(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
2010 年 9 月 30 日 (30.09.2010)

PCT

(10) 国际公布号
WO 2010/108421 A1

(51) 国际专利分类号:
H04L 12/24 (2006.01) *G06F 17/30* (2006.01)

(21) 国际申请号: PCT/CN2010/071145

(22) 国际申请日: 2010 年 3 月 19 日 (19.03.2010)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
200910129168.1 2009 年 3 月 27 日 (27.03.2009) CN

(71) 申请人 (对除美国外的所有指定国): **腾讯科技（深圳）有限公司 (TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED)** [CN/CN]; 中国广东省深圳市福田区振兴路赛格科技园 2 栋东 4 楼, Guangdong 518044 (CN)。

(72) 发明人; 及

(75) 发明人/申请人 (仅对美国): **龙一民 (LONG, Yimin)** [CN/CN]; 中国广东省深圳市福田区振兴路赛格科技园 2 栋东 4 楼, Guangdong 518044 (CN)。

(74) 代理人: **北京德琦知识产权代理有限公司 (DEQI INTELLECTUAL PROPERTY LAW CORPORATION)**; 中国北京市海淀区知春路 1 号学院国际大厦 7 层, Beijing 100083 (CN)。

(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, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, 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, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, 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, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM, 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(条约第 21 条(3))。

(54) Title: METHOD AND APPARATUS FOR AUTHENTICATING A WEBSITE

(54) 发明名称 : 鉴别网站的方法及装置

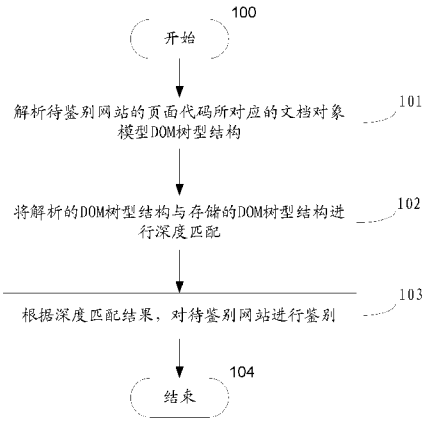

图 1 / Fig. 1

100 START
101 PARSING A DOCUMENT OBJECT MODEL (DOM) TREE STRUCTURE CORRESPONDING TO THE PAGE CODE OF A WEBSITE TO BE AUTHENTICATED
102 DEPTH MATCHING THE PARSED DOM TREE STRUCTURE WITH THE STORED DOM TREE STRUCTURE
103 AUTHENTICATING THE WEBSITE TO BE AUTHENTICATED ACCORDING TO THE DEPTH MATCHING RESULT
104 END

(57) Abstract: The present invention provides a method and apparatus for authenticating a website, and belongs to the communications field. The method includes: parsing a Document Object Model (DOM) tree structure corresponding to the page code of the website to be authenticated; matching the parsed DOM tree structure with the stored DOM tree structure; authenticating the website to be authenticated according to the matching result. The apparatus includes: a parsing module, a matching module and an authenticating module. By parsing a DOM tree structure corresponding to the page code of a website to be authenticated, and then matching the DOM tree structure of the website to be authenticated with the pre-stored DOM tree structure corresponding to the page code of the secure website, the present invention authenticates the website to be authenticated on the basis of the match results, and is able to duly and effectively identify risky websites.

(57) 摘要: 本发明公开了一种鉴别网站的方法及装置，属于通信领域。所述方法包括：解析待鉴别网站的页面代码所对应的 DOM 树型结构；将所述解析出的 DOM 树型结构与存储的 DOM 树型结构进行匹配；根据匹配结果，对所述待鉴别网站进行鉴别。所述装置包括：解析模块、匹配模块和鉴 别模块。本发明通过解析待鉴别网站的页面代码所对应的 DOM 树型结构，并将待鉴别网站的 DOM 树型结构与预先存储的安全网站的页面代码所对应的 DOM 树型结构进行匹配，根据匹配结果，对待鉴别网站进行鉴别，具有及时、有效地鉴别出风险网站的效果。

WO 2010/108421 A1

鉴别网站的方法及装置

技术领域

本发明涉及通信领域，特别涉及一种鉴别网站的方法及装置。

发明背景

随着互联网的快速发展，各种各样的网站丰富和方便了人们的日常生活，然而，相继也出现了一些不安全的网站，例如：钓鱼网站，即欺骗性的网站，该类网站意图引诱用户给出敏感信息，如用户名、口令、帐号、密码或信用卡详细信息等，以骗取用户帐户的财产；还存在一些病毒网站，利用人们在访问网站时传播病毒，因此，鉴别网站是否安全，成为了人们在应用网络时的一种需求。

现有技术鉴别钓鱼网站的一种方式是通过判断URL（Uniform Resource Locator，统一资源定位器）的可疑特征，例如：真实的某银行网站的URL为http://www.wmd.com，虚假的该银行网站的URL可能是http://www. wmd.com.mn，或者是对真实网站的URL进行了编码，例如http://www. wmd.com%2e%61%62%63%2e%63%6f%6d，用户虽然看到的是wmd.com，但浏览器会根据URL编码规则，将带有‘%’的内容解码，真实的目标地址是http://www. wmd.com.abc.com，真正的目标域名是abc.com而不是wmd.com。通过两者对比，即可鉴别钓鱼网站。另一种鉴别钓鱼网站的方式是通过维护黑名单库，该黑名单库中的URL由用户举报，或通过找出大量传播URL的可疑来源，将可疑来源传播的URL列入黑名单库。

在实现本发明的过程中，发明人发现现有技术至少存在以下几个缺点：

判断URL可疑特征的方式需URL含有可疑特征，例如带有编码符号“%”，或者URL的一部分和知名URL重合。假如URL不含可疑的编码特征，或者不含与知名URL重合的部分，则不能通过该方法进行有效鉴别，因此该种方式具有局限性。

对于维护黑名单库的方式，需要由识别出可疑URL的用户进行举报，或者需要对大量传播的URL的可疑来源进行查找，才能将可疑URL列入钓鱼网站鉴别的考察范围。这种方式鉴别钓鱼网站在时效性方面有滞后；而且对于传播量小、危害性高的钓鱼网站，无法及时鉴别。

发明内容

为了及时、有效地对网站进行鉴别，本发明实施例提供了一种鉴别网站的方法及装置。所述技术方案如下:

一方面，提供了一种鉴别网站的方法，所述方法包括:

解析待鉴别网站的页面代码所对应的DOM树型结构;

将所述解析出的DOM树型结构与存储的DOM树型结构进行匹配;

根据匹配结果，对所述待鉴别网站进行鉴别。

另一方面，提供了一种鉴别网站的装置，所述装置包括:

解析模块，用于解析待鉴别网站的页面代码所对应的文档对象模型DOM树型结构;

鉴别模块，用于根据匹配结果，对所述待鉴别网站进行鉴别。

本发明实施例提供的技术方案的有益效果是:

通过解析待鉴别网站的页面代码所对应的DOM树型结构，并将待鉴别网站的DOM树型结构与预先存储的DOM树型结构进行匹配，根据匹配结果，及时、有效地对待鉴别网站进行鉴别。

附图简要说明

为了更清楚地说明本发明实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅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。

图1是本发明实施例1提供的鉴别网站的方法流程图；

图2是本发明实施例2提供的鉴别网站的方法流程图；

图3是本发明实施例2提供的HTML代码与DOM树型结构示例示意图；

图4是本发明实施例3提供的第一种鉴别网站的装置结构示意图；

图5是本发明实施例3提供的第二种鉴别网站的装置结构示意图；

图6是本发明实施例3提供的第三种鉴别网站的装置结构示意图；

图7是本发明实施例3提供的第四种鉴别网站的装置结构示意图。

实施本发明的方式

为使本发明的目的、技术方案和优点更加清楚，下面将结合附图对本发明实施方式作进一步地详细描述。

实施例1

参见图1，本实施例提供了一种鉴别网站的方法，方法流程如下：

101：解析待鉴别网站的页面代码所对应的文档对象模型DOM树型结构；

102：将所述解析出的DOM树型结构与存储的DOM树型结构进行匹配；

103：根据匹配结果，对所述待鉴别网站进行鉴别。

本实施例提供的方法，通过解析待鉴别网站的页面代码所对应的

DOM树型结构，并将待鉴别网站的DOM树型结构与预先存储的DOM树型结构进行匹配，根据匹配结果，及时、有效地鉴别网站。

实施例2

本实施例提供了一种鉴别网站的方法，在一定时期内，网站显示的内容是会变化的，但是网站的布局结构保持一定或者大部分相似。如某网站昨天导航栏的位置今天依然是导航栏；昨天显示天气信息的位置今天仍旧显示天气信息；各个频道栏目所处的位置不变，变化的仅是其中链接的数目和标题等。因钓鱼网站绝大部分都与其对应的真实网站在视觉上非常相似，而网站的视觉和结构布局信息存在于网页的页面代码中，而网页的页面代码可以被解析，形成DOM（Document Object Model，文档对象模型）树，以网页代码中的HTML（Hyper Text Mark-up Language，超文本标记语言）代码解析成DOM树型结构为例，网页浏览器正是通过解析HTML代码，形成DOM树型结构，并基于其中的各个节点和节点的值，向终端用户展示网页信息，以及同用户进行网页操作的交互，同理，采用网页代码中的XML（Extensible Markup Language，可扩展标记语言）代码和XHTML（Extensible Hyper Text Markup Language，可扩展超文本标识语言）代码解析成DOM树型结构，同样也可实现鉴别网站。本实施例提供的方法，将根据网站的视觉相似程度，即两个DOM树型结构的相似程度，实时判断待鉴别网站的结构或者布局同哪个已知的安全网站类似，从而鉴别出待鉴别网站是否安全。

为了便于说明，本实施例以待鉴别网站（用户想要打开的网站）为A，而该待鉴别网站A为某知名网站B的钓鱼网站为例，以HTML语言为例，对本实施例提供的鉴别网站的方法进行说明，参见图2，该方法内容如下：

201：待鉴别网站A被打开之前，鉴别网站的装置根据待鉴别网站A

的网址去相应的服务器端拉取该待鉴别网站A对应的HTML代码;

其中，此处提到的网址为IP（Internet Protocol，网际协议）地址或域名地址，采用域名地址去相应服务器端拉取该待鉴别网站对应的页面代码时，需要通过域名服务器（DNS）将域名地址翻译成IP地址，然后通过IP地址才能找到服务器；而采用IP地址去相应服务器端拉取该待鉴别网站对应的页面代码时，则省去了域名服务器解析域名的过程，因此，在实际应用中，此处的网址采用IP地址则更为便捷，本实施例不对采用哪种地址做具体限定。

202：解析拉取的HTML代码所对应的DOM树型结构;

具体地，图3提供了将HTML代码解析成DOM树型结构的一个实例，如图3所示，左边是网站页面的HTML代码，右边是解析后形成的DOM树型结构。根据HTML代码的名称规范，图3中左边括号中的内容称为标签，标签一般以<标签名>开始，并已</标签名>结尾。一个标签结构中可以嵌套多个标签。标签的类型可以决定网站的网页布局，例如，图3中的TABLE（表格）标签在网页中呈现为表格，实际应用中还存在其它多种类型的标签，例如，图3未提及到的INPUT标签在网页中呈现为用户名或者密码的输入框。

HTML代码转换为DOM树结构时，节点的结构和顺序表示网站的视觉布局，节点的值表示用户所见的内容。如图3右边所示，TABLE为根节点，TABLE的子节点是TBODY（表格主体），TBODY有两个子节点，都是TR（表格中的行），表示表格中的一行。TR有两个子节点，都是TD（表格中的列），表示该行中的一列。TD节点的值，分别为文本天气，时间，城市和区域，这些文本便是用户在页面上看到的表格中的文本内容。

203：将解析的DOM树型结构与存储的DOM树型结构进行匹配;

其中，本实施例以存储的DOM树型结构为安全网站的页面代码所对应的DOM树型结构为例，相对于不安全的钓鱼网站，即风险网站而言，安全网站包括但不限于需要通过用户名、密码或其它帐户信息登录，且存在利益驱动、存在被模仿和伪造风险的网站，例如：银行网站等。本实施例提供的鉴别网站装置中，不仅预先存储了与钓鱼网站相对应的真实网站的HTML代码、HTML代码所对应的DOM树型结构，还存储了DOM树型结构中所有节点的个数。为了保证存储的安全网站的DOM树型结构更全面，更具有参照性，每隔一定周期（例如：每隔5秒钟），存储的安全网站的DOM树型结构将被刷新。

具体地，DOM树型结构分为深度方向（纵向）和广度方向（横向），将解析出的DOM树型结构与存储的DOM树型结构进行匹配时，本实施例提供了两种匹配方式，这两种匹配方式分别从DOM树型结构中的节点被赋权值及不被赋权值的情况下得出的，一方面，先考虑DOM树型结构中的节点不被赋权值的情况：

首先，在存储的DOM树型结构中，判断是否存在与解析出的DOM树型结构具有相同节点的DOM树型结构，如果是，则对该存储的DOM树型结构和解析出的DOM树型结构进行深度遍历，确定两个DOM树型结构中有多少个相同节点，得到所有相同节点的个数；设该待鉴别网站A的DOM树形结构为a，而鉴别钓鱼网站装置中存储的多个DOM树型结构中包括知名网站B的DOM树型结构b，则说明该待鉴别网站很有可能是存储的非钓鱼网站的仿造（即钓鱼网站）；当相同节点的个数m与知名网站B所对应的DOM树型结构b的总节点数n的比值满足一定阈值时，则匹配成功。其中，阈值能够代表两个网站的相似程度，如果不是同一个网站，阈值越高，则说明待鉴别网站为仿造网站的可能性越大，即，该待鉴别网站为风险网站的可能性越大，本实施例不对阈值做具体限

定，例如：设当m/n>=60%时，则判断待鉴别网站与存储的非钓鱼网站视觉相似，即匹配成功。

另一方面，在网页代码中，标签的类型可以决定网页的布局，将网页代码解析成DOM树型结构后，对于DOM树型结构中的所有节点，节点的类型及排列顺序可表示网站的视觉布局，因此，节点的类型及排列顺序均能够影响网站的视觉布局，不同类型的节点，对视觉布局的影响程度也并不相同，例如：钓鱼网站存在帐户、密码输入框、表格等主要特征的关键节点。考虑到节点类型对视觉布局的影响程度，本实施例采取了对DOM树型结构中的节点赋权值的匹配的形式：

在对DOM树型结构中的节点赋权值时，根据节点的不同类型，鉴别装置预先对DOM树型结构中的不同节点赋予了不同的权值。权值越高，则说明该节点的作用越关键，对视觉布局的影响程度越高，或对于非法获取用户帐户和密码的特征越强。例如，由高到低列举各不同节点的权值α：html代码中表单提交形式节点input的权值是2.0，即[input：2.0]，html代码中根节点的权值时1.8，即[table：1.8]，html代码中TABLE的子节点的权值是1.8，即[tbody: 1.8]，html代码中正文开始的第一个标签body的权值是1.8，即[body：1.8]，html代码中图片标签节点img的权值是1.5，即[img：1.5]，html代码中目标类型节点object的权值是1.5，即[object: 1.5]，html代码中表格中的行节点tr的权值是1.5，即[tr：1.5]，html代码中表格中的列节点td的权值是1.4，即[td：1.4]，html代码中块级元素节点div的权值是1.0，即[div: 1.0]，html代码中文本段落节点p的权值是0.8，即[p：0.8]，html代码中换行符节点br的权值是0.7，即[br：0.7]，html代码中行内元素节点span的权值是0.6，即[span：0.6]。仍以判断出存储的DOM树型结构中存在与解析出的DOM树型结构具有相同节点的DOM树型结构A为例，对该存储的DOM树型结构A和解析出的DOM树型结构

进行深度遍历，确定两个DOM树型结构中有多少个相同节点，得到所有相同节点的加权值x；则对所述具有相同节点的DOM树型结构进行深度遍历，得到所有相同节点的加权值；当所有相同节点的加权值x与DOM树型结构A的总节点数n的比值满足一定阈值，则匹配成功，例如：设x/n>=60%时，认为待鉴别网站与存储的非钓鱼网站视觉相似，即匹配成功。本实施例不对节点被赋予的权值做具体限定。

204：匹配成功，对匹配成功的两个DOM树型结构，比较两个DOM树型结构所对应的网址是否一致，如果是，则执行205，否则，执行206；

205：鉴别出该待鉴别网站并非钓鱼网站，流程结束；

206：鉴别出该待鉴别网站为钓鱼网站，阻止用户查看被鉴别为钓鱼网站的页面内容。

可选地，本实施例仅以预先存储的DOM树型结构为安全网站对应的DOM树型结构为例，对本实施例提供的方法做了说明，同理，本实施例也可以预先存储的DOM树型结构为风险网站所对应的DOM树型结构为例。

针对该种情况，当待鉴别网站所对应的DOM树型结构与存储的DOM树型结构匹配成功后，如果两个DOM树型结构对应的网址一致，则鉴别待鉴别网站为风险网站，如果两个DOM树型结构不一致，则鉴别该待鉴别网站为安全网站。在实际应用中，即使每隔一定周期，对预先存储的DOM树型结构进行更新，也未必能够覆盖所有风险网站的DOM树型结构，因此，即使待鉴别网站的网址与存储的风险网站的网址不一致，则该待鉴别网站也未必为安全网站。因此，采用该种方式鉴别网站的可靠性及安全性不高。

对于预先存储的DOM树型结构为安全网站对应的DOM树型结构的情况，虽然在鉴别结果上同样也会存在误差，但采用该种方式鉴别网站

时，不会将风险网站鉴别为安全网站，因此，可靠性及安全性相对也就更高。

综上所述，本实施例提供的方法，通过拉取待鉴别网站的HTML代码，解析待鉴别网站的HTML代码所对应的DOM树型结构，并将待鉴别网站的DOM树型结构与预先存储的非钓鱼网站的HTML代码所对应的DOM树型结构进行匹配，根据匹配结果，及待鉴别网站与非钓鱼网站的网址，判断待鉴别网站是否为钓鱼网站，从而能够实现及时、有效地鉴别钓鱼网站，并在鉴别出钓鱼网站时，阻止用户查看钓鱼网站内容，防止用户受骗，保护了用户的利益。

实施例3

参见图4，本实施例提供了一种鉴别网站的装置，该装置包括:

解析模块401，用于解析待鉴别网站的页面代码所对应的文档对象模型DOM树型结构;

匹配模块402，用于将解析出的DOM树型结构与存储的DOM树型结构进行匹配;

鉴别模块403，用于根据匹配结果，对待鉴别网站进行鉴别。

具体地，鉴别模块403，具体用于匹配的结果为成功，且待鉴别网站的网址与匹配成功的DOM树型结构所对应的网址不一致时，鉴别待鉴别网站为风险网站。

参见图5，对于DOM树型结构中的节点不被赋予权值的情况，上述匹配模块402，具体包括:

第一判断单元402a，用于在存储的DOM树型结构中，判断是否存在与解析出的DOM树型结构具有相同节点的DOM树型结构；如果是，则将与解析出的DOM树型结构具有相同节点的DOM树型结构作为符合匹配条件的DOM树型结构;

第一遍历单元402b，用于对符合匹配条件的DOM树型结构与解析出的DOM树型结构进行深度遍历，得到相同节点的个数；

第一匹配判断单元402c，用于判断相同节点的个数与符合匹配条件的DOM树型结构中的所有节点个数的比值是否满足一定阈值，如果是，则判断匹配成功。

参见图6，对于DOM树型结构中的节点被赋予权值的情况，上述匹配模块402，具体包括：

赋权单元402d，用于对存储的DOM树型结构中的每个节点赋权值；

第二判断单元402e，用于在存储的DOM树型结构中，判断是否存在与解析出的DOM树型结构具有相同节点的DOM树型结构；如果是，则将与解析出的DOM树型结构具有相同节点的DOM树型结构作为符合匹配条件的DOM树型结构；

第二遍历单元402f，用于对符合匹配条件的DOM树型结构与解析出的DOM树型结构进行深度遍历，得到所有相同节点的加权值；

第二匹配判断单元402g，用于判断所有相同节点的加权值与符合匹配条件的DOM树型结构中的所有节点个数的比值是否满足一定阈值，如果是，则判断匹配成功。

进一步地，参见图7，该鉴别网站装置还包括：

拉取模块404，用于根据待鉴别网站的网址拉取待鉴别网站的页面代码。

综上，本实施例提供的鉴别网站的装置，通过解析待鉴别网站的页面代码所对应的DOM树型结构，并将待鉴别网站的DOM树型结构与预先存储的安全网站的页面代码所对应的DOM树型结构进行匹配，根据匹配结果，及待鉴别网站与安全网站的网址，判断待鉴别网站是否为风险网站，从而能够实现及时、有效地鉴别网站，防止用户受骗，保护了用

户的利益。

上述本发明实施例序号仅仅为了描述，不代表实施例的优劣。

本发明实施例中的部分步骤，可以利用软件实现，相应的软件程序可以存储在可读取的存储介质中，如光盘或硬盘等。

以上所述仅为本发明的较佳实施例，并不用以限制本发明，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。

权利要求书

1、一种鉴别网站的方法，其特征在于，所述方法包括:

解析待鉴别网站的页面代码所对应的文档对象模型DOM树型结构;

将所述解析出的DOM树型结构与存储的DOM树型结构进行匹配;

根据匹配结果，对所述待鉴别网站进行鉴别。

2、根据权利要求1所述的方法，其特征在于，所述根据匹配结果，对网站进行鉴别，具体包括:

如果匹配成功，且所述待鉴别网站的网址与匹配成功的DOM树型结构所对应的网址不一致，则鉴别所述待鉴别网站为风险网站。

3、根据权利要求1所述的方法，其特征在于，所述将所述解析出的DOM树型结构与存储的DOM树型结构进行匹配，具体包括:

在存储的DOM树型结构中，判断是否存在与所述解析出的DOM树型结构具有相同节点的DOM树型结构，如果是，则将与所述解析出的DOM树型结构具有相同节点的DOM树型结构作为符合匹配条件的DOM树型结构，并对所述符合匹配条件的DOM树型结构与所述解析出的DOM树型结构进行深度遍历，得到相同节点的个数;

如果所述相同节点的个数与所述符合匹配条件的DOM树型结构的所有节点个数的比值满足一定阈值，则匹配成功。

4、根据权利要求1所述的方法，其特征在于，所述将所述解析出的DOM树型结构与存储的DOM树型结构进行匹配，具体包括:

对存储的DOM树型结构中的每个节点赋权值，在存储的DOM树型结构中，判断是否存在与所述解析出的DOM树型结构具有相同节点的DOM树型结构，如果是，则将所述与所述解析出的DOM树型结构具有相同节点的DOM树型结构作为符合匹配条件的DOM树型结构，并对所述符合匹配条件的DOM树型结构与所述解析出的DOM树型结构进行深

度遍历，得到所有相同节点的加权值;

如果所述所有相同节点的加权值与所述符合匹配条件的DOM树型结构中的所有节点个数的比值满足一定阈值，则匹配成功。

5、根据权利要求1至4任一项所述的方法，其特征在于，所述存储的DOM树型结构，具体为:

预先存储的安全网站的页面代码所对应的DOM树型结构，且每隔一定周期将被刷新。

6、根据权利要求1所述的方法，其特征在于，所述存储的DOM树型结构为:

预先存储的风险网站的页面代码所对应的DOM树型结构。

7、一种鉴别网站的装置，其特征在于，所述装置包括:

解析模块，用于解析待鉴别网站的页面代码所对应的文档对象模型DOM树型结构;

匹配模块，用于将所述解析出的DOM树型结构与存储的DOM树型结构进行匹配;

鉴别模块，用于根据匹配结果，对所述待鉴别网站进行鉴别。

8、根据权利要求7所述的装置，其特征在于，所述鉴别模块，具体用于匹配的结果为成功，且所述待鉴别网站的网址与匹配成功的DOM树型结构所对应的网址不一致时，鉴别所述待鉴别网站为风险网站。

9、根据权利要求7所述的装置，其特征在于，所述匹配模块，具体包括:

第一判断单元，用于在存储的DOM树型结构中，判断是否存在与所述解析出的DOM树型结构具有相同节点的DOM树型结构；如果是，则将与所述解析出的DOM树型结构具有相同节点的DOM树型结构作为符合匹配条件的DOM树型结构;

第一遍历单元，用于对所述符合匹配条件的DOM树型结构与所述解析出的DOM树型结构进行深度遍历，得到相同节点的个数；

第一匹配判断单元，用于判断所述相同节点的个数与所述符合匹配条件的DOM树型结构中的所有节点个数的比值是否满足一定阈值，如果是，则判断匹配成功。

10、根据权利要求7所述的装置，其特征在于，所述匹配模块，具体包括：

赋权单元，用于对存储的DOM树型结构中的每个节点赋权值；

第二判断单元，用于在存储的DOM树型结构中，判断是否存在与所述解析出的DOM树型结构具有相同节点的DOM树型结构；如果是，则将与所述解析出的DOM树型结构具有相同节点的DOM树型结构作为符合匹配条件的DOM树型结构；

第二遍历单元，用于对所述符合匹配条件的DOM树型结构与所述解析出的DOM树型结构进行深度遍历，得到所有相同节点的加权值；

第二匹配判断单元，用于判断所述所有相同节点的加权值与所述符合匹配条件的DOM树型结构中的所有节点个数的比值是否满足一定阈值，如果是，则判断匹配成功。

11、根据权利要求7所述的装置，其特征在于，所述装置，还包括：

拉取模块，用于根据所述待鉴别网站的网址拉取所述待鉴别网站的页面代码。

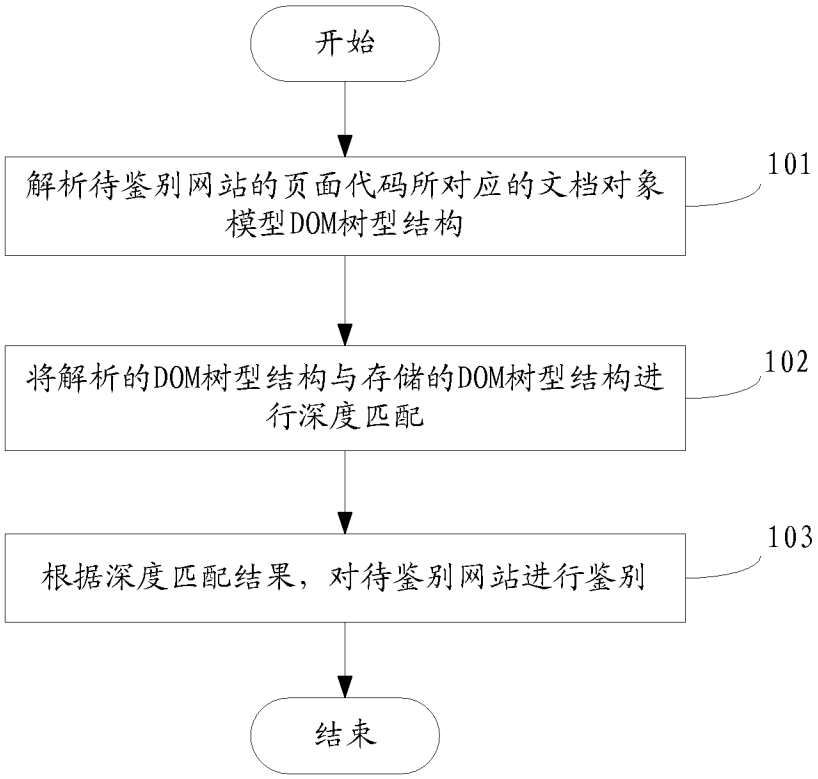
开始
解析待鉴别网站的页面代码所对应的文档对象模型DOM树型结构
101
将解析的DOM树型结构与存储的DOM树型结构进行深度匹配
102
根据深度匹配结果，对待鉴别网站进行鉴别
103
结束

图 1

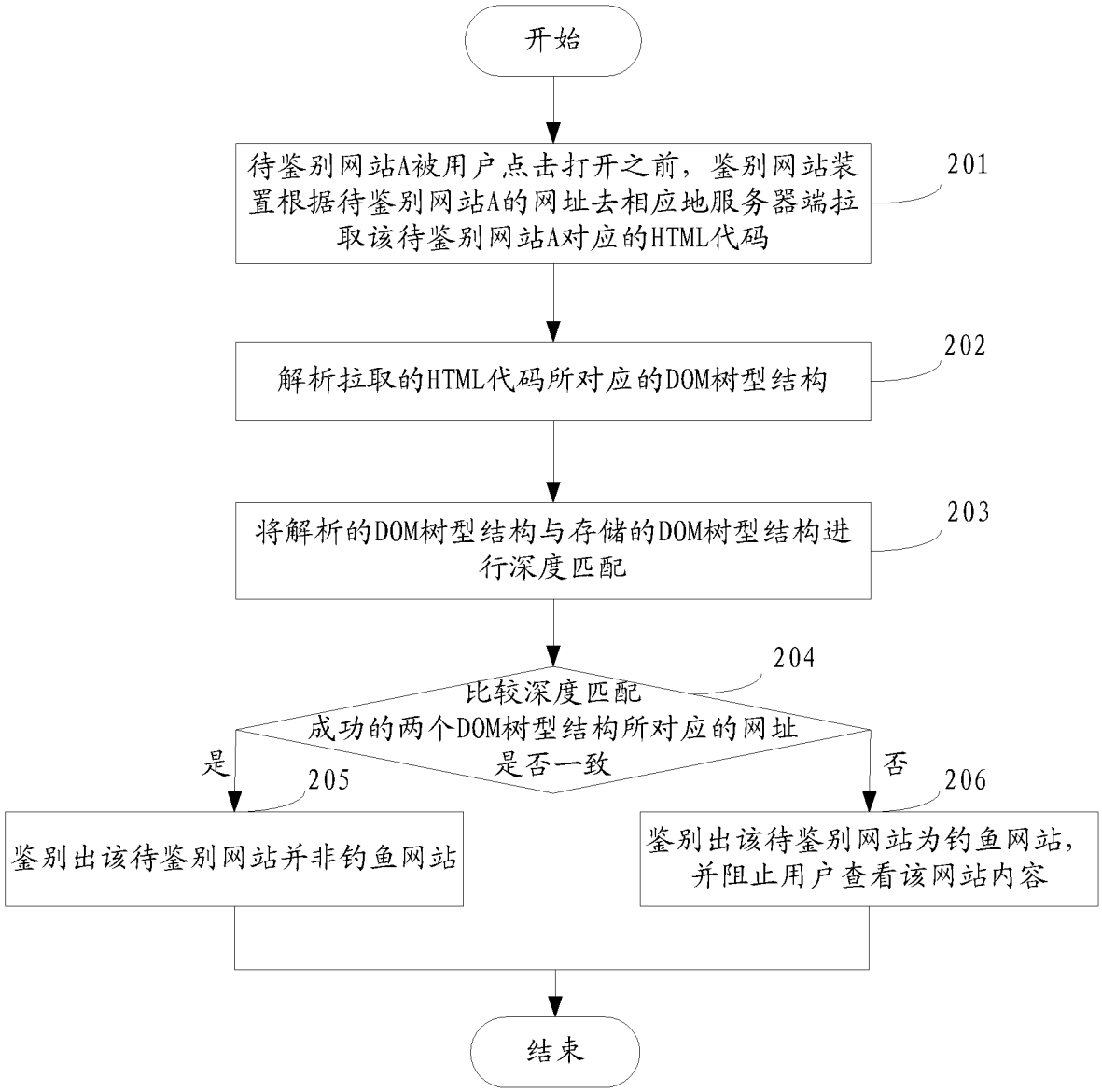
开始
待鉴别网站A被用户点击打开之前，鉴别网站装置根据待鉴别网站A的网址去相应地服务器端拉取该待鉴别网站A对应的HTML代码
201
解析拉取的HTML代码所对应的DOM树型结构
202
将解析的DOM树型结构与存储的DOM树型结构进行深度匹配
203
比较深度匹配成功的两个DOM树型结构所对应的网址是否一致
204
是
否
205
鉴别出该待鉴别网站并非钓鱼网站
206
鉴别出该待鉴别网站为钓鱼网站，并阻止用户查看该网站内容
结束

图 2

HTML代码

```
<TABLE>
  <TBODY>
    <TR>
      <TD>天气</TD>
      <TD>时间</TD>
    </TR>
      <TD>城市</TD>
      <TD>区域</TD>
    </TR>
  </TBODY>
</TABLE>
```

DOM树型结构

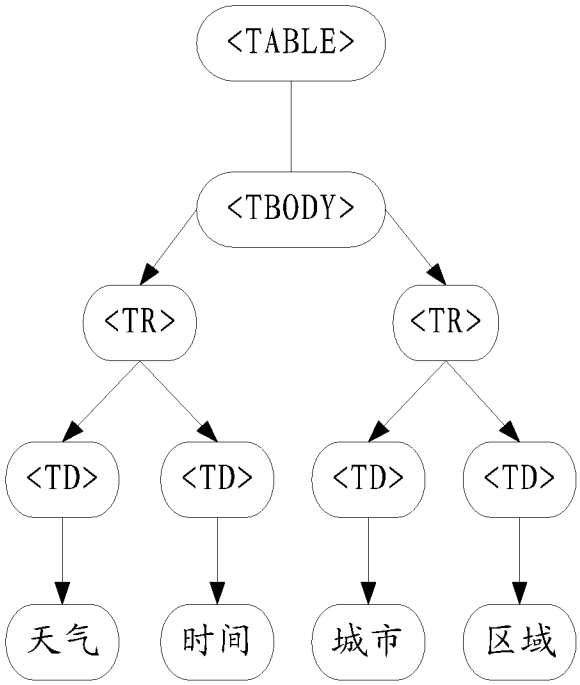

图 3

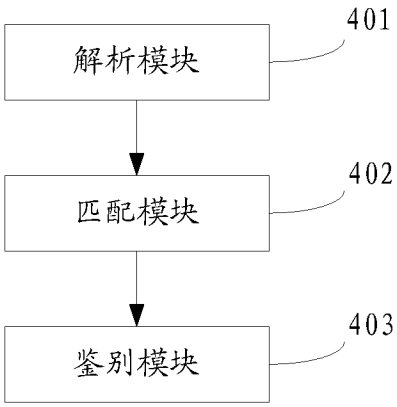

图 4

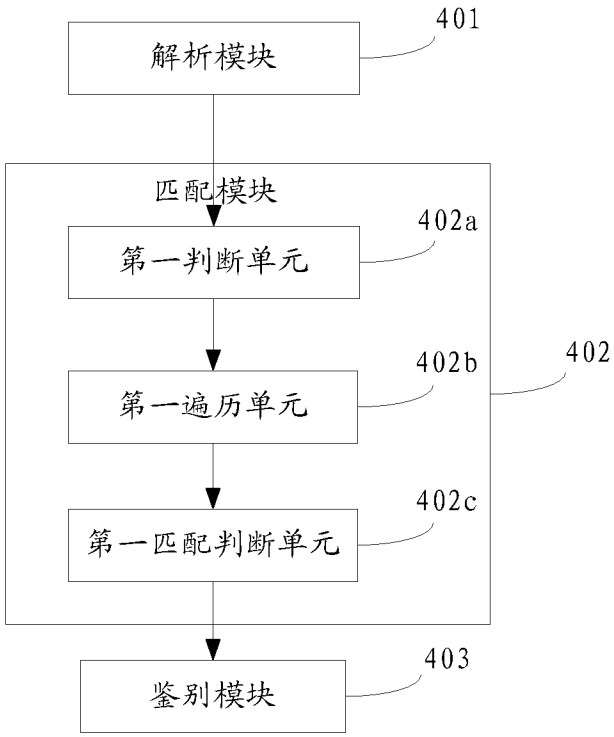
401
解析模块
匹配模块
402a
第一判断单元
402b
第一遍历单元
402
402c
第一匹配判断单元
403
鉴别模块

图 5

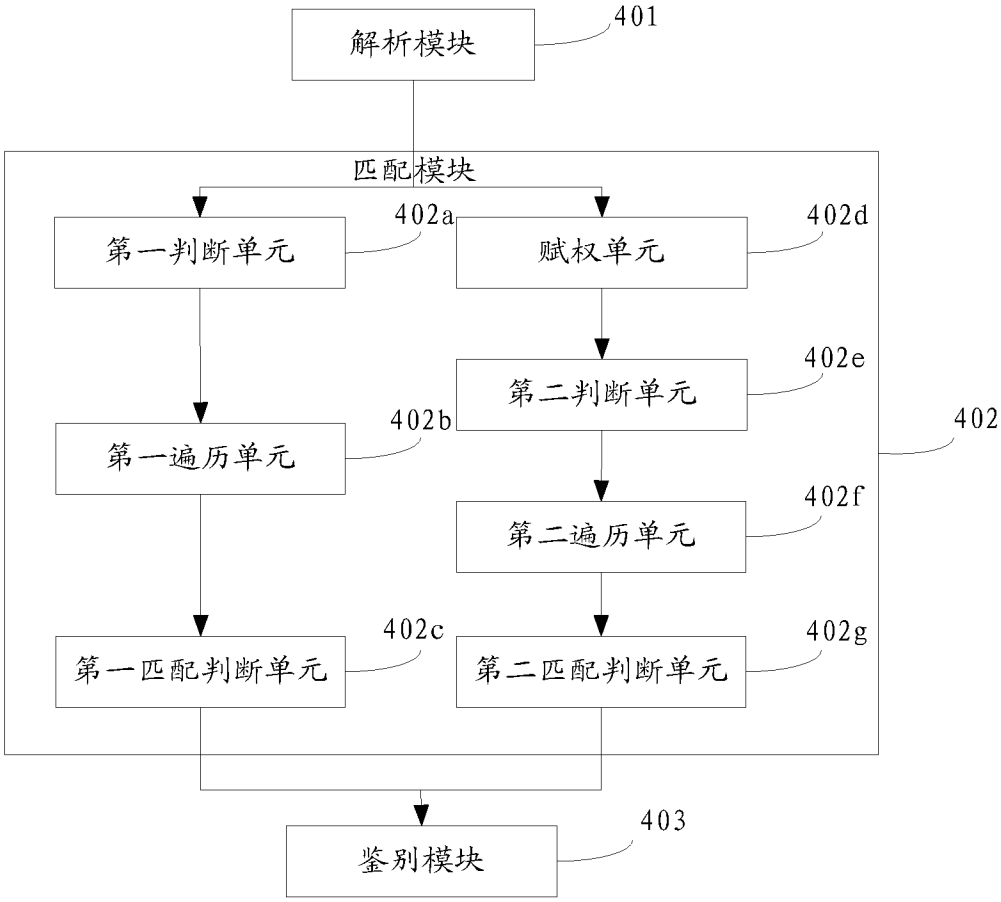
解析模块
401
匹配模块
第一判断单元
402a
赋权单元
402d
第二判断单元
402e
第一遍历单元
402b
402
第二遍历单元
402f
第一匹配判断单元
402c
第二匹配判断单元
402g
鉴别模块
403

图 6

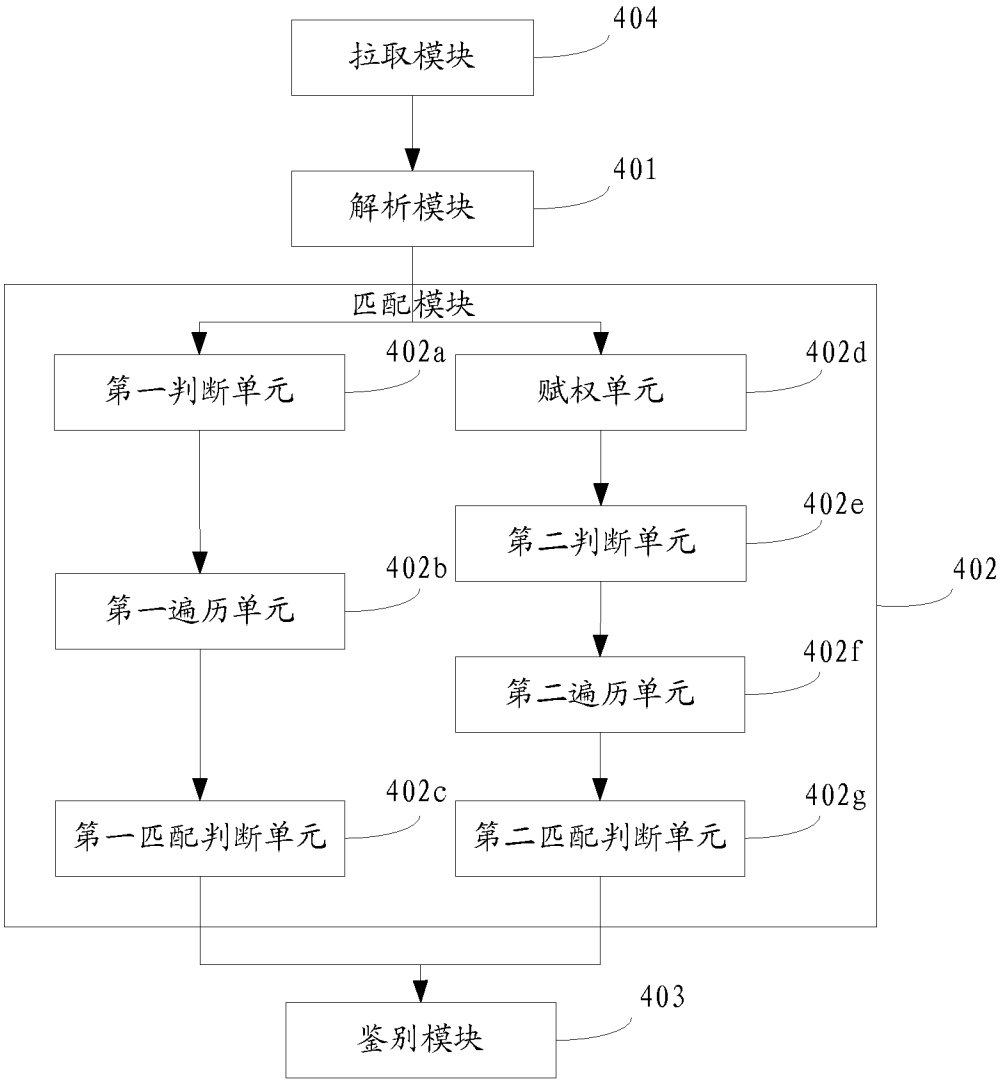
拉取模块
404
解析模块
401
匹配模块
第一判断单元
402a
赋权单元
402d
第二判断单元
402e
第一遍历单元
402b
402
第二遍历单元
402f
第一匹配判断单元
402c
第二匹配判断单元
402g
鉴别模块
403

图 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2010/071145

A. CLASSIFICATION OF SUBJECT MATTER

See Extra Sheet

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)

IPC: H04L, G06F

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

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

CPRSABS, CNTXT, CNKI, WPI, EPODOC: authenticat+, recogni+, verify, identify+, distinguish, differentiat+, discriminate+, validat+, phish, fish, risk, virus, secure+, safe, deceiv+, cheat, illegal, hacker, DOM, document w object w model, match, compar+, travers+, site, page, URL

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN101510887A（TENCENT TECHNOLOGY SHENZHEN CO LTD）19 Aug. 2009（19.08.2009）claims 1-10, the description pages 3-7	1-11
PX	CN101534306A（SHENZHEN TENCENT TECHNOLOGIES CO LTD）16 Sept. 2009（16.09.2009）abstract, the description page 3	1, 7
PA		2-6, 8-11
A	CN101147138A（DUAXES CORP）19 Mar. 2008（19.03.2008）the whole document	1-11
A	US2007174486A1（ORACLE INT CORP ET-AL）26 Jul. 2007（26.07.2007）the whole document	1-11
A	CN101145902A（UNIV SOUTHEAST）19 Mar. 2008（19.03.2008）the whole document	1-11

☐ 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
21 May 2010（21.05.2010）	**03 Jun. 2010 (03.06.2010)**
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 WANG Hongli Telephone No. (86-10)62411281

Form PCT/ISA /210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2010/071145

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN101510887A	19.08.2009	None	
CN101534306A	16.09.2009	None	
CN101147138A	19.03.2008	WO2006087837A1	24.08.2006
		EP1850234A1	31.10.2007
		KR20070103774A	24.10.2007
		JP2007503571T2	03.07.2008
		US2009178116A1	09.07.2009
US2007174486A1	26.07.2007	US7543024B2	02.06.2009
CN101145902A	19.03.2008	None	

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2010/071145

Continuation of : **A. CLASSIFICATION OF SUBJECT MATTER**

H04L 12/24（2006.01）i
G06F 17/30（2006.01）n

国际检索报告

国际申请号
PCT/CN2010/071145

A. 主题的分类

参见附加页

按照国际专利分类(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: H04L, 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）)

CPRSABS, CNTXT, CNKI: 鉴别, 验证, 认证, 识别, 钓鱼, 欺骗, 非法, 病毒, 安全, 恶意, 风险, 攻击, 网站, 网页, 网址, 代码, 文档对象模型, DOM, 匹配, 比较, 遍历

WPI, EPODOC: authenticat+, recogni+, verify, identify+, distinguish, differentiat+, discriminate+, validat+, phish, fish, risk, virus, secure+, safe, deceiv+, cheat, illegal, hacker, DOM, document w object w model, match, compar+, travers+, site, page, URL

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
P X	CN101510887A（腾讯科技（深圳）有限公司）19.8 月 2009（19.08.2009）权利要求 1-10，说明书 3-7 页	1-11
P X	CN101534306A（深圳市腾讯计算机系统有限公司）16.9 月 2009	1, 7
P A	（16.09.2009）摘要，说明书第 3 页	2-6, 8-11
A	CN101147138A（DUAXES 株式会社）19.3 月 2008（19.03.2008）全文	1-11
A	US2007174486A1（ORACLE INT CORP 等）26.7 月 2007（26.07.2007）全文	1-11
A	CN101145902A（东南大学）19.3 月 2008（19.03.2008）全文	1-11

☐ 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（如具体说明的）

"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
21.5 月 2010（21.05.2010）	**03.6 月 2010 (03.06.2010)**
ISA/CN 的名称和邮寄地址： 中华人民共和国国家知识产权局 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 王红丽 电话号码：(86-10) **62411281**

国际检索报告

关于同族专利的信息

国际申请号
PCT/CN2010/071145

检索报告中引用的专利文件	公布日期	同族专利	公布日期
CN101510887A	19.08.2009	无	
CN101534306A	16.09.2009	无	
CN101147138A	19.03.2008	WO2006087837A1	24.08.2006
		EP1850234A1	31.10.2007
		KR20070103774A	24.10.2007
		JP2007503571T2	03.07.2008
		US2009178116A1	09.07.2009
US2007174486A1	26.07.2007	US7543024B2	02.06.2009
CN101145902A	19.03.2008	无	

续：**A.** 主题的分类

H04L 12/24（2006.01）i

G06F 17/30（2006.01）n