(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2008 年11 月6 日 (06.11.2008)

PCT

(10) 国际公布号
WO 2008/131597 A1

(51) 国际专利分类号:
G06F 17/30 (2006.01)

(21) 国际申请号: PCT/CN2007/001474

(22) 国际申请日: 2007 年4 月29 日 (29.04.2007)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人及
(72) 发明人: **林海涛(LIN, Haitao)** [CN/CN]; 中国北京市海淀区西直门北大街时代之光1号楼810室, Beijing 100044 (CN)。

(74) 代理人: **北京三友知识产权代理有限公司(BEIJING SANYOU INTELLECTUAL PROPERTY AGENCY LTD.)**; 中国北京市金融街35号国际企业大厦A座16层, Beijing 100032 (CN)。

[见续页]

(54) **Title:** SEARCH ENGINE AND METHOD FOR FILTERING AGENCY INFORMATION

(54) **发明名称:** 搜索引擎及其对中介信息的过滤方法

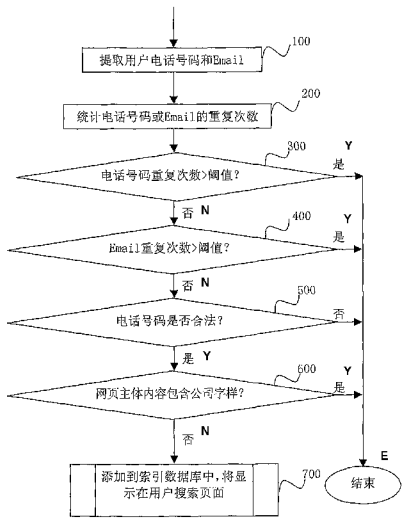

图3 / Fig . 3

00 EXTRACTING USER'S PHONE NUMBER AND EMAIL
200 COUNTING THE NUMBER OF REPETITION OF THE PHONE NUMBER OR EMAIL
300 THE NUMBER OF REPETITION OF THE PHONE NUMBER > A THRESHOLD VALUE
400 THE NUMBER OF REPETITION OF THE EMAIL > A THRESHOLD VALUE
500 IS THE PHONE NUMBER LEGAL?
600 DOES THE MAIN CONTENT OF THE WEB PAGE CONTAIN WORD OF COMPANY?
700 ADDING IT TO THE INDEX DATABASE, DISPLAYING IT ON A USER SEARCH PAGE
Y YES
N NO
E END

(57) Abstract: A search engine and a method for filtering agency information, wherein the method comprising: grasping web pages from internet, sending them to a web page database; extracting link information, extracting the titles and contents of the web pages from the database, and extracting agency feature information further; analyzing the agency feature information extracted, if the set agency information judging condition is satisfied, the information corresponding to the agency feature information is determined as agency information; filtering the agency information from the search result.

(57) 摘要:

一种搜索引擎及其对中介信息的过滤方法，该方法包括：从互联网抓取网页，送入网页数据库；进行链接信息提取，从网页数据库提取网页标题和网页内容，并从网页内容中进一步提取中介特征信息；对提取的中介特征信息进行分析，如果满足设定的中介信息判断条件，则判断该中介特征信息对应的信息为中介信息；在搜索结果中过滤掉中介信息。

WO 2008/131597 A1

(81) 指定国（除另有指明，要求每一种可提供的国家保护）：AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国（除另有指明，要求每一种可提供的地区保护）：ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), 欧洲 (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)。

本国际公布：

— 包括国际检索报告。

搜索引擎及其对中介信息的过滤方法

技术领域

本发明涉及计算机搜索引擎技术，特别涉及搜索引擎及其对中介信息的过滤方法。

背景技术

互联网提供了即时丰富的信息（以及人与人沟通参与/娱乐的平台），深层影响着现代人的生活。但随着网站数量和内容的急增， 互联网就像是没有目录的巨大百科全书，让人们无法找寻自己想要的信息。而搜索引擎的出现，为这本百科全书加上了目录和索引。只需要在搜索框中敲入关键词汇， 就能够获得相关的信息或网址。面对浩瀚的网络资源， 搜索引擎为所有网上冲浪的用户提供了一个入口，毫不夸张的说，几乎所有的用户都可以从搜索出发到达自己想去的网上任何一个地方。因此它也成为除了电子邮件以外最多人使用的网上服务。

图 1 列出了现有技术中一个典型的搜索引擎的系统架构图，搜索引擎的各部分都会相互交错相互依赖。其处理流程大致如下：

网络蜘蛛从互联网上抓取网页，抓取过程如下：(1)手工向 URL 数据库中加入一个或多个起始网页的 URL(统一资源定位符，又称为网页地址)，这些 URL 也称为种子；(2) 网络蜘蛛程序从 URL 数据库中获取一个 URL,抓取这个 URL 对应的网页内容，然后把网页内容放入网页数据库中；（3）把抓取到的网页中的满足要求的 URL 提取出，放入 URL 数据库中。判断 URL 是否满足要求的方法为模式匹配；（4）重复步骤（2）-（3），直到网页数据库不再有新的记录加入。

系统从网页数据库中取得网页原始页面，从网页中提取文本信息，即把 HTML 语法标记全部去除。然后把提取后的文本信息送入文本索引模块建立索引，建立索引的过程为首先计算页面内容中及超链中每一个关键词的相关度（或重要性），然后用这些相关信息建立网页索引数据库，形成索引数据库。文本索引

建立的过程中，需要参考网站的链接信息，主要是用来防止非法网站，例如网站自身的多重循环链接。索引数据库建立的同时，也从网页数据库进行链接信息提取，把链接信息（包括锚文本、链接本身等信息）送入链接数据库，为网页评级提供依据。

用户通过提交查询请求给查询服务器，服务器在索引数据库中进行相关网页的查找，同时网页评级把查询请求和链接信息结合起来对搜索结果进行相关度的评价，通过查询服务器按照相关度进行排序，并提取关键词的内容摘要，最后由页面生成系统将搜索结果的链接地址和页面内容摘要等内容组织起来返回给用户。

如图 1 所示的搜索引擎的系统架构中，网络蜘蛛（Spider）和链接信息提取（Parser）模块是最主要的部分。其中：

所述网络蜘蛛（Spider）使用多线程并发搜索技术，主要完成文档访问代理、路径选择引擎和访问控制引擎。网络蜘蛛（Spider）主要由 URL 服务器、爬行器、存储器、URL 解析器四大功能部件和资源库（网页数据库）、锚库、URL 数据库三大数据资源构成，另外还要借助标引器的一个辅助功能。具体过程是，URL 服务器从 URL 数据库中获取要去抓取的 URL，爬行器根据 URL 抓取 Web 页并送给存储器，存储器压缩 Web 页并存入网页数据库，然后由标引器分析每个 Web 页的所有链接并把相关的重要信息存储在锚（anchors）文件中。URL 解析器读锚文件并解析 URL，然后依次转成 docID。再把锚文本变成顺排索引，送入索引数据库。具体过程如图 2 所示，图 2 中分析器可以看成是标引器的一部分，或者说标引器的一个辅助功能部分。由于网络蜘蛛的处理流程属于公知技术，在此并不详述。

所述链接信息提取模块用于读取网页数据库，解压缩文档然后进行分析。每个文档都被转成一套单词出现频率， 称之为采样数。采样数记录单词及在文档中出现的位置，字体的大小以及大小写信息。搜索引擎有两种类型的采样数：

（1）标题：此标题为 HTML 或 URL 的标题以及 HTML 文件中的 Meta 信息。

通过分析各个单词，建立索引。用户就可以通过此索引搜索到此条信息。

（2）内容：获取页面的所有内容，通过分析各个单词，建立索引。用户就可以通过此索引搜索到此条信息。

由此可以看到，通用的搜索引擎仅仅对网页中的标题和内容进行提取并建立索引，并不对内容中的信息进一步提取。

随着搜索引擎能够获取的网页的迅速增加，用户输入搜索关键词后，往往会返回过多信息，其中包括很多无关或无用信息，用户必须从结果中进行筛选，大大影响了用户的搜索效率。因此，为了方便使用搜索引擎，使用户高效率地从搜索引擎中得到有用的信息，对搜索结果的处理就显得越来越重要。例如，在对于房屋出租信息的搜索结果，很多用户都希望过滤掉中介的信息。但目前的搜索引擎还未能解决这个问题。

发明内容

本发明实施例的目的在于提供一种搜索引擎及其对中介信息的过滤方法，使得在搜索结果过滤掉部分或全部中介信息。

为了实现上述目的，本发明提供一种搜索引擎，包括：网络蜘蛛、链接信息提取模块及查询服务器；

所述链接信息提取模块用于从网页数据库提取网页标题、网页内容及中介特征信息，并通过设定的中介信息判断条件判断该中介特征信息对应的信息是否为中介信息；

所述搜索引擎从其索引数据库中过滤掉中介信息对应的索引。

本发明还提供一种搜索引擎，包括：网络蜘蛛、链接信息提取模块及查询服务器；

所述链接信息提取模块用于从网页数据库提取网页标题、网页内容，并对所述网页内容进行分析，判断包含中介倾向信息的内容为中介信息。

所述搜索引擎从其索引数据库中过滤掉中介信息对应的索引。

本发明还提供一种搜索引擎对中介信息的过滤方法，包括：

从互联网抓取网页，送入网页数据库；

进行链接信息提取，从所述网页数据库提取网页标题和网页内容，并从网页内容中进一步提取中介特征信息；

对提取的中介特征信息进行分析，如果满足设定的中介信息判断条件，则判断该中介特征信息对应的信息为中介信息；

在搜索结果中过滤掉中介信息。

本发明还提供一种搜索引擎对中介信息的过滤方法，包括：

从互联网抓取网页，送入网页数据库；

进行链接信息提取，从所述网页数据库提取网页标题和网页内容，并对提取的网页内容进行分析，如果该网页内容中包含中介倾向信息，则判断该中介倾向信息对应的信息为中介信息；

在搜索结果中过滤掉中介信息。

本发明实施例的搜索引擎及其对中介信息的过滤方法可以过滤掉搜索结果中的部分或全部中介信息，有效防止了中介信息对用户的干扰，提高了搜索结果的可用性，为用户提供了更大的方便。

附图说明

此处所说明的附图用来提供对本发明的进一步理解，构成本申请的一部分，并不构成对本发明的限定。在附图中：

图 1 为现有技术中典型的搜索引擎的系统架构图；

图 2 为现有技术中网络蜘蛛的处理流程示意图；

图 3 为本发明实施例的过滤中介信息的流程示意图。

具体实施方式

为使本发明的目的、技术方案和优点更加清楚，下面结合附图对本发明的具体实施例进行详细说明。在此，本发明的示意性实施例及其说明用于解释本发明，但并不作为对本发明的限定。

实施例 1

如果希望搜索引擎对搜索结果进行有针对性的筛选，必须让搜索引擎“了解”页面的内容。例如，对于房屋出租等信息的搜索结果，如果希望过滤掉中介信息，则需要了解中介信息的一般特征。中介信息一般具有如下特征中的一个或多个：

（1）同一个中介会发布很多条不同的信息。以出租房屋为例，中介一般会发布很多个不同地点的租房信息。

（2）发布的信息中包含公司信息。例如公司地址和公司联系方式等。

（3）发布的信息中包含不合理的信息。例如包括不正确的电话号码（包括手机号、固定电话号码、小灵通号码等），非常低的价格等。

本发明实施例基于通用的垂直搜索，对搜索引擎的链接信息提取部分（链接信息提取模块）进行修改。本实施例中的搜索引擎主要包括网络蜘蛛（Spider）、链接信息提取模块（Parser）和查询服务器。其中，所述网络蜘蛛（Spider）和查询服务器采用通用的处理技术，在此不作详述。所述链接信息提取模块针对中介信息的特征进行了改进，除了获取网页标题和内容之外，还对内容中的信息进一步的提取，以提取用于识别中介信息的中介特征信息（如电话号码、Email 和价格等），同时也可对提取的内容进行进一步处理：通过提取并分析中介特征信息，可以找出同一个中介发布的很多条中介信息以及包含有不合理信息的中介信息；通过对提取的网页内容的分析处理，可以找出进一步的包含公司信息或其它有中介倾向的信息。

改进后的链接信息提取模块除了可获取网页标题和内容之外，还增加了如下功能：

（1） 提取用于判断中介信息的中介特征信息（以电话号码和 Email 为例进行说明）：

I. 提取用户电话号码，提取方式为模式匹配，即针对每个网页寻找“手机”、“移动电话”、“电话”、“小灵通”、“Mobile Phone”、“Cell Phone”等,

一旦发现就提取这些字符串后面的第一个连续的数字。第一个连续的数字就是用户的电话号码。

II. 提取用户 Email，提取方式为模式匹配，即针对每个网页寻找“电子邮箱”、“Email”等,一旦发现就提取这些字符串后面的连续的字符串，遇到空格停止提取。提取到的字符串就是用户的 Email。

（2） 提取电话号码和用户 Email 后，统计相同的电话号码或 Email 的重复次数。统计与时间有关，一般统计过去 n 个月（24>n>1，例如 3 个月）的电话号码或 eamil 的重复次数。

（3）对于电话号码和 eamil 的重复次数各设置一个阈值，如果超过这个阈值，就认为信息是中介发布。例如对电话号码的重复次数设置阈值为 10，当一个电话号码重复次数大于 10 时,则认为该电话号码对应的信息全部是中介信息。

（4）对于电话号码进行分析，根据号码前缀规则，判断出不存在或不合法的号码。

例如中国的网站上以 010 开始的号码，第 4 个数字必须为 5、6、8。否则，则认为这个号码对应的信息全部是中介信息。

（5）对网页主体内容进行分析，以识别中介信息。

由于中介发布的信息中有的还包含“公司”、“大量房源” 字样等具有中介倾向的信息，因此链接信息提取模块通过对提取的内容进行分析处理，可以进一步的识别出这些中介信息。例如，可对网页主体内容进行分析，如果包含“本公司”、“公司地址”、“我公司”“大量房源”等字样，则认为这条信息为中介信息。

链接信息提取模块提取如上信息后，仅对判断为非中介信息的信息建立索引，或者链接信息提取模块提取如上信息后，建立索引，但对于判定为是中介信息的所有信息从索引数据库中删除。建立索引采用的是通用的“倒排索引”技术（由于倒排索引技术是本技术领域的公知技术，在此不作详述）。

这样，索引数据库中就过滤掉了中介信息对应的索引，用户通过提交查询

请求给查询服务器，服务器在索引数据库中进行相关网页的查找，返回的搜索结果中便基本过滤掉了中介信息。

图 3 为本发明一实施例的搜索引擎对中介信息的过滤流程示意图。如图 3 所示，包括如下步骤：

步骤 100，提取中介特征信息（如电话号码和 Email），具体包括如下信息：

i. 手机号码；

ii. 固定电话号码；

iii. 小灵通号码；

iv. Email。

步骤 200，对于提取到的相同的信息进行计数。本实施例实现的方式为在搜索引擎的后台数据库中建立一个表，第一个字段为电话号码或者 Email，第二个字段为重复出现的次数。每提取一个信息后，先查询这个表，如果已经存在记录，则把相应的重复出现次数加 1；如果没有记录，则插入一个记录，把相应的重复出现次数设置为 1。

步骤 300、步骤 400，如果某个手机、固定电话、小灵通或者 Email 的重复次数多于 10 次，则不对此手机、电话、小灵通或者 Email 所对应的所有发布信息建立索引，或者把此手机、电话、小灵通或者 Email 所对应的所有发布信息从搜索引擎的索引数据库中删除。

步骤 500，判断手机、电话或小灵通号码是否合法，判断的规则是根据中国各个地方的号码规则表，例如北京的电话号码为 8 位。对于不符合规则的，把此手机、电话、小灵通所对应的所有发布信息从搜索引擎的索引数据库中删除。

步骤 600，判断提取的网页内容是否有中介倾向。如果网页内容包含“本公司”“大量房源”或者包含多个不同的地址（例如：现有东直门、西直门、中关村多处住房），则不对此条信息建立索引，或者把此条信息从搜索引擎的索引数据库中删除。

在本实施例的上述步骤 300-600 的每一步骤中，对判断为中介信息的信息也可以先不进行特殊处理，而在所有条件都判断完之后再对所有判断的中介信息从搜索引擎的索引数据库中删除；或者在所有条件都判断完之后再将非中介信息添加到索引数据库中，供用户查询使用，对判断为中介信息的信息则不建立索引。但本发明并不限于这些方式，只要能将判断的中介信息从索引数据库中过滤掉，都应涵盖在本发明的范围之内。

另外，图 3 所示的本实施例中的如上各步骤并没有先后顺序上的限制，并且，中介特征信息并不限于实施例中给出的电话号码或者 Email，还可以为价格等其它信息。

本领域普通技术人员可以理解实现上述实施例方法中的全部或部分步骤可以通过程序来指令相关的硬件来完成，该程序可以存储于一计算机可读取存储介质中，比如 ROM/RAM、磁碟、光盘等。

通过以上的处理，就可以把搜索引擎的索引数据库记录提供给查询服务器，供用户查询使用。此时，由于索引数据库中已经基本不包含中介信息的索引，因此这样处理后，搜索结果中的中介信息可由处理之前的 90%降低为 10%以下。

如上所述，本发明实施例的搜索引擎及其对中介信息的过滤方法可以过滤掉搜索结果中的部分或全部中介信息，有效防止了中介信息对用户的干扰，提高了搜索结果的可用性，为用户提供了更大的方便。

以上具体实施方式仅用于说明本发明，而非用于限定本发明。凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

权 利 要 求

1. 一种搜索引擎对中介信息的过滤方法，其特征在于，该方法包括：

从互联网抓取网页，送入网页数据库；

进行链接信息提取，从所述网页数据库提取网页标题和网页内容，并从网页内容中进一步提取中介特征信息；

对提取的中介特征信息进行分析，如果满足设定的中介信息判断条件，则判断该中介特征信息对应的信息为中介信息；

在搜索结果中过滤掉该中介信息。

2. 根据权利要求1所述的方法，其特征在于：

对所述中介特征信息的提取方式为模式匹配方式。

3. 根据权利要求1所述的方法，其特征在于：

所述的中介特征信息为电话号码及/或电子邮件信息；

对提取的中介特征信息进行分析是指：统计预定时间内网页中相同电话号码及/或电子邮件的重复次数；

所述设定的中介信息判断条件为：所述相同电话号码及/或电子邮件信息的重复次数超过各自对应的阈值。

4. 根据权利要求1所述的方法，其特征在于：

所述的中介特征信息为电话号码；

所述设定的中介信息判断条件为：所述电话号码为错误电话号码。

5. 根据权利要求1所述的方法，其特征在于：

所述中介特征信息为价格信息；

所述设定的中介信息判断条件为：所述价格低于设定的阈值。

6. 根据权利要求1-5中任意一项所述的方法，其特征在于，在搜索结果中过滤掉中介信息是指：

从搜索引擎的索引数据库中删除中介信息或者仅对判断为非中介信息的信息建立索引，以从索引数据库中过滤掉中介信息；

搜索引擎基于过滤掉中介信息的索引数据库进行检索，获得检索结果。

7. 根据权利要求1-5中任意一项所述的方法，其特征在于，该方法还包括：

对提取的网页内容进行分析，如果该网页内容中包含中介倾向信息，则判断该中介倾向信息对应的信息为中介信息。

8. 一种搜索引擎对中介信息的过滤方法，其特征在于：

从互联网抓取网页，送入网页数据库；

进行链接信息提取，从所述网页数据库提取网页标题和网页内容，并对提取的网页内容进行分析，如果该网页内容中包含中介倾向信息，则判断该中介倾向信息对应的信息为中介信息；

在搜索结果中过滤掉中介信息。

9. 根据权利要求8所述的方法，其特征在于，该方法还包括：

从网页内容中进一步提取中介特征信息；

对提取的中介特征信息进行分析，如果满足设定的中介信息判断条件，则判断该中介特征信息对应的网页信息为中介信息。

10. 根据权利要求9所述的方法，其特征在于：

对所述中介特征信息的提取方式为模式匹配方式。

11. 根据权利要求9所述的方法，其特征在于：

所述的中介特征信息为电话号码及/或电子邮件信息；

对提取的中介特征信息进行分析是指：统计预定时间内网页中相同电话号码及/或电子邮件的重复次数；

所述设定的中介信息判断条件为：所述相同电话号码及/或电子邮件信息的重复次数超过各自对应的阈值。

12. 根据权利要求9所述的方法，其特征在于：

所述的中介特征信息为电话号码；

所述设定的中介信息判断条件为：所述电话号码为错误电话号码。

13. 根据权利要求9所述的方法，其特征在于：

所述中介特征信息为价格信息；

所述设定的中介信息判断条件为：所述价格低于设定的阈值。

14. 根据权利要求8-13中任意一项所述的方法，其特征在于，在搜索结果中过滤掉中介信息是指：

从搜索引擎的索引数据库中删除中介信息或者仅对判断为非中介信息的信息建立索引，以从索引数据库中过滤掉中介信息；

搜索引擎基于过滤掉中介信息的索引数据库进行检索，获得检索结果。

15. 一种搜索引擎，包括网络蜘蛛和查询服务器，其特征在于，该搜索引擎还包括链接信息提取模块；

所述链接信息提取模块用于从网页数据库提取网页标题、网页内容及中介特征信息，并通过设定的中介信息判断条件判断该中介特征信息对应的信息是否为中介信息；

所述搜索引擎从索引数据库中过滤掉中介信息对应的索引。

16. 根据权利要求15所述的搜索引擎，其特征在于：

所述链接信息提取模块还用于对所述网页内容进行分析，判断包含中介倾向信息的内容为中介信息。

17. 根据权利要求15所述的搜索引擎，其特征在于：

所述链接信息提取模块对中介特征信息的提取方式为模式匹配方式。

18. 根据权利要求15所述的搜索引擎，其特征在于：

所述的中介特征信息为电话号码及/或电子邮件信息；

所述设定的中介信息判断条件为：所述链接信息提取模块统计的预计时间内相同电话号码及/或电子邮件信息的重复次数超过各自对应的阈值。

19. 根据权利要求15所述的搜索引擎，其特征在于：

所述的中介特征信息为电话号码；

所述设定的中介信息判断条件为：所述电话号码为错误电话号码。

20. 根据权利要求15所述的搜索引擎，其特征在于：

所述中介特征信息为价格信息；

所述设定的中介信息判断条件为：所述价格低于设定的阈值。

21. 根据权利要求15-20中任意一项所述的搜索引擎，其特征在于：

所述搜索引擎从其索引数据库中过滤掉中介信息对应的索引是指：

链接信息提取模块仅对判断为非中介信息的信息建立索引；或者

链接信息提取模块提取信息并建立索引后，对判断为是中介信息的信息从索引数据库中删除。

22. 一种搜索引擎，包括网络蜘蛛和查询服务器，其特征在于，该搜索引擎还包括链接信息提取模块；

所述链接信息提取模块用于从网页数据库提取网页标题、网页内容，并对所述网页内容进行分析，判断包含中介倾向信息的内容为中介信息。

所述搜索引擎从索引数据库中过滤掉中介信息对应的索引。

23. 根据权利引擎22所述的搜索引擎，其特征在于：

所述链接信息提取模块还用于从网页数据库提取中介特征信息，并通过设定的中介信息判断条件判断该中介特征信息对应的信息是否为中介信息。

24. 根据权利要求22所述的搜索引擎，其特征在于：

所述链接信息提取模块对中介特征信息的提取方式为模式匹配方式。

25. 根据权利要求22所述的搜索引擎，其特征在于：

所述的中介特征信息为电话号码及/或电子邮件信息；

所述设定的中介信息判断条件为：所述链接信息提取模块统计的预计时间内相同电话号码及/或电子邮件信息的重复次数超过各自对应的阈值。

26. 根据权利要求22所述的搜索引擎，其特征在于：

所述的中介特征信息为电话号码；

所述设定的中介信息判断条件为：所述电话号码为错误电话号码。

27. 根据权利要求22所述的搜索引擎，其特征在于：

所述中介特征信息为价格信息；

所述设定的中介信息判断条件为：所述价格低于设定的阈值。

28. 根据权利要求22-27中任意一项所述的搜索引擎，其特征在于：

所述搜索引擎从其索引数据库中过滤掉中介信息对应的索引是指：

链接信息提取模块仅对判断为非中介信息的信息建立索引；或者链接信息提取模块提取信息并建立索引后，对判断为是中介信息的信息从索引数据库中删除。

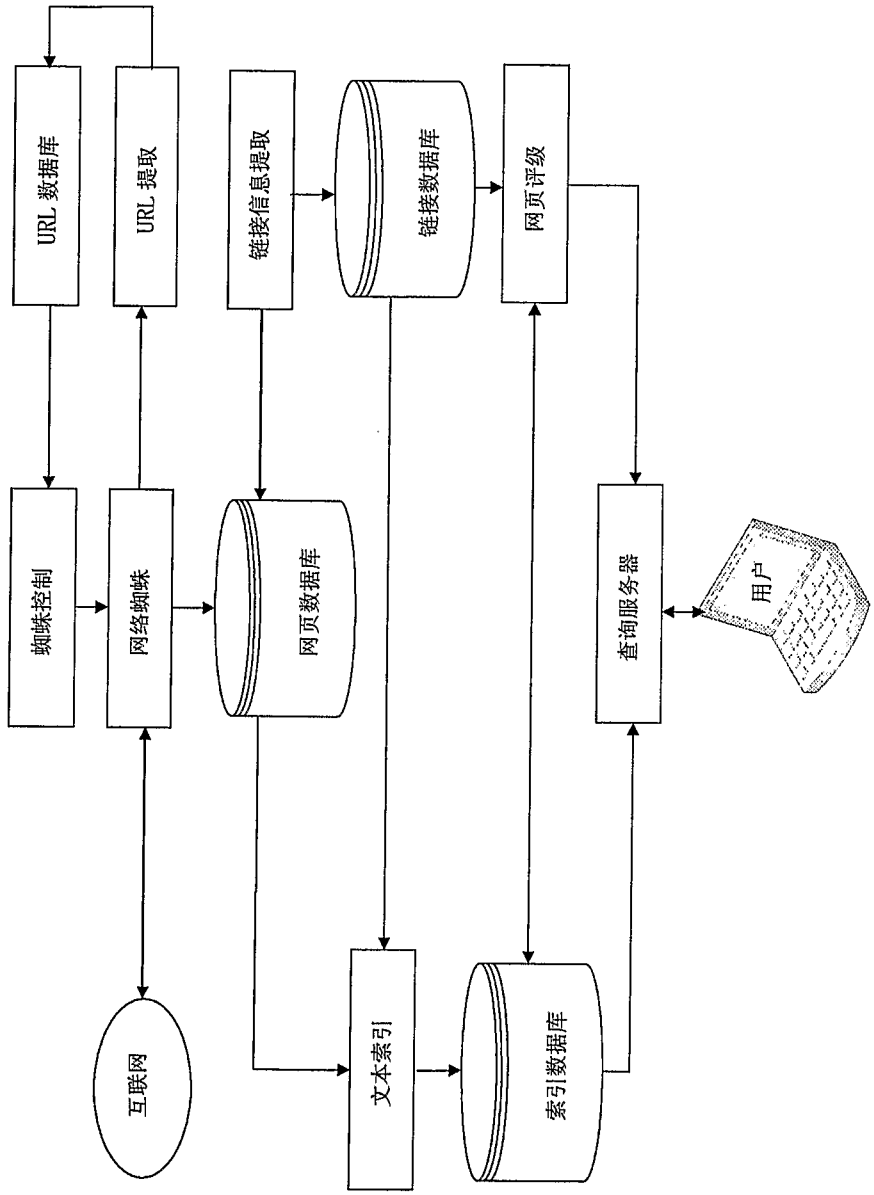
URL 数据库
URL 提取
链接信息提取
链接数据库
网页评级
蜘蛛控制
网络蜘蛛
网页数据库
查询服务器
用户
互联网
文本索引
索引数据库

图1

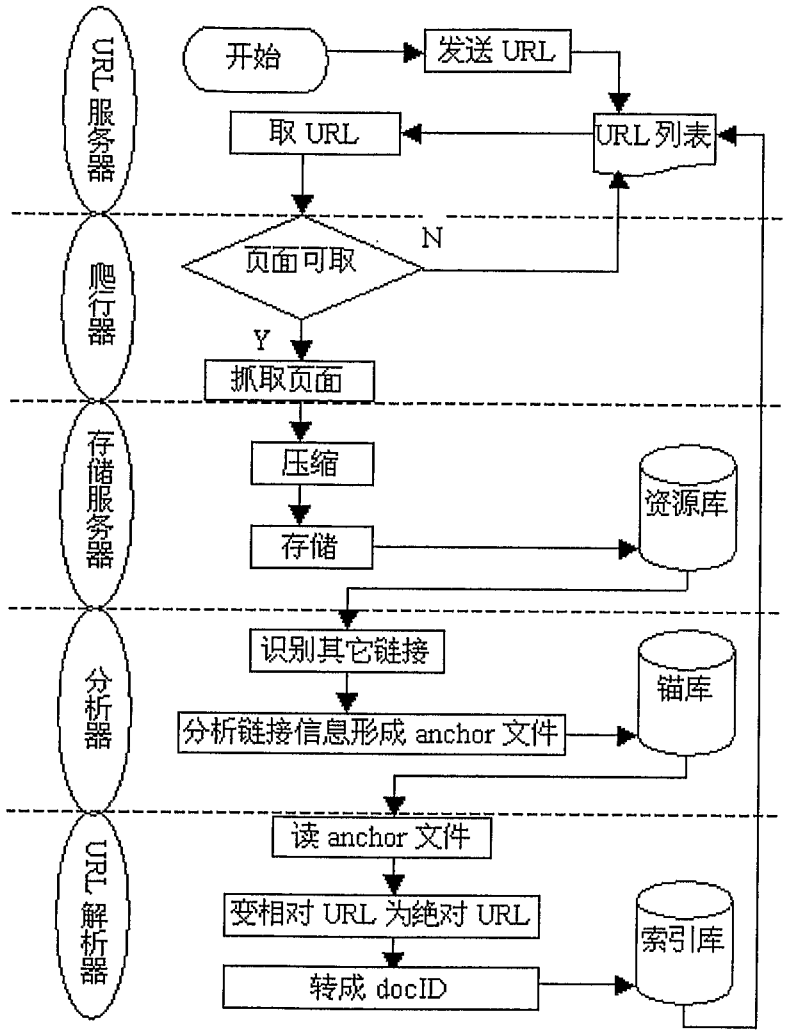
URL 服务器
爬行器
存储服务器
分析器
URL 解析器
开始
发送 URL
URL 列表
取 URL
页面可取
N
Y
抓取页面
压缩
存储
资源库
识别其它链接
分析链接信息形成 anchor 文件
锚库
读 anchor 文件
变相对 URL 为绝对 URL
转成 docID
索引库

图2

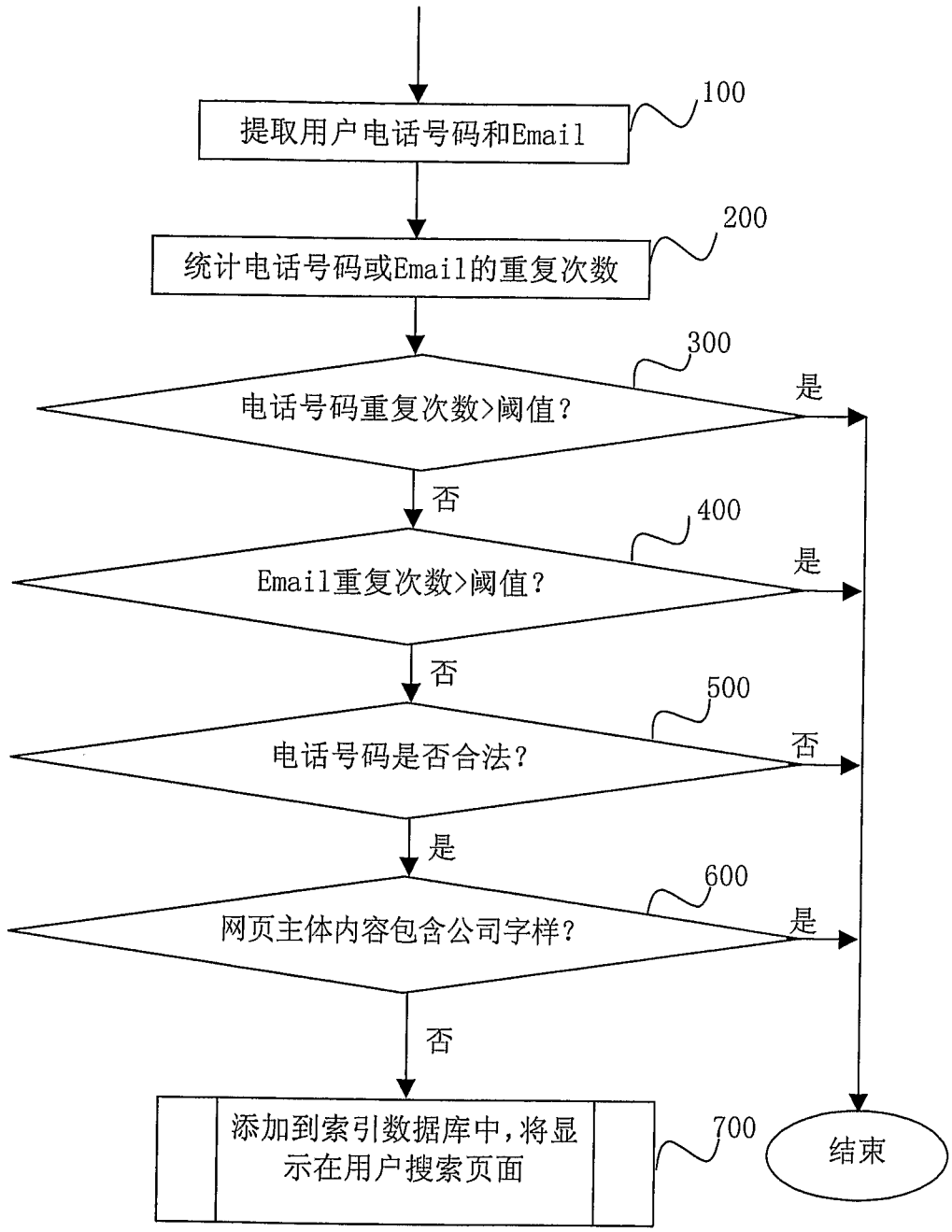
提取用户电话号码和Email
100
统计电话号码或Email的重复次数
200
电话号码重复次数>阈值?
300
是
否
Email重复次数>阈值?
400
是
否
电话号码是否合法?
500
否
是
网页主体内容包含公司字样?
600
是
否
添加到索引数据库中,将显示在用户搜索页面
700
结束

图3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2007/001474

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

IPC: G06F

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPODOC; WPI; PAJ; CNPAT;CNKI

search, engine, filter, delete, grasp, extract, pick, feature, web page, judge, determine, agency

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Zhang Maoyuan, Zou Chunyan. Research of Web Page Filter with Natural Language Processing. Computer & Digital Engineering. March 2003, vol. 31, No. 3, pages 11, 24-28, ISSN 1672-9722	1-28
A	CN1536483A (CHEN W) 13 Oct. 2004(13.10.2004) the whole document	1-28
A	US2006136411A1 (MICROSOFT CORP) 22 Jun. 2006(22.06.2006) the whole document	1-28

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim (S) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
11 Jan. 2008(11.01.2008)	**14 Feb. 2008 (14.02.2008)**
Name and mailing address of the ISA/CN The State Intellectual Property Office, the P.R.China 6 Xitucheng Rd., Jimen Bridge, Haidian District, Beijing, China 100088 Facsimile No. 86-10-62019451	Authorized officer DU, Yi Telephone No. (86-10)62411635

Form PCT/ISA/210 (second sheet) (April 2007)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2007/001474

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN1536483A	13.10.2004	NONE	
US2006136411A1	22.06.2006	NONE	

国际检索报告

国际申请号
PCT/CN2007/001474

A. 主题的分类

G06F 17/30 (2006.01) i

按照国际专利分类表(IPC)或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC: G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用）)

EPODOC; WPI; PAJ; CNPAT;CNKI

搜索,检索,引擎,过滤,滤除,抓取,提取,特征,网页,判断,中介

search, engine, filter, delete, grasp, extract, pick, feature, web page, judge, determine, agency

C. 相关文件

类　型*	引用文件，必要时，指明相关段落	相关的权利要求
X	张茂元, 邹春燕. 基于自然语言处理的网页过滤方法研究. 计算机与数字工程. 2003 年 3 月, vol. 31, No. 3, 第 11,24-28 页, ISSN 1672-9722	1-28
A	CN1536483A (陈文中) 13.10 月 2004(13.10.2004) 全文	1-28
A	US2006136411A1 (MICROSOFT CORP) 22.6 月 2006(22.06.2006) 全文	1-28

☐ 其余文件在 C 栏的续页中列出。　　☒ 见同族专利附件。

* 引用文件的具体类型：

"A" 认为不特别相关的表示了现有技术一般状态的文件

"E" 在国际申请日的当天或之后公布的在先申请或专利

"L" 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件

"O" 涉及口头公开、使用、展览或其他方式公开的文件

"P" 公布日先于国际申请日但迟于所要求的优先权日的文件

"T" 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

"X" 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

"Y" 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

"&" 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
11.1 月 2008(11.01.2008)	**14.2 月 2008 (14.02.2008)**
中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路 6 号 100088 传真号：(86-10)62019451	受权官员 杜轶 电话号码：(86-10) 62411635

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2007/001474

检索报告中引用的专利文件	公布日期	同族专利	公布日期
CN1536483A	13.10.2004	无	
US2006136411A1	22.06.2006	无	