(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 112632355 A
(43)申请公布日 2021.04.09

(21)申请号 202011354462.5

(22)申请日 2020.11.26

(71)申请人 武汉虹旭信息技术有限责任公司
地址 443000 湖北省武汉市江夏区藏龙岛谭湖二路1号虹信无线通信产业园

(72)发明人 杨志龙

(74)专利代理机构 北京路浩知识产权代理有限公司 11002
代理人 张睿

(51)Int.Cl.
G06F 16/951(2019.01)
G06F 16/953(2019.01)

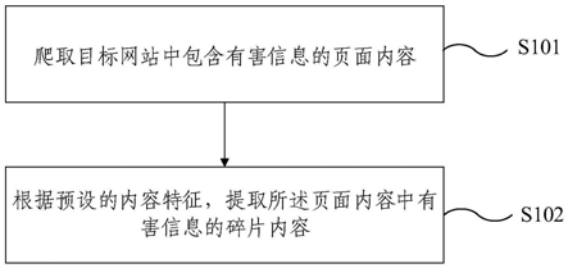
权利要求书1页 说明书10页 附图2页

(54)发明名称

有害信息的碎片内容处理方法及装置

(57)摘要

本发明提供一种有害信息的碎片内容处理方法及装置，该方法包括：爬取目标网站中包含有害信息的页面内容；根据预设的内容特征，提取所述页面内容中有害信息的碎片内容；其中，所述内容特征，包括关键词。本发明提供的有害信息的碎片内容处理方法及装置，通过爬取目标网站中包含有害信息的页面内容，根据内容特征匹配上述页面内容，获取上述页面内容包含的所有有害信息的碎片内容，能更准确的识别有害信息的碎片化内容，能更有效的阻止有害信息的传播，能降低网络的安全风险。



CN 112632355 A

1.一种有害信息的碎片内容处理方法,其特征在于,包括:

爬取目标网站中包含有害信息的页面内容;

根据预设的内容特征,提取所述页面内容中有害信息的碎片内容;

其中,所述内容特征,包括关键词。

2.根据权利要求1所述的有害信息的碎片内容处理方法,其特征在于,所述提取所述页面内容中有害信息的碎片内容之后,还包括:

确定所述有害信息的碎片内容的类型。

3.根据权利要求2所述的有害信息的碎片内容处理方法,其特征在于,所述确定所述有害信息的碎片内容的类型之后,还包括:

根据所述有害信息的碎片内容的类型,存储所述有害信息的碎片内容。

4.根据权利要求3所述的有害信息的碎片内容处理方法,其特征在于,所述根据所述有害信息的碎片内容的类型,存储所述有害信息的碎片内容之后,还包括:

响应于检索请求,根据存储的所述有害信息的碎片内容返回所述检索请求对应的检索结果。

5.根据权利要求1所述的有害信息的碎片内容处理方法,其特征在于,所述爬取目标网站中包含有害信息的页面内容,具体包括:

接收爬取指令后,根据所述爬取指令,爬取所述目标网站中包含有害信息的页面内容。

6.根据权利要求1至5任一所述的有害信息的碎片内容处理方法,其特征在于,所述根据预设的内容特征,提取所述页面内容中有害信息的碎片内容之前,还包括:

接收并存储所述预设的内容特征。

7.根据权利要求2至4任一所述的有害信息的碎片内容处理方法,其特征在于,所述确定所述有害信息的碎片内容的类型之后,还包括:

根据各有害信息的碎片内容的类型,对所述各有害信息的碎片内容进行分析。

8.一种有害信息的碎片内容处理装置,其特征在于,包括:

爬取模块,用于爬取目标网站中包含有害信息的页面内容;

提取模块,用于根据预设的内容特征,提取所述页面内容中有害信息的碎片内容;

其中,所述内容特征,包括关键词。

9.一种电子设备,包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序,其特征在于,所述处理器执行所述程序时实现如权利要求1至7任一项所述有害信息的碎片内容处理方法的步骤。

10.一种非暂态计算机可读存储介质,其上存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现如权利要求1至7任一项所述有害信息的碎片内容处理方法的步骤。

有害信息的碎片内容处理方法及装置

技术领域

[0001] 本发明涉及移动互联网技术领域，尤其涉及一种有害信息的碎片内容处理方法及装置。

背景技术

[0002] 移动互联网中存在的有害信息，指可能对现存法律秩序和其他公共秩序造成破坏或者威胁的数据。

[0003] 随着网络科学技术的发展，不法分子将有害信息分解为碎片化的信息片段，通过有害信息的碎片内容的传播，扩大有害信息的传播范围，进而达到非法的目的。

[0004] 现有技术通过对有害信息进行识别，实现对有害信息的过滤及管理。但是，有害信息的碎片内容具有乱序和无规则的特点，使得现有技术难以准确识别有害信息的碎片内容，进而无法实现对有害信息的碎片内容的有效过滤及管理。

发明内容

[0005] 本发明提供一种有害信息的碎片内容处理方法及装置，用以解决现有技术中难以准确的识别有害信息的碎片内容的缺陷，实现更准确地识别有害信息的碎片内容。

[0006] 本发明提供一种有害信息的碎片内容处理方法，包括：

[0007] 爬取目标网站中包含有害信息的页面内容；

[0008] 根据预设的内容特征，提取所述页面内容中有害信息的碎片内容；

[0009] 其中，所述内容特征，包括关键词。

[0010] 根据本发明提供一种有害信息的碎片内容处理方法，所述提取所述页面内容中有害信息的碎片内容之后，还包括：

[0011] 确定所述有害信息的碎片内容的类型。

[0012] 根据本发明提供一种有害信息的碎片内容处理方法，所述确定所述有害信息的碎片内容的类型之后，还包括：

[0013] 根据所述有害信息的碎片内容的类型，存储所述有害信息的碎片内容。

[0014] 根据本发明提供一种有害信息的碎片内容处理方法，所述根据所述有害信息的碎片内容的类型，存储所述有害信息的碎片内容之后，还包括：

[0015] 响应于检索请求，根据存储的所述有害信息的碎片内容返回所述检索请求对应的检索结果。

[0016] 根据本发明提供一种有害信息的碎片内容处理方法，所述爬取目标网站中包含有害信息的页面内容，具体包括：

[0017] 接收爬取指令后，根据所述爬取指令，爬取所述目标网站中包含有害信息的页面内容。

[0018] 根据本发明提供一种有害信息的碎片内容处理方法，所述根据预设的内容特征，提取所述页面内容中有害信息的碎片内容之前，还包括：

[0019] 接收并存储所述预设的内容特征。

[0020] 根据本发明提供一种有害信息的碎片内容处理方法，所述确定所述有害信息的碎片内容的类型之后，还包括：

[0021] 根据各有害信息的碎片内容的类型，对所述各有害信息的碎片内容进行分析。

[0022] 本发明还提供一种有害信息的碎片内容处理装置，包括：

[0023] 爬取模块，用于爬取目标网站中包含有害信息的页面内容；

[0024] 提取模块，用于根据预设的内容特征，提取所述页面内容中有害信息的碎片内容；

[0025] 其中，所述内容特征，包括关键词。

[0026] 本发明还提供一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现如上述任一种所述有害信息的碎片内容处理方法的步骤。

[0027] 本发明还提供一种非暂态计算机可读存储介质，其上存储有计算机程序，该计算机程序被处理器执行时实现如上述任一种所述有害信息的碎片内容处理方法的步骤。

[0028] 本发明提供的有害信息的碎片内容处理方法及装置，通过爬取目标网站中包含有害信息的页面内容，根据内容特征匹配上述页面内容，获取上述页面内容包含的所有有害信息的碎片内容，能更准确的识别有害信息的碎片化内容，能更有效的阻止有害信息的传播，能降低网络的安全风险。

附图说明

[0029] 为了更清楚地说明本发明或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作一简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0030] 图1是本发明提供的有害信息的碎片内容处理方法的流程示意图；

[0031] 图2是本发明提供的有害信息的碎片内容处理装置的结构示意图；

[0032] 图3是本发明提供的电子设备的结构示意图。

具体实施方式

[0033] 为使本发明的目的、技术方案和优点更加清楚，下面将结合本发明中的附图，对本发明中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0034] 在本发明的描述中，需要说明的是，除非另有明确的规定和限定，术语“安装”、“相连”、“连接”应做广义理解，例如，可以是固定连接，也可以是可拆卸连接，或一体地连接；可以是机械连接，也可以是电连接；可以是直接相连，也可以通过中间媒介间接相连，可以是两个元件内部的连通。对于本领域的普通技术人员而言，可以具体情况理解上述术语在本发明中的具体含义。

[0035] 为了克服现有技术的上述问题，本发明提供一种有害信息的碎片内容处理方法及装置，其发明构思是，爬取目标网站中包含有害信息的页面内容后，根据内容特征提取页面

内容中有害信息的碎片内容，能更准确的识别有害信息的碎片化内容。

[0036] 图1是本发明提供的有害信息的碎片内容处理方法的流程示意图。下面结合图1描述本发明的有害信息的碎片内容处理方法。如图 1所示，该方法包括：步骤S101、爬取目标网站中包含有害信息的页面内容。

[0037] 有害信息，指在计算机信息系统及其存储介质中存在或出现的，含有危害社会秩序的内容。

[0038] 由于网络的空间特殊性，网站中有害信息存储于网站的服务器中，有害信息在被访问前不易为人所知。

[0039] 本发明实施例通过基于目标数据模式的网络爬虫爬取目标网站中包含有害信息的页面内容。

[0040] 需要说明的是，网络爬虫可以针对网页上的数据，选择性地抓取符合一定的模式的页面内容。

[0041] 具体地，可以设定有害信息对应的相关敏感词，根据上述相关敏感词爬取目标网站中包含有害信息的相关敏感词的页面内容。例如：可以设定上述敏感词为“赌”、“日赚”或“刷单”等。有害信息的相关敏感词可以为一个，也可以为多个。

[0042] 需要说明的是，为了提高识别有害信息的碎片内容的准确率，通过网络爬虫爬取的包含有害信息的页面内容，可以指目标网站中包含有害信息的相关敏感词的页面中的所有内容，例如：目标网站中，包括“刷单”字段的页面中的所有内容。

[0043] 为了提高识别有害信息的碎片内容的识别效率，通过网络爬虫爬取的包含有害信息的页面内容，还可以指目标网站中包含有害信息相关敏感词的一部分页面内容，例如：目标网站中，包括“赌”字的前后300字的页面内容。

[0044] 步骤S102、根据预设的内容特征，提取页面内容中有害信息的碎片内容。

[0045] 其中，内容特征，包括关键词。

[0046] 有害信息的碎片内容，可以指混杂在正常信息中的有害信息，还可以指被分解为碎片化的有害信息的信息片段，还可以是一段包含有害信息的URL。

[0047] 有害信息的碎片内容与其他内容混杂在一起，可以将有害信息伪装为正常信息，使得有害信息的传播更加隐秘，准确识别有害信息更加困难。

[0048] 本发明实施例可以根据预设的内容特征，提取通过网络爬虫爬取的目标网站中包含有害信息的页面内容中有害信息的碎片内容。

[0049] 具体地，通过查找页面内容，可以获取页面内容中，与预设的内容特征相匹配的内容，可以提取上述与预设的内容特征相匹配的内容作为有害信息的碎片内容。

[0050] 预设的内容特征，可以包括关键词。

[0051] 关键词可以匹配有害信息的碎片内容。通过查找页面内容，可以获取页面内容中所有包含关键词的句子或一段URL，提取上述包含关键词的句子、句子所在的段落或一段URL，作为有害信息的碎片内容。

[0052] 关键词匹配有害信息的碎片内容的规则可以是：在预设的字数阈值内，同时出现组成关键词的关键字。例如：若关键词为“赌博”，在页面内容中任意20个字内，同时出现关键字“赌”和关键字“博”，则可以认为关键词匹配到有害信息的碎片内容。具体地，若在20个字内的页面内容中包含有：“赌博网站地址”或“赌一个博的网络小站地址”的句子，则可以

提取的有害信息的碎片内容为:包含有“赌博网站地址”或“赌一个博的网络小站地址”的句子或段落。预设的字数阈值可以根据实际情况确定,在本发明实施例中不作具体限定。

[0053] 需要说明的是,关键词可以根据实际情况确定,在本发明实施例中不作具体限定。

[0054] 需要说明的是,页面内容中包含关键词的句子所在的段落,可能包含其他的非法信息,为了提高识别有害信息的碎片内容的准确率,可以根据实际情况,提取句子或句子所在的段落,作为有害信息的碎片内容。

[0055] 关键词匹配有害信息的碎片内容的规则还可以是,在页面内容对应的一段URL中,同时出现组成关键词的关键字。

[0056] 关键词匹配有害信息的碎片内容的规则可以通过参考有害信息被分解的方式获得。分解有害信息的方式可以包括但不限于:在有害信息中添加其他文字或符号、调整有害信息的文字顺序或结合上述两种方式。

[0057] 预设的内容特征,还可以包括有害特征图形。有害特征图形可以匹配有害信息的关键内容。

[0058] 通过查找页面内容,可以获取页面内容中所有包含有害特征图形的图片,提取上述图片或图片及图片前后的文字,作为有害信息的碎片内容。

[0059] 查找页面内容时,还可以结合关键词和有害特征图形,获取页面内容中所有包含关键词和有害特征图形的内容,作为有害信息的碎片内容。

[0060] 需要说明的是,提取页面内容中有害信息的碎片内容后,还可以获取上述有害信息的碎片内容所在页面的页面信息,包括:页面的地址等。

[0061] 需要说明的是,可以采用开源的JavaScript技术提取页面内容中有害信息的碎片内容。

[0062] 需要说明的是,本发明实施例提供的有害信息的碎片内容的提取方法适用于基于互联网的有害信息的碎片内容的提取,尤其适用于基于移动互联网的有害信息的碎片内容的提取。

[0063] 本发明实施例通过爬取目标网站中包含有害信息的页面内容,根据内容特征匹配上述页面内容,获取上述页面内容包含的所有有害信息的碎片内容,能更准确的识别有害信息的碎片化内容,能更有效的阻止有害信息的传播,能降低网络的安全风险。

[0064] 基于上述各实施例的内容,提取页面内容中有害信息的碎片内容之后,还包括:确定有害信息的碎片内容的类型。

[0065] 有害信息的碎片内容具有乱序和无规则的特点,无法直接对有害信息的碎片内容进行分析和管理。

[0066] 本发明实施例提取页面内容中有害信息的碎片内容后,可以将有害信息的碎片内容分类。

[0067] 根据不同的分类标准,可以将有害信息的碎片内容分类。分类标准在本发明实施例中不作具体限定,以下为几种有害信息的碎片内容的具体分类标准。

[0068] 可以根据预设的内容特征进行分类。每个预设的内容特征包括的关键词可以对应若干个类型,例如:关键词“赌博”可以对应“违法类有害信息”,关键词“刷单”可以对应“诈骗类有害信息”和“违法有害信息”。根据关键词的类型,可以确定关键词对应的有害信息的碎片内容的类型。

[0069] 可以根据目标网站中的不同页面进行分类。提取页面内容中有害信息的碎片内容后，可以以目标网站中的一个页面作为一个类型，对有害信息的碎片内容进行分类。

[0070] 可以根据不同的目标网站进行分类。对多个目标网站进行有害信息的碎片内容的提取后，可以以一个目标网站作为一个类型，对有害信息的碎片内容进行分类。

[0071] 确定有害信息的碎片内容的类型后，还可以进一步细化每一类型中有害信息的碎片内容的类型，确定有害信息的碎片内容的子类型。例如：根据不同的目标网站对有害信息的碎片内容进行分类后，还可以在每一目标网站类型中，根据关键词的类型，确定有害信息的碎片内容的子类型。

[0072] 需要说明的是，可以采用开源的JavaScript技术确定有害信息的碎片内容的类型。

[0073] 本发明实施例根据不同维度对乱序和无规则的有害信息的碎片内容分类，确定有害信息的碎片内容的类型，从而能将乱序和无规则的有害信息的碎片内容变为有序，能为有害信息的碎片内容的分析和管理提供数据基础。

[0074] 基于上述各实施例的内容，确定有害信息的碎片内容的类型之后，还包括：根据有害信息的碎片内容的类型，存储有害信息的碎片内容。

[0075] 确定有害信息的碎片内容的类型后，可以根据有害信息的碎片内容的类型，将有害信息的碎片内容分别存储于每一类型的数据集。

[0076] 存储有害信息的碎片内容时，还可以存储有害信息的碎片内容所在页面的页面信息。

[0077] 需要说明的是，若同一有害信息的碎片内容对应多个类型，则可以在上述多个类型中每一类型的数据集，分别存储上述有害信息的碎片内容。

[0078] 需要说明的是，可以采用mysql数据库或hbase分布式系统存储有害信息的碎片内容。

[0079] 本发明实施例根据有害信息的碎片内容的类型，存储有害信息的碎片内容，能建立有害信息的碎片内容的分类数据集，能将乱序和无规则的有害信息的碎片内容变为有序，能为有害信息的碎片内容的检索提供支持，能为有害信息的碎片内容的分析和管理提供数据基础。

[0080] 基于上述各实施例的内容，根据有害信息的碎片内容的类型，存储有害信息的碎片内容之后，还包括：响应于检索请求，根据存储的所述有害信息的碎片内容返回检索请求对应的检索结果。

[0081] 需要说明的是，本发明实施例的执行主体是服务器。

[0082] 具体地，客户端可以向服务器发送检索请求。服务器接收客户端发送的检索请求后，可以根据检索请求返回检索请求对应的信息。

[0083] 检索请求可以是携带相关检索条件的请求。

[0084] 检索条件可以是根据已知类型或页面信息，检索已知类型或页面信息对应的有害信息的碎片内容。

[0085] 检索条件还可以是根据已知有害信息的碎片内容，检索已知有害信息的碎片内容对应的类型或页面信息。

[0086] 检索请求的具体内容可以根据实际情况确定，在本发明实施例中不作具体限制。

[0087] 检索结果可以是有害信息的碎片内容。

[0088] 检索结果还可以是有害信息的碎片内容的类型或存在有害信息的碎片内容的页面的页面信息。

[0089] 需要说明的是,可以采用开源的JavaScript技术根据检索请求返回检索请求对应的信息。

[0090] 需要说明的是,客户端可以为PC机。PC机的操作系统可以为 Windows XP系统、Windows 7系统或Windows 8系统,可以支持 Firefox或chrome等浏览器。

[0091] 上述PC机包括检索模块,可以通过检索模块发送检索请求。

[0092] 具体地,检索模块可以基于浏览器发送检索请求。

[0093] 可以在检索模块中输入检索请求,并通过检索模块将检索请求发送至服务器。

[0094] 需要说明的是,服务器还可以接收从外设输入的检索请求。

[0095] 本发明实施例通过接收检索请求,根据检索请求返回对应的信息,能根据需求获得所需的信息,能通过检索获得的信息对有害信息的碎片内容进行分析。

[0096] 基于上述各实施例的内容,爬取目标网站中包含有害信息的页面内容,具体包括:接收爬取指令后,根据爬取指令,爬取目标网站中包含有害信息的页面内容。

[0097] 具体地,客户端可以向服务器发送爬取指令。服务器接收客户端发送的爬取指令后,根据爬取指令,通过基于目标数据模式的网络爬虫爬取目标网站中包含有害信息的页面内容。

[0098] 爬取指令中可以携带预先设定的有害信息的相关敏感词。

[0099] 爬取指令中还可以携带执行爬取任务的其他爬取条件,包括:爬取开始时间、爬取结束时间、爬取周期和爬取范围等中的任意若干个。

[0100] 例如:若爬取指令携带的爬取开始时间为当日0时,爬取周期为 48小时,爬取范围为目标网站A、B和C,敏感词为“赌”,则可以执行自当日0时开始,在48小时内,实时爬取目标网站A、B和C 中包含敏感词为“赌”的有害信息的页面内容的爬取任务;若爬取指令携带的爬取开始时间为每日0时,爬取结束时间为每日早晨6时,爬取周期为7天,爬取范围为目标网站A,敏感词为“刷单”,则可以执行每天0时至早晨6时,连续7天爬取目标网站A中包含敏感词为“刷单”的有害信息的页面内容的爬取任务。

[0101] 本发明实施例根据接收的爬取指令,爬取目标网站中包含有害信息的页面内容,能获取可能包含有害信息的碎片内容的页面内容,能缩小识别有害信息的碎片内容的识别范围,能为识别有害信息的碎片内容提供数据基础。

[0102] 基于上述各实施例的内容,根据预设的内容特征,提取页面内容中有害信息的碎片内容之前,还包括:接收并存储预设的内容特征。

[0103] 具体地,客户端中的检索模块可以向服务器发送预设的内容特征。

[0104] 检索模块可以基于浏览器发送预设的内容特征。

[0105] 可以在检索模块中输入预设的内容特征,并通过检索模块将上述预设的内容特征发送至服务器。

[0106] 需要说明的是,预设的内容特征可以根据实际需求自定义获得。

[0107] 服务器可以接收客户端发送的预设的内容特征,并将上述预设的内容特征存储至存储库中。

[0108] 需要说明的是,服务器还可以接收从外设输入的预设的内容特征。

[0109] 需要说明的是,服务器可以采用mysql数据库或hbase分布式系统存储上述预设的内容特征。

[0110] 本发明实施例通过接收并存储预设的内容特征,能使得服务器根据预设的内容特征获取对应的有害信息的碎片内容,能更准确的识别有害信息的碎片内容。

[0111] 基于上述各实施例的内容,确定有害信息的碎片内容的类型之后,还包括:根据各有害信息的碎片内容的类型,对各有害信息的碎片内容进行分析。

[0112] 具体地,可以根据各有害信息的碎片内容的类型,对各类型的有害信息的碎片内容进行统计。通过对各类型的有害信息的碎片内容进行统计,可以获取每一类型或每一类型中的每一子类型中有害信息的碎片内容的数量。

[0113] 根据各类型有害信息的碎片内容的数量,可以分析有害信息的碎片内容之间的关联性。以下为几种有害信息的碎片内容的具体分析方法。

[0114] 获取不同目标网站中有害信息的碎片内容的数量后,若某目标网站提取到的有害信息的碎片内容数量较多,则说明该目标网站存在更大的安全风险。进一步地,对存在更大安全风险的目标网站可以优先进行相应的处理。

[0115] 获取同一目标网站中不同页面中有害信息的碎片内容的数量后,若某页面提取到的有害信息的碎片内容数量较多,则说明该页面存在更大的安全风险。进一步地,对存在更大安全风险的页面可以优先进行相应的处理。

[0116] 获取同一目标网站中不同类型的有害信息的碎片内容的数量后,若目标网站提取到的某一类型的有害信息的碎片内容数量较多,则说明该类型为目标网站存在安全风险的类型。进一步地,可以根据存在安全风险的类型进行相应的处理。

[0117] 获取同一目标网站中不同类型的有害信息的碎片内容的数量后,若目标网站提取到的多个类型有害信息的碎片内容数量较多,则说明上述多个类型之间存在关联。进一步地,可以根据类型之间的关联进行相应的处理。

[0118] 需要说明的是,根据各类型中有害信息的碎片内容的数量,分析有害信息的碎片内容之间的关联性可以不限于上述举例说明的情况。

[0119] 本发明实施例根据各有害信息的碎片内容的类型,对各有害信息的碎片内容进行分析,能获得针对乱序和无规则的有害信息的碎片内容的分析结果,能根据分析结果作出有针对性的处理,能实现对有害信息的碎片内容的更有效的管理,能更有效的阻止有害信息的传播,能降低网络的安全风险。

[0120] 图2是本发明提供的有害信息的碎片内容处理装置的结构示意图。下面结合图2对本发明提供的有害信息的碎片内容处理装置进行描述,下文描述的有害信息的碎片内容处理装置与上文描述的有害信息的碎片内容处理方法可相互对应参照。如图2所示,该装置包括:爬取模块201和提取模块202,其中:

[0121] 爬取模块201,用于爬取目标网站中包含有害信息的页面内容。

[0122] 提取模块202,用于根据预设的内容特征,提取页面内容中有害信息的碎片内容。

[0123] 其中,内容特征,包括关键词。

[0124] 具体地,爬取模块201和提取模块202电连接。

[0125] 爬取模块201可以通过基于目标数据模式的网络爬虫爬取目标网站中包含有害信

息的页面内容。

[0126] 具体地，可以设定有害信息对应的相关敏感词，根据上述相关敏感词爬取目标网站中包含有害信息的相关敏感词的页面内容。例如：可以设定上述敏感词为“赌”、“日赚”或“刷单”等。有害信息的相关敏感词可以为一个，也可以为多个。

[0127] 需要说明的是，为了提高识别有害信息的碎片内容的准确率，通过网络爬虫爬取的包含有害信息的页面内容，可以指目标网站中包含有害信息的相关敏感词的页面中的所有内容，例如：目标网站中，包括“刷单”字段的页面中的所有内容。

[0128] 为了提高识别有害信息的碎片内容的识别效率，通过网络爬虫爬取的包含有害信息的页面内容，还可以指目标网站中包含有害信息相关敏感词的一部分页面内容，例如：目标网站中，包括“赌”字的前后300字的页面内容。

[0129] 爬取模块201还可以用于接收客户端发送的爬取指令。

[0130] 提取模块202可以接收预设的内容特征后，根据预设的内容特征，提取通过网络爬虫爬取的目标网站中包含有害信息的页面内容中有害信息的碎片内容。

[0131] 预设的内容特征，可以包括关键词。

[0132] 关键词可以匹配有害信息的碎片内容。通过查找页面内容，可以获取页面内容中所有包含关键词的句子或一段URL，提取上述包含关键词的句子、句子所在的段落或一段URL，作为有害信息的碎片内容。

[0133] 关键词匹配有害信息的碎片内容的规则可以是：在预设的字数阈值内，同时出现组成关键词的关键字。

[0134] 需要说明的是，关键词可以根据实际情况确定，在本发明实施例中不作具体限定。

[0135] 需要说明的是，页面内容中包含关键词的句子所在的段落，可能包含其他的非法信息，为了提高识别有害信息的碎片内容的准确率，可以根据实际情况，提取句子或句子所在的段落，作为有害信息的碎片内容。

[0136] 关键词匹配有害信息的碎片内容的规则还可以是，在页面内容对应的一段URL中，同时出现组成关键词的关键字。

[0137] 预设的内容特征，还可以包括有害特征图形。

[0138] 通过查找页面内容，可以获取页面内容中所有包含有害特征图形的图片，提取上述图片或图片及图片前后的文字，作为有害信息的碎片内容。

[0139] 查找页面内容时，还可以结合关键词和有害特征图形，获取页面内容中所有包含关键词和有害特征图形的内容，作为有害信息的碎片内容。

[0140] 提取模块202还可以用于确定有害信息的碎片内容的类型。

[0141] 需要说明的是，本发明实施例的有害信息的碎片内容处理装置还可以包括存储模块。

[0142] 存储模块，可以用于根据有害信息的碎片内容的类型，存储有害信息的碎片内容。

[0143] 存储模块，还可以用于存储预设的内容特征。

[0144] 存储模块，还可以用于响应于检索请求，根据存储的有害信息的碎片内容返回检索请求对应的检索结果。

[0145] 本发明实施例通过爬取目标网站中包含有害信息的页面内容后，查找上述页面内容，获取页面内容中所有包含关键词的词句，提取上述词句或词句所在的段落，作为有害信

息的碎片内容，能更准确的识别有害信息的碎片化内容，能更有效的阻止有害信息的传播，能降低网络的安全风险。

[0146] 图3示例了一种电子设备的实体结构示意图，如图3所示，该电子设备可以包括：处理器(processor)310、通信接口(Communications Interface)320、存储器(memory)330和通信总线340，其中，处理器310，通信接口320，存储器330通过通信总线340完成相互间的通信。处理器310可以调用存储器330中的逻辑指令，以执行有害信息的碎片内容处理方法，该方法包括：爬取目标网站中包含有害信息的页面内容；根据预设的内容特征，提取页面内容中有害信息的碎片内容；其中，内容特征，包括关键词。

[0147] 此外，上述的存储器330中的逻辑指令可以通过软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0148] 另一方面，本发明还提供一种计算机程序产品，所述计算机程序产品包括存储在非暂态计算机可读存储介质上的计算机程序，所述计算机程序包括程序指令，当所述程序指令被计算机执行时，计算机能够执行上述各方法所提供的有害信息的碎片内容处理方法，该方法包括：爬取目标网站中包含有害信息的页面内容；根据预设的内容特征，提取页面内容中有害信息的碎片内容；其中，内容特征，包括关键词。

[0149] 又一方面，本发明还提供一种非暂态计算机可读存储介质，其上存储有计算机程序，该计算机程序被处理器执行时实现以执行上述各提供的有害信息的碎片内容处理方法，该方法包括：爬取目标网站中包含有害信息的页面内容；根据预设的内容特征，提取页面内容中有害信息的碎片内容；其中，内容特征，包括关键词。

[0150] 以上所描述的装置实施例仅仅是示意性的，其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性的劳动的情况下，即可以理解并实施。

[0151] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到各实施方式可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件。基于这样的理解，上述技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品可以存储在计算机可读存储介质中，如ROM/RAM、磁碟、光盘等，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行各个实施例或者实施例的某些部分所述的方法。

[0152] 最后应说明的是：以上实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；

而这些修改或者替换,并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

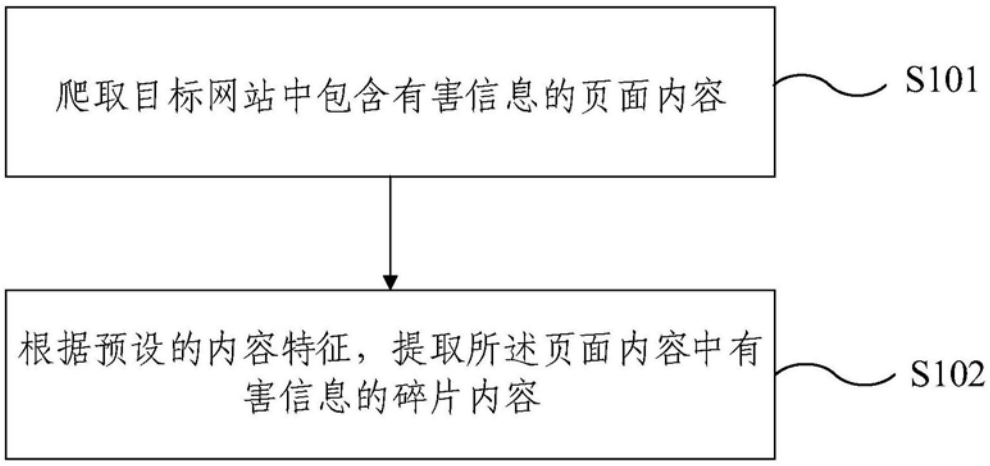

爬取目标网站中包含有害信息的页面内容
S101
根据预设的内容特征，提取所述页面内容中有害信息的碎片内容
S102

图1

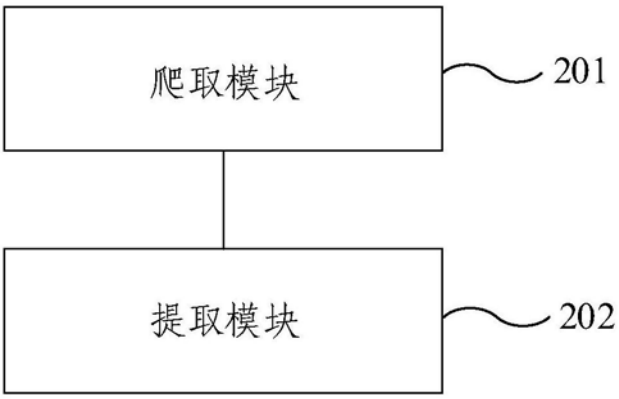

爬取模块
201
提取模块
202

图2

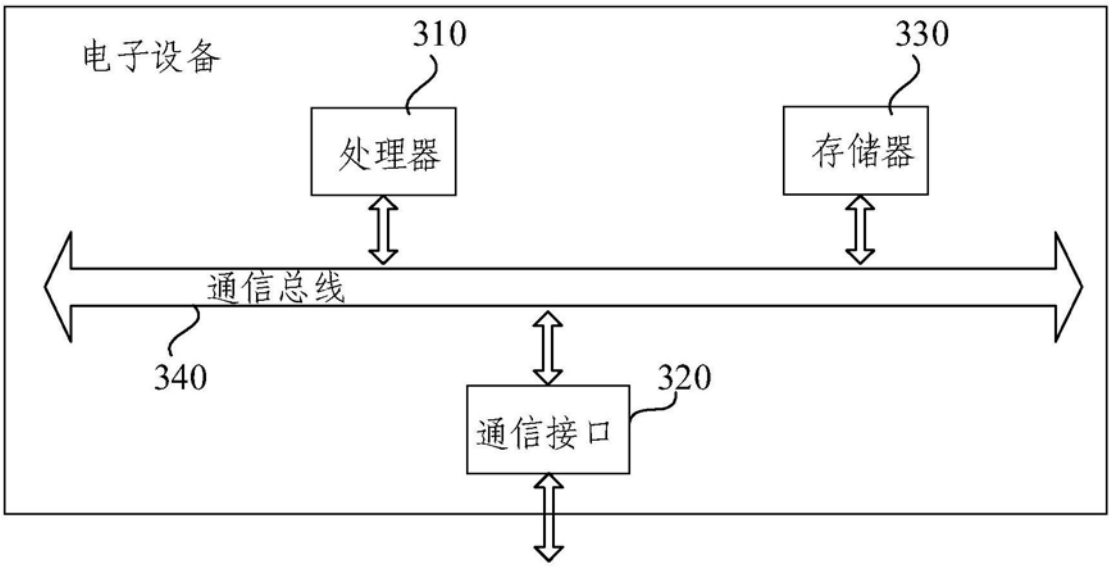


图3