的多媒体文件信息。比如，对于 Mp3 搜索引擎来说，除了提供 Mp3 文件的 URL 链接以外，还需要提供 Mp3 文件的歌曲名称、歌手名称、专辑名称等信息。再比如，对于视频文件搜索引擎来说，还需要提供视频文件的名称、演员名称等信息。保证这些信息的完整和合理排序，是一个良好的多媒体文件搜索引擎的基础。

图 1 为现有技术中的多媒体文件搜索引擎的排序示意图。首先由爬虫（Crawler）从互联网获取多媒体文件的下载链接，然后由检测器（Detector）对这些下载链接进行检测以检测出其中的活链，检测器并且对活链打分排序后送索引器，再由索引器（Index）建立查询索引，最后由用户根据所建立的索引从互联网上进行下载等直接操作。其中，排序问题基本可以转化为对搜索结果的打分问题，主要考虑两个方面：

1、对爬虫在网页上抓取的链接本身和锚文本（anchor）进行打分；

2、对 Mp3、WMA 等文件的 Tag 信息进行打分，Tag 信息为多媒体文件通常带有的歌曲名、歌手、专辑等信息。

一般来说，可以结合考虑以上两种方面来解决基本的排序问题。然而，随着搜索技术的发展，搜索欺骗（spam）技术也层出不穷，很多网站针对 Tag 信息作出了各种欺骗搜索，这样根据 Tag 进行的打分往往会不准确，会给欺骗网站打很高的分数，甚至帮助欺骗网站打广告，从而严重降低了用户体验度。

另外，由于爬虫抓取的网页下载链接和锚文本的重复几率都比较大，因此利用锚文本往往无法区分两个不相同的多媒体文件。比如，很多锚文本都是“点击”或 “试听”、“试看”等文本，利用这些信息无法区分其所对应的多媒体文件。

不仅与此，由于网页和 Tag 的欺骗手段千变万化，并且随着时间发展而更隐蔽，因此用固定的规则很难达到防止欺骗和区分重复记录的效果。

发明内容

有鉴于此，本发明的主要目的是提出一种多媒体文件搜索引擎的排序方法，以动态地降低甚至克服搜索过程中的欺骗。

为达到上述目的，本发明的技术方案是这样实现的：

一种多媒体文件搜索引擎的排序方法，预先设置至少一个原子规则，并进一步设置由原子规则所表示的排序规则，该方法还包括以下步骤：

A、爬虫从互联网中获取多媒体文件的下载链接信息，所述下载链接信息中至少包括多媒体文件的下载链接；

B、检测器将所述排序规则解析成原子规则，并根据解析出的原子规则对所述下载链接信息进行检测并打分；

C、索引器根据所述打分的结果对多媒体文件的下载链接进行排序。

所述原子规则包括以下逻辑规则中的任一个或其中至少一个的任意组合：

信息百分比大于预先设定值、信息百分比包含预先设定值、信息百分比不等于预先设定值、信息百分比小于预先设定值、信息百分比等于预先设定值、丢弃信息、不丢弃信息，其中信息百分比为某个信息在总信息中的比例。

所述设置由原子规则所表示的排序规则为：将排序规则设置为原子规则的正则表达式；

步骤 B 所述检测器将排序规则解析成原子规则为：检测器分析所述正则表达式，以将所述排序规则解析成原子规则。

所述排序规则保存在文本文件中，

步骤 B 中检测器在初始化时从所述文本文件中调入排序规则，以对排序规则进行解析。

所述文本文件为可扩展标记语言（XML）文件。

所述下载链接信息中进一步包括 Tag 信息，所述排序规则包括：

如果所述 Tag 信息里面包含有链接，则不从 Tag 信息中获取内容并将该多媒体文件的下载链接打分降低。

所述下载链接信息中进一步包括锚文本，所述排序规则包括:

如果同一个网站的相同锚文本超过预定比例，则不从该网站的锚文本获取内容。

所述下载链接信息进一步包括锚文本和/或 Tag 信息。

所述多媒体文件包括音乐文件、视频文件或图像文件。

所述音乐文件包括: MP3 文件、WMV 文件或 RM 文件。

从上述技术方案中可以看出，在本发明中，首先预先设置至少一个原子规则，并进一步设置由原子规则所表示的排序规则，然后由爬虫从互联网中获取多媒体文件的下载链接信息，所述下载链接信息中至少包括多媒体文件的下载链接；接着由检测器将排序规则解析成原子规则，并根据解析出的原子规则对下载链接信息进行检测并打分；最后由索引器根据所述打分的结果对多媒体文件的下载链接进行排序。

由此可见，本发明中的排序规则是由简单的原子规则所表示的，因此更新起来非常方便迅速，可以动态地载入排序规则，而不需要改写代码，实现了代码和规则相分离，因此针对各种欺骗能够快速地制定响应规则，从而动态地降低甚至克服搜索过程中的欺骗。

另外，在本发明中，如果同一个网站的相同锚文本超过预定比例，则不从该网站的锚文本获取内容，从而克服现有技术中的锚文本重复所带来的无法区分多媒体文件以及锚文本欺骗的缺陷。

而且，在本发明中，如果 Tag 信息里面包含有链接，则不从 Tag 信息中获取内容并将该多媒体文件的下载链接打分降低。因此，本发明可以克服与 Tag 信息相关的欺骗。

附图说明

图 1 为现有技术中的多媒体文件搜索引擎的排序示意图。

图 2 为根据本发明的多媒体文件搜索引擎的示范性排序方法。

具体实施方式

为使本发明的目的、技术方案和优点表达得更加清楚明白，下面结合附图及具体实施例对本发明再作进一步详细的说明。

本发明的主要思想是：首先预先设置至少一个原子规则，并进一步设置由原子规则所表示的排序规则，然后由爬虫从互联网中获取多媒体文件的下载链接信息，所述下载链接信息中至少包括多媒体文件的下载链接；接着由检测器将排序规则解析成原子规则，并根据解析出的原子规则对下载链接信息进行检测并打分；最后由索引器根据所述打分的结果对多媒体文件的下载链接进行排序。

图 2 为根据本发明的多媒体文件搜索引擎的示范性排序方法。如图 2 所示，该方法包括：

步骤 201：设置至少一个原子规则，并进一步设置由原子规则所表示的排序规则；

无论如何复杂的逻辑规则，都可以用简单的逻辑组合而成。可以预先规定最初的几个逻辑规则，这些基本的逻辑规则就是原子规则。比如将“大于”、“包含”、“不等于”、小于”等写成原子规则，这样几乎所有的打分操作和防欺骗操作都可以由这些原子规则组成。比如，原子规则可以包括：信息百分比大于预先设定值、信息百分比包含预先设定值、信息百分比不等于预先设定值、信息百分比小于预先设定值、信息百分比等于预先设定值、丢弃信息、不丢弃信息等。

设置好原子规则后，可以设置由原子规则所表示的排序规则。

当下载链接信息中进一步包括 Tag 信息时，排序规则可以为：如果 Tag 信息里面包含有链接，则不从 Tag 信息中获取内容，并将该多媒体文件的下载链接打分降低。应用该规则，可以避免 Tag 信息中的欺骗，并且克服由于锚文本重复所带来的无法区分多媒体文件的问题。

当下载链接信息中进一步包括锚文本时，排序规则可以为：如果同一个

网站的相同锚文本超过预定比例，则不从该网站的锚文本获取内容。应用该规则，可以克服由锚文本相同所造成的无法区分多媒体文件的问题。

优选将排序规则设置为原子规则的正则表达式。同时，优选利用 XML 存放复杂结构化的数据的优点，把排序规则变成以上的原子规则的正则表达式，然后借助正则表达式解析器就可以实现规则的文本化。

XML 是一种简单、与平台无关并被广泛采用的标准，是用来定义其它语言的一种元语言。简单的说，XML 是提供一种描述结构化数据的方法，它不但完成了 HTML 不能完成的任务，更为互联网世界提供了定义各行各业的“专业术语”的工具。XML 可用于各种不同的应用程序，其实质是：XML 是一种表示数据的方式。有时候数据是为数据库准备的，有些时候则是供人阅读的。与这两方面应用相关的技术，比如数据验证和 XML 转换也已经随着 XML 自身一起发展起来。XML 包括验证或者确认的能力、文档结构和文档（在某种意义上的）内容。可通过多种方式使用 XML 封装的数据。一种常见的处理方式是通过使用可扩展样式表语言转换（Extensible Stylesheet Language Transformations，XSLT），开发人员可以使用 XSLT 定义对 XML 文档的操作，以生成特定的结果。这种动态转换信息的能力允许从单个源文档产生多种输出，无论输出到不同的数据库还是输出到不同的浏览器。

可见，利用 XML 可以很方便地将排序规则变成以上的原子规则的正则表达式。

步骤 202：爬虫从互联网中获取多媒体文件的下载链接信息，所述下载链接信息中至少包括多媒体文件的下载链接；

在这里，当用户执行多媒体文件搜索时，爬虫首先从互联网中采集多媒体文件的下载链接信息，并且将这些下载链接信息按照一定的哈希（Hash）顺序进行保存，下载链接信息中除了多媒体文件的下载链接（URL）之外，还可以进一步包括锚文本和/或 Tag 信息。

步骤 203：检测器将排序规则解析成原子规则，并根据解析出的原子规

则对所述下载链接信息进行检测并打分;

检测器首先将排序规则解析成原子规则。当将排序规则设置为原子规则的正则表达式时，检测器首先分析所述正则表达式，以将所述排序规则解析成原子规则，然后根据解析出的原子规则对所述下载链接信息进行检测并打分。

其中，检测器对保存的下载链接信息进行检测，从中检测出活链，并且只根据解析出的原子规则对这些活链进行打分。比如，对于来自同一个网站的 URL，如果超过某一门限值的锚文本是相同的，就不从该网站的锚文本获取信息，并且将打分降低。

步骤 204：索引器根据所述打分的结果对多媒体文件的下载链接进行排序。

一般地，将打分较高的多媒体文件的下载链接排序在索引的前列，从而能够让用户更快发现。还可以将下载链接及与其相关的附加信息一起排序在索引的前列。

下面，以一个具体的实例对本发明进行更加详细的说明。比如，以 Mp3 搜索引擎为例进行说明。

假设对于某 Mp3 搜索引擎，爬虫从网络上爬行了许多 Mp3 文件的下载链接信息，这些下载链接信息包括:

1、URL 字符串，其中有可能包含歌曲名字;

2、锚文本，锚文本中有可能有歌曲名字，但是也有可能是错误甚至是欺骗;

3、Tag 信息内容，这些 Tag 信息内容中可能有欺骗。

此时，需要从以上三个信息里面提取两个信息：歌曲信息（如歌手名、歌曲名）和排序分数。

首先，定义一些原子规则。表 1 是本发明的原子规则的示范性示意表。这些原子规则包括：信息百分比大于某值、信息百分比小于某值、信息百分比等于某值、丢弃信息、不丢弃信息。其中，信息百分比大于某值、信息百

分比小于某值、信息百分比等于某值等可以归于对信息百分比（part of）的操作。信息百分比的参数包括4个，参数1为Tag、URL、或锚文本；参数2为所有信息或者网站信息；参数3为百分比数目；参数4为大于或小于或等于。比如，part of（Tag，网站，30，大于）就表示：Tag信息在对应所有属于本网站的Tag，比例是否大于30%。丢弃（Drop）用于表示是否丢弃某个信息，它的参数可以有两个，参数1为Tag、URL、或锚文本；参数2为真/假。比如，Drop（Tag，真）表示丢弃此链接的Tag的信息。

原子操作	说明	输入	输出	举例
信息百分比 PartOf	判断某个信息在总信息中的比例	参数1: Tag/URL/锚文本 参数2: 所有/网站 参数3: 百分比数目 参数4: 大于/小于/等于	真 /假	partof(Tag，网站，30，大于) 表示：Tag信息在对应所有属于本网站的Tag，比例是否大于30%
丢弃 Drop	是否丢弃某个信息	参数1:tag/url/anchor 参数2:真/假	无	Drop(Tag，真)表示丢弃此链接的Tag的信息.

表1

另外，根据一定的统计，发现如下排序规则对打分有很大作用：

排序规则1、来自同一个网站的URL，如果超过某一门限值（比如30%）的锚文本是相同的，就不从该网站的锚文本获取信息。

此规则的依据来自一个假设，就是同一个网站不可能很多的链接都是同一首歌曲。比如，某网站锚文本几乎都是“点击”，还有些网站的锚文本很多都是“试听”。而正好有首Mp3就叫“好歌试听”，如果不抛弃此锚文本信息，容易造成误判。甚至一些网站的锚文本全由知名歌曲名所组成，以

加大本网站的分数。应用本规则以后，就能够防止锚文本的重复所造成的无法区分不同 Mp3 歌曲，也能够克服与锚文本相关的欺骗。

排序规则 2、如果一个歌曲的 Tag 信息里面有链接，就不从 Tag 获取信息，并且把该歌曲的 URL 打分降低。

示范性地，将排序规则 1 写成由原子操作组成的表达式。比如，当采用 XML 文本时：

排序规则 1: 锚文本如果超过 30%相同，就丢弃锚文本信息

操作: Drop(Anchor，$1)，$1=Part of(Anchor，网站，30，大于)

XML 如下：

```
<Rule>
  <R Name="Partof" Return="$1">
      <Para> Anchor </Para>
<Para>网站</Para>
<Para>30</Para>
<Para>大于</Para>
</R>
  <R Name="Drop" Return="">
  <Para>Anchor</Para>
  <Para>$1</Para>
</R>
</Rule>
```

这样，根据 XML 就将排序规则 1 由原子规则所表示。显然，排序规则 2 类似也能由原子规则所表示。检测器启动的时候可以从 XML 中加载规则到内存中，然后每个 Mp3 链接都检查此规则，并且按照规则完成相应的功能。

如果排序规则发生改变，只需要修改 XML 并且重启进程，可以非常快的修改规则，并且不用修改与搜索相关的代码。当然，由于原子规则是灵活的，因此这些原子规则还可以组合成其他的规则，因此程序最初必须实现一

定量的原子规则，才可以起到由配置文件动态加载的作用。

根据测试，如果排序规则发生改变，在多服务器的情况下，实际运行中的更新周期可以控制在 24 小时之内，因此可以提供动态地打分结果和良好的体验，快速地防欺骗。

以上过程中，以 Mp3 搜索引擎为例子对本发明进行了详细说明。显然，这仅是示范性的，并不用于对本发明的保护范围进行限定。实质上，本发明可以适用于音乐文件搜索引擎、视频文件搜索引擎、图像文件搜索引擎等各种多媒体文件搜索引擎。优选地，在搜索音乐文件时，这些音乐文件的格式包括 MP3 文件格式、WMA 文件格式或 RM 文件格式等。

以上所述，仅为本发明的较佳实施例而已，并非用于限定本发明的保护范围。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

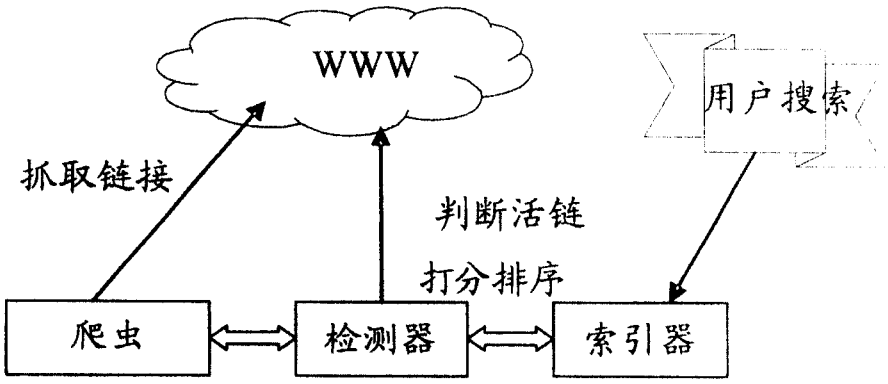


图 1

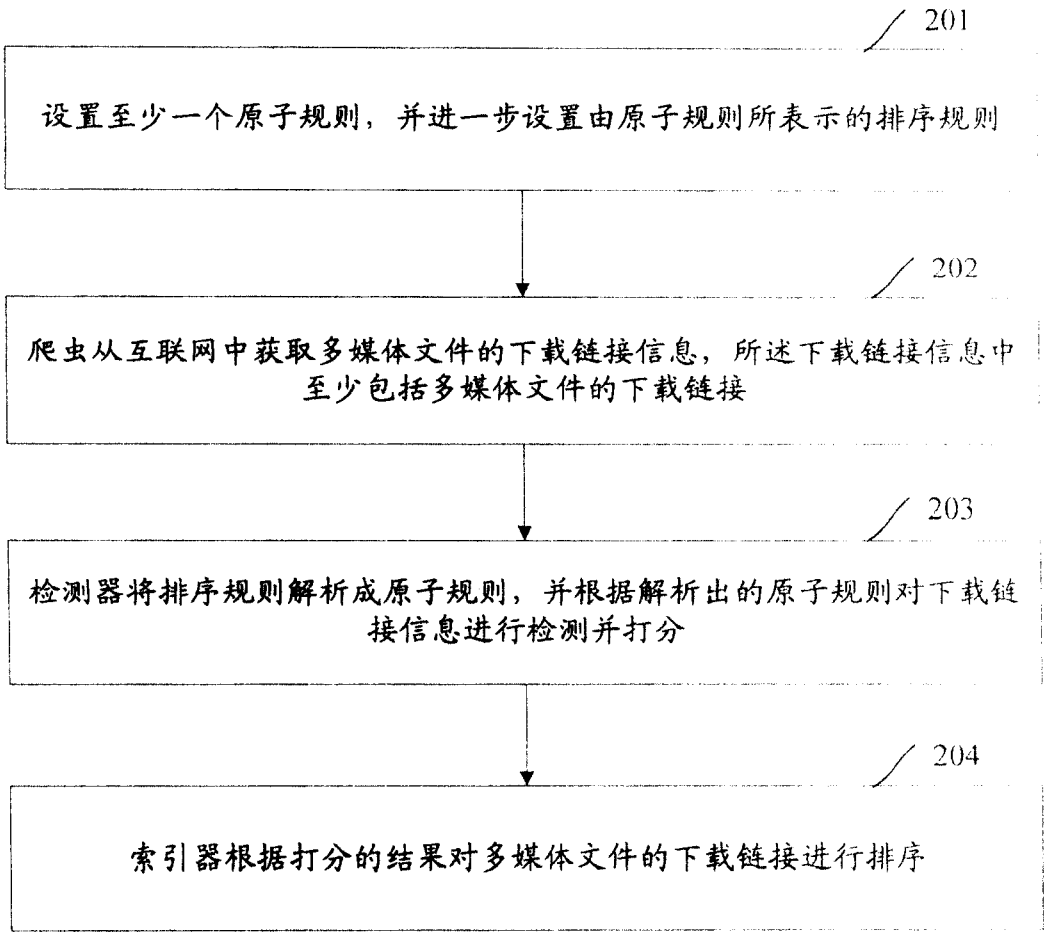

201
设置至少一个原子规则，并进一步设置由原子规则所表示的排序规则
202
爬虫从互联网中获取多媒体文件的下载链接信息，所述下载链接信息中至少包括多媒体文件的下载链接
203
检测器将排序规则解析成原子规则，并根据解析出的原子规则对下载链接信息进行检测并打分
204
索引器根据打分的结果对多媒体文件的下载链接进行排序

图 2