

(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
2020 年 8 月 20 日 (20.08.2020)



(10) 国际公布号
WO 2020/164276 A1

(51) 国际专利分类号:
G06F 16/951 (2019.01)

(21) 国际申请号: PCT/CN2019/118144

(22) 国际申请日: 2019 年 11 月 13 日 (13.11.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910113261.7 2019 年 2 月 13 日 (13.02.2019) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO.,LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(72) 发明人: 吴启(WU, Qi); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。 王雪春(WANG, Xuechun); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市世纪恒程知识产权代理事务所(CENFO INTELLECTUAL PROPERTY AGENCY); 中国广东省深圳市南山区粤海街道高新技术产业园北区松坪山路 3 号奥特讯电力大厦 201, Guangdong 518057 (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, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS,

(54) Title: WEBPAGE DATA CRAWLING METHOD, APPARATUS AND SYSTEM, AND COMPUTER-READABLE STORAGE MEDIUM

(54) 发明名称: 网页数据爬取方法、装置、系统及计算机可读存储介质

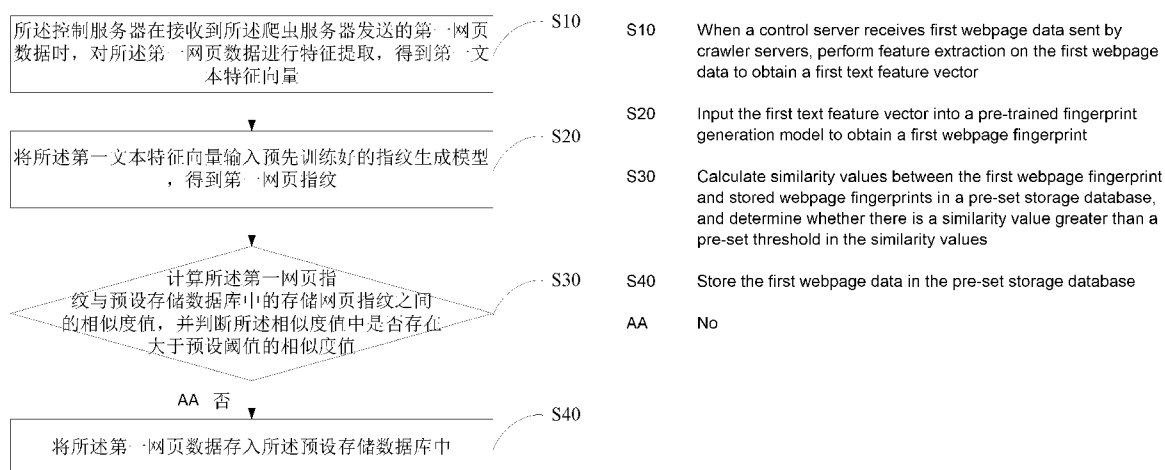


图 2

(57) Abstract: A webpage data crawling method, relating to the field of data crawling. The method is applied to a webpage data crawling system. The system comprises a control server and a plurality of crawler servers connected to the control server. The method comprises: when the control server receives first webpage data sent by the crawler servers, performing feature extraction on the first webpage data to obtain a first text feature vector (S10); inputting the first text feature vector into a pre-trained fingerprint generation model to obtain a first webpage fingerprint (S20); calculating similarity values between the first webpage fingerprint and stored webpage fingerprints in a pre-set storage database, and determining whether there is a similarity value greater than a pre-set threshold in the similarity values (S30); and if not, storing the first webpage data in the pre-set storage database (S40). Further provided are a webpage data crawling apparatus and system, and a computer-readable storage medium. The present application can reduce the repetition rate of webpage data obtained based on crawling of a web crawler.

JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, 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, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, 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, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, 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))。

(57) 摘要: 一种网页数据爬取方法, 涉及数据爬取领域, 该方法应用于网页数据爬取系统, 该系统包括控制服务器、及与控制服务器连接的多个爬虫服务器, 该方法包括: 所述控制服务器在接收到所述爬虫服务器发送的第一网页数据时, 对所述第一网页数据进行特征提取, 得到第一文本特征向量(S10); 将所述第一文本特征向量输入预先训练好的指纹生成模型, 得到第一网页指纹(S20); 计算所述第一网页指纹与预设存储数据库中的存储网页指纹之间的相似度值, 并判断所述相似度值中是否存在大于预设阈值的相似度值(S30); 若不存在, 将所述第一网页数据存入所述预设存储数据库中(S40)。还提供一种网页数据爬取装置、系统及计算机可读存储介质, 能降低基于网络爬虫爬取得到的网页数据的重复率。

网页数据爬取方法、装置、系统及计算机可读存储介质

本申请要求于2019年2月13日提交中国专利局、申请号为201910113261.7、发明名称为“网页数据爬取方法、装置、系统及计算机可读存储介质”的中国专利申请的优先权，其全部内容通过引用结合在申请中。

技术领域

本申请涉及数据爬取技术领域，尤其涉及一种网页数据爬取方法、装置、系统及计算机可读存储介质。

背景技术

随着网络技术的大力发展，通过互联网获取数据已成为当前人们获取信息资源的重要途径，而网络爬虫已成为主流的网页数据的获取手段。然而随着互联网信息的肆意转载和多网站投放，网络爬虫爬取出的数据大多掺杂很多冗余及重复的数据，给后续的数据分析造成一定的影响。因此，亟需一种可去除网络爬虫爬取得到的网页数据中的重复数据的方法，以降低爬取得到的网页数据的重复率。

发明内容

本申请的主要目的在于提供一种网页数据爬取方法、装置、系统及计算机可读存储介质，旨在降低基于网络爬虫爬取得到的网页数据的重复率。

为实现上述目的，本申请提供一种网页数据爬取方法，所述网页数据爬取方法应用于网页数据爬取系统，所述网页数据爬取系统包括控制服务器、及与所述控制服务器连接的多个爬虫服务器，所述网页数据爬取方法包括：

所述控制服务器在接收到所述爬虫服务器发送的第一网页数据时，对所述第一网页数据进行特征提取，得到第一文本特征向量；

将所述第一文本特征向量输入预先训练好的指纹生成模型，得到第一网页指纹；

计算所述第一网页指纹与预设存储数据库中的存储网页指纹之间的相似度值，并判断所述相似度值中是否存在大于预设阈值的相似度值；

若所述相似度值中不存在大于预设阈值的相似度值，则将所述第一网页数据存入所述预设存储数据库中。

此外，为实现上述目的，本申请还提供一种网页数据爬取装置，所述网页数据爬取装置包括：

第一特征提取模块，用于所述控制服务器在接收到所述爬虫服务器发送的第一网页数据时，对所述第一网页数据进行特征提取，得到第一文本特征向量；

第一指纹生成模块，用于将所述第一文本特征向量输入预先训练好的指纹生成模型，得到第一网页指纹；

相似度值判断模块，用于计算所述第一网页指纹与预设存储数据库中的存储网页指纹之间的相似度值，并判断所述相似度值中是否存在大于预设阈值的相似度值；

网页数据存储模块，用于若所述相似度值中不存在大于预设阈值的相似度值，则将所述第一网页数据存入所述预设存储数据库中。

此外，为实现上述目的，本申请还提供一种网页数据爬取系统，所述网页数据爬取系统包括控制服务器、及与所述控制服务器连接的多个爬虫服务器，还存储器、处理器以及存储在所述存储器上并可被所述处理器执行的计算机可读指令，其中所述计算机可读指令被所述处理器执行时，实现如上所述的网页数据爬取方法的步骤。

此外，为实现上述目的，本申请还提供一种计算机可读存储介质，所述计算机可读存储介质上存储有计算机可读指令，其中所述计算机可读指令被处理器执行时，实现如上所述的网页数据爬取方法的步骤。

本申请提供一种网页数据爬取方法、装置、系统及计算机可读存储介质，该方法应用于基于分布式爬虫技术构建的网页数据爬取系统，该网页数据爬取系统包括控制服务器、及与该控制服务器连接的多个爬虫服务器，该方法包括：控制服务器在接收到爬虫服务器发送的第一网页数据时，对该第一网页数据进行特征提取，得到第一文本特征向量；将第一文本特征向量输入预先训练好的指纹生成模型，得到第一网页指纹；计算第一网页指纹与预设存储数据库中的存储网页指纹之间的相似度值，并判断计算得到的相似度值中是否存在大于预设阈值的相似度值；若不存在，才将该第一网页数据存入该预设存储数据库中。本申请通过构建指纹生成模型，对爬取到的网页数据生成对应指纹，然后与存储数据库中已爬取保存的网页数据的指纹进行对比，以判断是否重复，进而对不重复的数据才进行保存，从而可避免重复的网页

数据入库，因此，本申请可解决现有技术中基于网络爬虫爬取得到的数据重复性较高的问题。同时，通过与分布式爬虫技术相结合，可实现网页数据的快速抓取，保证在较短的时间内完成数据的采集，提高了数据采集的效率。

附图说明

图1为本申请实施例方案涉及的硬件运行环境的终端结构示意图；

图2为本申请网页数据爬取方法第一实施例的流程示意图；

图3为本申请网页数据爬取方法第二实施例的流程示意图；

图4为本申请网页数据爬取装置第一实施例的功能模块示意图。

本申请目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

应当理解，此处所描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

请参照图1，图1为本申请实施例方案涉及的硬件运行环境的终端结构示意图。

本申请实施例终端为控制服务器，该控制服务器可以是PC（personal computer，个人计算机）、笔记本电脑、服务器等终端设备。

如图1所示，该终端可以包括：处理器1001，例如CPU（Central Processing Unit，中央处理器），通信总线1002，用户接口1003，网络接口1004，存储器1005。其中，通信总线1002用于实现这些组件之间的连接通信；用户接口1003可以包括显示屏（Display）、输入单元比如键盘（Keyboard），可选用户接口1003还可以包括标准的有线接口、无线接口。网络接口1004可选的可以包括标准的有线接口、无线接口（如无线保真 Wireless-Fidelity，Wi-Fi 接口）；存储器1005可以是高速随机存取存储器（random access memory，RAM），也可以是稳定的存储器（non-volatile memory），例如磁盘存储器，存储器1005可选的还可以是独立于前述处理器1001的存储装置。本领域技术人员可以理解，图1中示出的硬件结构并不构成对本申请的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置。

继续参照图1，图1中作为一种计算机存储介质的存储器1005中可以包括操作系统、网络通信模块以及计算机可读指令。在图1中，网络通信模块可用于连接服务器，与服务器进行数据通信；而处理器1001可以用于调用存

储器 1005 中存储的计算机可读指令，并执行本申请实施例提供的网页数据爬取方法。

基于上述硬件结构，提出本申请网页数据爬取方法的各个实施例。

本申请提供一种网页数据爬取方法。

请参照图 2，图 2 为本申请网页数据爬取方法第一实施例的流程示意图。

在本实施例中，所述网页数据爬取方法应用于网页数据爬取系统，所述网页数据爬取系统包括控制服务器、及与所述控制服务器连接的多个爬虫服务器，所述网页数据爬取方法包括：

步骤 S10，所述控制服务器在接收到所述爬虫服务器发送的第一网页数据时，对所述第一网页数据进行特征提取，得到第一文本特征向量；

在本实施例中，该网页数据爬取方法应用于网页数据爬取系统，该网页数据爬取系统是基于分布式爬虫 Scrapy-Redis 技术构建的，其中，Scrapy-Redis 是一个基于 Redis 的 Scrapy 分布式组件，它利用 Redis 对待爬取的 URL 任务进行存储和调度，并对爬取产生的网页数据进行存储以供后续处理使用。该网页数据爬取系统包括控制服务器（Master）和多个爬虫服务器（Slave），多个爬虫服务器分别与控制服务器通信连接。其中，主控制服务器的职责是负责网页数据的指纹生成、指纹判重，URL（Uniform Resource Locator，统一资源定位符）任务的分配、URL 判重，以及网页数据的存储。爬虫服务器，其主要负责执行爬虫程序爬取网页数据，并将爬取到的网页数据和爬取过程中新的 URL 任务提交到主控制服务器的 Redis 数据库中。

在本实施例中，该网页数据爬取方法由控制服务器实现。在数据爬取过程中，爬虫服务器可从控制服务器上基于 Redis 技术构建的待爬取队列中获取待爬取的 URL 任务，然后根据该获取到的待爬取的 URL 任务去爬取对应的网页数据（记为第一网页数据），进而将该爬取到的第一网页数据发送至控制服务器。此时，控制服务器在接收到爬虫服务器发送的第一网页数据时，对该第一网页数据进行特征提取，得到第一文本特征向量。具体的，步骤 S10 包括：

步骤 a1，所述控制服务器在接收到所述爬虫服务器发送的第一网页数据时，对所述第一网页数据进行分词处理，得到第一分词集；

控制服务器在接收到爬虫服务器发送的第一网页数据时，对该第一网页数据进行分词处理，得到第一分词集。需要说明的是，分词处理可以通过分

词工具实施，例如汉语词法分析系统 ICTCLAS，清华大学中文词法分析程序 THULAC，语言技术平台 LTP 等。分词主要是根据中文语言的特点，将所述样本数据中的每条中文文本切割成一个一个的单词，并进行词性标注。

步骤 a2，根据预设停用词表过滤所述第一分词集中的停用词，得到第二分词集；

然后，对第一分词集中的分词进行词性标注，并根据预设停用词表过滤第一分词集中的停用词，得到第二分词集。其中，停用词主要包括两类：第一类是使用过于频繁的一些单词，例如“我”、“就”等，这类词几乎在每个文档中均会出现；第二类是在文本中出现频率很高，但无实际意义的单词，这类词只有将其放入一个完整的句子中才有一定作用，包括语气助词、副词、介词、连接词等，如“的”、“在”等。通过过滤停用词，可避免无意义词的影响，从而有利于提高文本特征向量生成的准确度，进而提高网页重复率检测的准确度。

步骤 a3，根据预设规则计算所述第二分词集中各分词的权重，并根据所述第二分词集中各分词及其权重得到第一文本特征向量。

在得到第二分词集之后，根据预设规则计算第二分词集中各分词的权重，并根据第二分词集中各分词及其权重得到第一文本特征向量。具体的，步骤 a3 包括：

步骤 a31，计算所述第二分词集中各分词的词频和逆向文件频率；

在本实施例中，先计算第二分词集中各分词的词频和逆向文件频率，其中，词频 TF 表示某个分词在第一数据网页中出现的频率，逆向文件频率 IDF 表示该关键词在所有数据网页中的分布情况，是一个词语普遍重要性的度量。TF 和 IDF 的计算公式如下：

$$TF = \frac{n_i}{n}, \quad IDF = \log \frac{N}{1 + N_i},$$

其中， n_i 表示该分词在第一数据网页中出现的次数， n 表示第一数据网页中的关键词总数， N 表示数据网页集的网页总数， N_i 表示在数据网页集在该分词 i 的网页数。

步骤 a32，根据所述各分词的词频和逆向文件频率计算得到各分词的词频-逆向文件频率，并将所述词频-逆向文件频率作为权重；

然后，根据各分词的词频和逆向文件频率计算得到各分词的词频-逆向文

件频率，并将词频-逆向文件频率作为权重。其中，词频-逆向文件频率 TF-IDF 为词频和逆向文件频率的乘积，即 $TF-IDF = TF \times IDF$ 。

步骤 a33，根据所述权重的大小选取预设数量的权重，并根据选取的权重及与所述选取的权重对应的分词，生成第一文本特征向量。

在得到各分词的权重之后，根据权重大小选取预设数量（如 k 个）的权重，即选取根据权重大小排在前 k 位的权重，分别记为 $w_1, w_2, \dots, w_k$ 。然后根据选取的权重及与这些选取的权重所对应的分词，生成第一文本特征向量。将与选取的权重所对应的分词分别记为 $s_1, s_2, \dots, s_k$ ，生成与第一网页数据对应的第一文本特征向量 $V = ((s_1, w_1), (s_k, w_2), \dots, (s_k, w_k))$ 。可选地，预设数量为 10-20 个，即 k 为 10 到 20 之间的自然数。

当然，在具体实施例中，还可以根据权重大小，过滤掉权重小于预设阈值的权重，进而根据过滤后的权重及其对应的分词，生成第一文本特征向量。

步骤 S20，将所述第一文本特征向量输入预先训练好的指纹生成模型，得到第一网页指纹；

在得到第一文本特征向量之后，将第一文本特征向量输入预先训练好的指纹生成模型，得到第一网页指纹。其中，指纹生成模型的具体训练过程可参照下述第二实施例，此处不作赘述。

步骤 S30，计算所述第一网页指纹与预设存储数据库中的存储网页指纹之间的相似度值，并判断所述相似度值中是否存在大于预设阈值的相似度值；

根据训练好的指纹生成模型，可对爬取到的每一网页数据生成唯一对应的一个网页指纹，有相同网页指纹的网页数据可被认为是一样的，相似内容的网页数据也有着相似的网页指纹，因此可根据第一网页数据对应的网页指纹（即第一网页指纹）与预设存储数据库中的存储网页指纹之间的相似度值，来判断第一网页数据与已爬取入库的网页数据是否有重复。具体的，由于网页指纹是一哈希字符串，因此可通过计算网页指纹之间的海明距离作为相似度值，然后检测计算得到的相似度值中是否大于预设阈值的相似度值。其中，海明距离是指在信息编码中，两个合法代码对应位上编码不同的位数称为码距。具体的，海明距离的计算方法可参照现有技术，此处不作赘述。该预设阈值可根据实际情况进行设定，此处不作具体限定。

步骤 S40，若所述相似度值中不存在大于预设阈值的相似度值，则将所述第一网页数据存入所述预设存储数据库中。

若第一网页指纹与预设存储数据库中的存储网页指纹之间的相似度值中不存在大于预设阈值的相似度值，则说明第一网页数据与已爬取的网页数据不重复，此时，则将该第一网页数据存入该预设存储数据库中，以供后续使用。

若第一网页指纹与预设存储数据库中的存储网页指纹之间的相似度值中已存在大于预设阈值的相似度值，则说明第一网页数据与某一个或某些已爬取的网页数据之间的重复度较大，该第一网页数据极可能是通过转载后经小范围添加和修改得到的新网页数据，而与该第一网页数据相类似的网页数据已经爬取入库了，此时，则删除第一该网页数据，拒绝存入预设存储数据库中。

本申请提供一种网页数据爬取方法，应用于基于分布式爬虫技术构建的网页数据爬取系统，该网页数据爬取系统包括控制服务器、及与该控制服务器连接的多个爬虫服务器，该方法包括：控制服务器在接收到爬虫服务器发送的第一网页数据时，对该第一网页数据进行特征提取，得到第一文本特征向量；将第一文本特征向量输入预先训练好的指纹生成模型，得到第一网页指纹；计算第一网页指纹与预设存储数据库中的存储网页指纹之间的相似度值，并判断计算得到的相似度值中是否存在大于预设阈值的相似度值；若不存在，才将该第一网页数据存入该预设存储数据库中。本申请通过构建指纹生成模型，对爬取到的网页数据生成对应指纹，然后与存储数据库中已爬取保存的网页数据的指纹进行对比，以判断是否重复，进而对不重复的数据才进行保存，从而可避免重复的网页数据入库，因此，本申请可解决现有技术中基于网络爬虫爬取得到的数据重复性较高的问题。同时，通过与分布式爬虫技术相结合，可实现网页数据的快速抓取，保证在较短的时间内完成数据的采集，提高了数据采集的效率。

进一步的，参照图 3，图 3 为本申请网页数据爬取方法第二实施例的流程示意图。

基于上述图 2 所示的第一实施例，在步骤 S10 之前，该网页数据爬取方法还包括：

步骤 S50，获取训练样本，所述训练样本包括多个经标注的网页数据组合；

本实施例提供了指纹生成模型的具体训练过程。首先，获取训练样本，其中，该训练样本包括多个经标注的网页数据组合。可选地，每个网页数据

组合包括两个网页数据，标注过程为对各网页数据组合进行真实相似度值的标注。网页数据可以来源各个网站，也可以根据实时需要进行编辑创建。

步骤 S60，对所述训练样本进行特征提取，得到第二文本特征向量；

然后，对训练样本进行特征提取，得到第二文本特征向量，具体的，由于各个训练样本为包括两个网页数据的组合，此时，在对训练样本进行特征提取时，则先分别对各训练样本中的各个网页数据进行分词处理，并进行词性标注，然后根据预设停用词表过滤掉停用词，得到对应的分词集，进而根据计算该分词集中各分词的词频和逆向文件频率，根据词频和逆向文件频率计算得到词频-逆向文件频率，并将词频-逆向文件频率作为权重，进而根据该分词集中各分词及其权重得到第二文本特征向量。对应的，每个训练样本所对应的文本特征向量为两个。其中，第二文本特征向量的具体生成过程，与上述第一实施例中第一文本特征向量的生成过程基本相同，此处不作赘述。

步骤 S70，将所述第二文本特征向量输入初始神经网络模型，得到对应的第二网页指纹；

在得到第二文本特征向量之后，将第二文本特征向量输入初始神经网络模型，得到对应的第二网页指纹。具体的，将与该训练样本对应的两个第二文本特征向量依次输入初始神经网络模型，得到第二网页指纹。对应的，每个训练样本对应的第二网页指纹也是两个。

步骤 S80，根据所述第二网页指纹、所述训练样本的真实相似度值，通过预设算法计算得到损失值；

根据第二网页指纹、训练样本的真实相似度值，通过预设算法计算得到损失值。具体的，步骤 S80 包括：

步骤 b1，根据所述第二网页指纹计算所述训练样本中的网页数据之间的相似度值，记作预测相似度值；

在本实施例中，先根据第二网页指纹计算训练样本中的网页数据之间的相似度值，记作预测相似度值，具体的，可通过计算海明距离的方法得到相似度值。

步骤 b2，根据所述预测相似度值、所述训练样本的真实相似度值，通过预设损失函数计算得到损失值；

在计算得到预测相似度值之后，根据该预测相似度值、训练样本的真实相似度值，通过预设损失函数计算得到损失值。其中，该预设损失函数为：

$$L(\hat{c}, c) = \|\hat{c} - c\|^2$$

其中， $\hat{c}$ 为预测相似度值， c 为训练样本的真实相似度值。

步骤 S90, 根据所述损失值通过梯度下降算法对所述初始神经网络模型的参数进行更新，并对所述训练样本进行迭代训练，得到训练好的指纹生成模型。

最后，根据计算得到的损失值通过梯度下降算法对该初始神经网络模型的参数进行更新，并对每一个训练样本进行迭代训练，即根据损失值来更新初始神经网络模型中各层结点的梯度，进而更新各结点的权值参数，不断输入训练样本所对应的第二文本特征向量进行迭代直至网络收敛，直至该损失值稳定下降到一个较小范围（如低于一预设阈值或达到最小值），此时，可得到训练好的神经网络模型，即训练好的指纹生成模型。通过梯度下降算法可求解大规模样本数据的优化问题，具体的梯度下降算法可参照现有技术，此处不做赘述。

进一步的，基于上述各实施例，提出本申请网页数据爬取方法的第三实施例。

在本实施例中，在步骤 S40 之后，该网页数据爬取方法还可以包括：

步骤 A, 在接收到所述爬虫服务器发送的待爬取统一资源定位符 URL 时，计算所述待爬取 URL 的哈希值；

步骤 B, 检测所述哈希值是否已存在于预设键-值对存储 Redis 数据库中；

在本实施例中，爬虫服务器在爬取数据的过程中，会从抓取到的网页中提取到新的 URL（Uniform Resource Locator，统一资源定位符），进而将新的 URL（记为待爬取 URL）发送给控制服务器，此时，控制服务器在接收到爬虫服务器发送的待爬取 URL 时，计算待爬取 URL 的哈希值。进而，检测该哈希值是否已存在于预设 Redis（键-值对存储）数据库中。其中，哈希值的计算方法可参照现有技术，此处不作赘述。

若所述哈希值未存在于预设 Redis 数据库中，则执行步骤 C：将所述待爬取 URL 添加至预设待爬取队列，以供所述爬虫服务器从所述预设待爬取队列获取所述待爬取 URL，并根据所述待爬取 URL 爬取新的网页数据；

若所述哈希值已存在于所述预设 Redis 数据库中，则执行步骤 D：删除所述待爬取 URL。

若该哈希值未存在于预设 Redis 数据库中，则说明该待爬取 URL 未被抓

取过，此时，则将该待爬取 URL 添加至预设待爬取队列，以供爬虫服务器后续从预设待爬取队列获取该待爬取 URL，并根据该待爬取 URL 爬取新的网页数据。

若该哈希值存在于预设 Redis 数据库中，则说明该待爬取 URL 之前已被抓取过，此时，则删除该待爬取 URL，拒绝将待爬取 URL 添加至预设待爬取队列中。

本实施例中，通过对重复的 URL 进行去重处理，可避免多次抓取同一网页，从而可进一步降低爬取得到的网页数据的重复率。

进一步的，基于上述各实施例，提出本申请网页数据爬取方法的第三实施例。

在本实施例中，在第一实施例或第二实施例的步骤 S40 之后，该网页数据爬取方法还可以包括：

步骤 E，统计在预设周期内所述爬虫服务器的数据发送量和所述预设存储数据库的数据入库量，并根据所述数据发送量和所述数据入库量生成对应的统计报告；

步骤 F，将所述统计报告发送至预设工作终端，以供工作人员进行数据重复分析。

在本实施例中，为方便工作人员了解网页数据的重复情况，控制服务器可统计在预设周期内爬虫服务器的数据发送量和预设存储数据库的数据入库量，并根据数据发送量和数据入库量生成对应的统计报告。其中，预设周期可以设为一天、或一周等，可根据实际情况进行设定，此处不作限定。统计时间可根据预设周期来定，例如，预设周期为一周时，则每隔一周进行一次统计。统计报告中，可以显示数据发送量、数据入库量及数据入库率（数据入库率=数据入库量/数据发送量）等。

在生成统计报告之后，将该统计报告发送至预设工作终端，以供工作人员进行数据重复分析。

进一步的，由于爬虫服务器可以从多个不同的网站进行爬取，对应的，控制服务器接收到的网页数据也是来源于不同的网站，在生成统计报告时可以对以网站（数据源）为维度进行分类分析，以得出不同网站的数据入库率；对于入库率低的网站，可认为是数据重复比较严重的网站，工作人员可以去除该网站的数据爬取，从而减少服务器的资源消耗。

此外，还可以每隔预设时间，对上述统计报告做进一步的汇总统计，得到汇总统计报告。如当每隔一周进行一次统计，得到统计报告后，可每隔三个月，对上述每周的统计报告进行汇总统计，得到季度统计报告。该季度统计报告可以图表的形式展示数据发送量、数据入库量及数据入库率的变化趋势。进而，将该汇总统计报告发送至该预设工作端，以供工作人员了解从宏观上了解数据的重复情况。

本申请还提供一种网页数据爬取装置。

参照图4，图4为本申请网页数据爬取装置第一实施例的功能模块示意图。

在本实施例中，所述网页数据爬取装置包括：

第一特征提取模块10，用于所述控制服务器在接收到所述爬虫服务器发送的第一网页数据时，对所述第一网页数据进行特征提取，得到第一文本特征向量；

第一指纹生成模块20，用于将所述第一文本特征向量输入预先训练好的指纹生成模型，得到第一网页指纹；

相似度值判断模块30，用于计算所述第一网页指纹与预设存储数据库中的存储网页指纹之间的相似度值，并判断所述相似度值中是否存在大于预设阈值的相似度值；

网页数据存储模块40，用于若所述相似度值中不存在大于预设阈值的相似度值，则将所述第一网页数据存入所述预设存储数据库中。

其中，上述网页数据爬取装置各虚拟功能模块存储于图1所示网页数据爬取系统的存储器1005中，用于实现计算机可读指令的所有功能；各模块被处理器1001执行时，可实现对爬取到的网页数据生成对应指纹，然后与预设存储数据库中已爬取保存的网页数据的指纹进行对比，以判断是否重复，进而剔除重复的网页数据的功能。

进一步的，所述网页数据爬取装置还包括：

训练样本获取模块，用于获取训练样本，所述训练样本包括多个经标注的网页数据组合；

第二特征提取模块，用于对所述训练样本进行特征提取，得到第二文本特征向量；

第二指纹生成模块，用于将所述第二文本特征向量输入初始神经网络模型，得到对应的第二网页指纹；

损失值计算模块，用于根据所述第二网页指纹、所述训练样本的真实相似度值，通过预设算法计算得到损失值；

指纹生成模型训练模块，用于根据所述损失值通过梯度下降算法对所述初始神经网络模型的参数进行更新，并对所述训练样本进行迭代训练，得到训练好的指纹生成模型。

进一步的，所述损失值计算模块包括：

预测相似度值计算单元，用于根据所述第二网页指纹计算所述训练样本中的网页数据之间的相似度值，记作预测相似度值；

损失值计算单元，用于根据所述预测相似度值、所述训练样本的真实相似度值，通过预设损失函数计算得到损失值；

其中，所述预设损失函数为：

$$L(\hat{c}, c) = \|\hat{c} - c\|^2$$

其中， $\hat{c}$ 为所述预测相似度值， c 为所述训练样本的真实相似度值。

进一步的，所述第一特征提取模块 10 包括：

分词处理单元，用于所述控制服务器在接收到所述爬虫服务器发送的第一网页数据时，对所述第一网页数据进行分词处理，得到第一分词集；

停用词过滤单元，用于根据预设停用词表过滤所述第一分词集中的停用词，得到第二分词集；

第一特征获取单元，用于根据预设规则计算所述第二分词集中各分词的权重，并根据所述第二分词集中各分词及其权重得到第一文本特征向量。

进一步的，所述第一特征获取单元包括：

第一计算子单元，用于计算所述第二分词集中各分词的词频和逆向文件频率；

第二计算子单元，用于根据所述各分词的词频和逆向文件频率计算得到各分词的词频-逆向文件频率，并将所述词频-逆向文件频率作为权重；

第一特征获取子单元，用于根据所述权重的大小选取预设数量的权重，并根据选取的权重及与所述选取的权重对应的分词，生成第一文本特征向量。

进一步的，所述网页数据爬取装置还包括：

哈希值计算模块，用于在接收到所述爬虫服务器发送的待爬取统一资源定位符 URL 时，计算所述待爬取 URL 的哈希值；

哈希值检测模块，用于检测所述哈希值是否已存在于预设键-值对存储

Redis 数据库中；

URL 添加模块，用于若所述哈希值未存在于预设 Redis 数据库中，则将所述待爬取 URL 添加至预设待爬取队列，以供所述爬虫服务器从所述预设待爬取队列获取所述待爬取 URL，并根据所述待爬取 URL 爬取新的网页数据；

URL 删除模块，用于若所述哈希值已存在于所述预设 Redis 数据库中，则删除所述待爬取 URL。

进一步的，所述网页数据爬取装置还包括：

统计报告生成模块，用于统计在预设周期内所述爬虫服务器的数据发送量和所述预设存储数据库的数据入库量，并根据所述数据发送量和所述数据入库量生成对应的统计报告；

统计报告发送模块，用于将所述统计报告发送至预设工作终端，以供工作人员进行数据重复分析。

其中，上述网页数据爬取装置中各个模块的功能实现与上述网页数据爬取方法实施例各步骤相对应，其功能和实现过程在此处不再一一赘述。

本申请还提供一种网页数据爬取系统，该网页数据爬取系统包括控制服务器、及与所述控制服务器连接的多个爬虫服务器，还包括存储器、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可读指令，所述计算机可读指令被所述处理器执行时实现如以上任一项实施例所述的网页数据爬取方法的步骤。

本申请网页数据爬取系统的具体实施例与上述网页数据爬取方法各实施例基本相同，在此不作赘述。

本申请还提供一种计算机可读存储介质，所述计算机可读存储介质可以为非易失性可读存储介质，该计算机可读存储介质上存储有计算机可读指令，所述计算机可读指令被处理器执行时实现如以上任一项实施例所述的网页数据爬取方法的步骤。

本申请计算机可读存储介质的具体实施例与上述网页数据爬取方法各实施例基本相同，在此不作赘述。

需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者系统不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者系统所固有的要素。在没有更多限制的情况下，

由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、方法、物品或者系统中还存在另外的相同要素。

上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。

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上述实施例方法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，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的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，该计算机软件产品存储在如上所述的一个存储介质(如 ROM/RAM、磁碟、光盘)中，包括若干指令用以使得一台设备(可以是手机，计算机，服务器，空调器，或者网络设备等)执行本申请各个实施例所述的方法。

以上仅为本申请的优选实施例，并非因此限制本申请的专利范围，凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本申请的专利保护范围内。

权 利 要 求 书

1、一种网页数据爬取方法，其特征在于，所述网页数据爬取方法应用于网页数据爬取系统，所述网页数据爬取系统包括控制服务器、及与所述控制服务器连接的多个爬虫服务器，所述网页数据爬取方法包括以下步骤：

所述控制服务器在接收到所述爬虫服务器发送的第一网页数据时，对所述第一网页数据进行特征提取，得到第一文本特征向量；

将所述第一文本特征向量输入预先训练好的指纹生成模型，得到第一网页指纹；

计算所述第一网页指纹与预设存储数据库中的存储网页指纹之间的相似度值，并判断所述相似度值中是否存在大于预设阈值的相似度值；

若所述相似度值中不存在大于预设阈值的相似度值，则将所述第一网页数据存入所述预设存储数据库中；

其中，所述控制服务器在接收到所述爬虫服务器发送的第一网页数据时，对所述第一网页数据进行特征提取，得到第一文本特征向量的步骤包括：

所述控制服务器在接收到所述爬虫服务器发送的第一网页数据时，对所述第一网页数据进行分词处理，得到第一分词集；

根据预设停用词表过滤所述第一分词集中的停用词，得到第二分词集；

根据预设规则计算所述第二分词集中各分词的权重，并根据所述第二分词集中各分词及其权重得到第一文本特征向量。

2、如权利要求1所述的网页数据爬取方法，其特征在于，所述控制服务器在接收到从所述爬虫服务器发送的第一网页数据时，对所述第一网页数据进行预处理，得到第一文本特征向量的步骤之前，所述方法还包括：

获取训练样本，所述训练样本包括多个经标注的网页数据组合；

对所述训练样本进行特征提取，得到第二文本特征向量；

将所述第二文本特征向量输入初始神经网络模型，得到对应的第二网页指纹；

根据所述第二网页指纹、所述训练样本的真实相似度值，通过预设算法计算得到损失值；

根据所述损失值通过梯度下降算法对所述初始神经网络模型的参数进行更新，并对所述训练样本进行迭代训练，得到训练好的指纹生成模型。

3、如权利要求 2 所述的网页数据爬取方法，其特征在于，所述根据所述第二网页指纹、所述训练样本的真实相似度值，通过预设损失函数计算得到损失值的步骤包括：

根据所述第二网页指纹计算所述训练样本中的网页数据之间的相似度值，记作预测相似度值；

根据所述预测相似度值、所述训练样本的真实相似度值，通过预设损失函数计算得到损失值；

其中，所述预设损失函数为： $L(\hat{c}, c) = \|\hat{c} - c\|^2$

其中， $\hat{c}$ 为所述预测相似度值， c 为所述训练样本的真实相似度值。

4、如权利要求 1 所述的网页数据爬取方法，其特征在于，所述根据预设规则计算所述第二分词集中各分词的权重，并根据所述第二分词集中各分词及其权重得到第一文本特征向量的步骤包括：

计算所述第二分词集中各分词的词频和逆向文件频率；

根据所述各分词的词频和逆向文件频率计算得到各分词的词频-逆向文件频率，并将所述词频-逆向文件频率作为权重；

根据所述权重的大小选取预设数量的权重，并根据选取的权重及与所述选取的权重对应的分词，生成第一文本特征向量。

5、如权利要求 1 所述的网页数据爬取方法，其特征在于，所述网页数据爬取方法还包括：

在接收到所述爬虫服务器发送的待爬取统一资源定位符 URL 时，计算所述待爬取 URL 的哈希值；

检测所述哈希值是否已存在于预设键-值对存储 Redis 数据库中；

若所述哈希值未存在于预设 Redis 数据库中，则将所述待爬取 URL 添加至预设待爬取队列，以供所述爬虫服务器从所述预设待爬取队列获取所述待爬取 URL，并根据所述待爬取 URL 爬取新的网页数据；

若所述哈希值已存在于所述预设 Redis 数据库中，则删除所述待爬取 URL。

6、如权利要求 1 所述的网页数据爬取方法，其特征在于，所述网页数据爬取方法还包括：

统计在预设周期内所述爬虫服务器的数据发送量和所述预设存储数据库的数据入库量，并根据所述数据发送量和所述数据入库量生成对应的统计报告；

将所述统计报告发送至预设工作终端，以供工作人员进行数据重复分析。

7、一种网页数据爬取装置，其特征在于，所述网页数据爬取装置包括：

第一特征提取模块，用于所述控制服务器在接收到所述爬虫服务器发送的第一网页数据时，对所述第一网页数据进行特征提取，得到第一文本特征向量；

第一指纹生成模块，用于将所述第一文本特征向量输入预先训练好的指纹生成模型，得到第一网页指纹；

相似度值判断模块，用于计算所述第一网页指纹与预设存储数据库中的存储网页指纹之间的相似度值，并判断所述相似度值中是否存在大于预设阈值的相似度值；

网页数据存储模块，用于若所述相似度值中不存在大于预设阈值的相似度值，则将所述第一网页数据存入所述预设存储数据库中；

其中，所述第一特征提取模块包括：分词处理单元，用于所述控制服务器在接收到所述爬虫服务器发送的第一网页数据时，对所述第一网页数据进行分词处理，得到第一分词集；停用词过滤单元，用于根据预设停用词表过滤所述第一分词集中的停用词，得到第二分词集；第一特征获取单元，用于根据预设规则计算所述第二分词集中各分词的权重，并根据所述第二分词集中各分词及其权重得到第一文本特征向量。

8、如权利要求7所述的网页数据爬取装置，其特征在于，所述网页数据爬取装置还包括：

训练样本获取模块，用于获取训练样本，所述训练样本包括多个经标注的网页数据组合；

第二特征提取模块，用于对所述训练样本进行特征提取，得到第二文本特征向量；

第二指纹生成模块，用于将所述第二文本特征向量输入初始神经网络模型，得到对应的第二网页指纹；

损失值计算模块，用于根据所述第二网页指纹、所述训练样本的真实相似度值，通过预设算法计算得到损失值；

指纹生成模型训练模块，用于根据所述损失值通过梯度下降算法对所述初始神经网络模型的参数进行更新，并对所述训练样本进行迭代训练，得到训练好的指纹生成模型。

9、如权利要求 8 所述的网页数据爬取装置，其特征在于，所述损失值计算模块包括：

预测相似度值计算单元，用于根据所述第二网页指纹计算所述训练样本中的网页数据之间的相似度值，记作预测相似度值；

损失值计算单元，用于根据所述预测相似度值、所述训练样本的真实相似度值，通过预设损失函数计算得到损失值；

其中，所述预设损失函数为： $L(\hat{c}, c) = \|\hat{c} - c\|^2$

其中， $\hat{c}$ 为所述预测相似度值， c 为所述训练样本的真实相似度值。

10、如权利要求 7 所述的网页数据爬取装置，其特征在于，所述第一特征提取模块包括：

分词处理单元，用于所述控制服务器在接收到所述爬虫服务器发送的第一网页数据时，对所述第一网页数据进行分词处理，得到第一分词集；

停用词过滤单元，用于根据预设停用词表过滤所述第一分词集中的停用词，得到第二分词集；

第一特征获取单元，用于根据预设规则计算所述第二分词集中各分词的权重，并根据所述第二分词集中各分词及其权重得到第一文本特征向量。

11、如权利要求 7 所述的网页数据爬取装置，其特征在于，所述网页数据爬取装置还包括：

哈希值计算模块，用于在接收到所述爬虫服务器发送的待爬取统一资源定位符 URL 时，计算所述待爬取 URL 的哈希值；

哈希值检测模块，用于检测所述哈希值是否已存在于预设键-值对存储 Redis 数据库中；

URL 添加模块，用于若所述哈希值未存在于预设 Redis 数据库中，则将所述待爬取 URL 添加至预设待爬取队列，以供所述爬虫服务器从所述预设待爬取队列获取所述待爬取 URL，并根据所述待爬取 URL 爬取新的网页数据；

URL 删除模块，用于若所述哈希值已存在于所述预设 Redis 数据库中，则删除所述待爬取 URL。

12、如权利要求 7 所述的网页数据爬取装置，其特征在于，所述网页数据爬取装置还包括：

统计报告生成模块，用于统计在预设周期内所述爬虫服务器的数据发送量和所述预设存储数据库的数据入库量，并根据所述数据发送量和所述数据

入库量生成对应的统计报告；

统计报告发送模块，用于将所述统计报告发送至预设工作终端，以供工作人员进行数据重复分析。

13、一种网页数据爬取系统，其特征在于，所述网页数据爬取系统包括控制服务器、及与所述控制服务器连接的多个爬虫服务器，还包括存储器、处理器以及存储在所述存储器上并可被所述处理器执行的计算机可读指令，其中所述计算机可读指令被所述处理器执行时，实现以下步骤：

所述控制服务器在接收到所述爬虫服务器发送的第一网页数据时，对所述第一网页数据进行特征提取，得到第一文本特征向量；

将所述第一文本特征向量输入预先训练好的指纹生成模型，得到第一网页指纹；

计算所述第一网页指纹与预设存储数据库中的存储网页指纹之间的相似度值，并判断所述相似度值中是否存在大于预设阈值的相似度值；

若所述相似度值中不存在大于预设阈值的相似度值，则将所述第一网页数据存入所述预设存储数据库中；

所述计算机可读指令被所述处理器执行时，还实现以下步骤：所述控制服务器在接收到所述爬虫服务器发送的第一网页数据时，对所述第一网页数据进行分词处理，得到第一分词集；根据预设停用词表过滤所述第一分词集中的停用词，得到第二分词集；根据预设规则计算所述第二分词集中各分词的权重，并根据所述第二分词集中各分词及其权重得到第一文本特征向量。

14、如权利要求 13 所述的网页数据爬取系统，其特征在于，所述计算机可读指令被所述处理器执行时，还实现以下步骤：

获取训练样本，所述训练样本包括多个经标注的网页数据组合；

对所述训练样本进行特征提取，得到第二文本特征向量；

将所述第二文本特征向量输入初始神经网络模型，得到对应的第二网页指纹；

根据所述第二网页指纹、所述训练样本的真实相似度值，通过预设算法计算得到损失值；

根据所述损失值通过梯度下降算法对所述初始神经网络模型的参数进行更新，并对所述训练样本进行迭代训练，得到训练好的指纹生成模型。

15、如权利要求 14 所述的网页数据爬取系统，其特征在于，所述计算机

可读指令被所述处理器执行时，还实现以下步骤：

根据所述第二网页指纹计算所述训练样本中的网页数据之间的相似度值，记作预测相似度值；

根据所述预测相似度值、所述训练样本的真实相似度值，通过预设损失函数计算得到损失值；

其中，所述预设损失函数为： $L(\hat{c}, c) = \|\hat{c} - c\|^2$

其中， $\hat{c}$ 为所述预测相似度值， c 为所述训练样本的真实相似度值。

16、如权利要求13所述的网页数据爬取系统，其特征在于，所述计算机可读指令被所述处理器执行时，还实现以下步骤：

计算所述第二分词集中各分词的词频和逆向文件频率；

根据所述各分词的词频和逆向文件频率计算得到各分词的词频-逆向文件频率，并将所述词频-逆向文件频率作为权重；

根据所述权重的大小选取预设数量的权重，并根据选取的权重及与所述选取的权重对应的分词，生成第一文本特征向量。

17、一种计算机可读存储介质，其特征在于，所述计算机可读存储介质上存储有计算机可读指令，其中所述计算机可读指令被处理器执行时，实现如下步骤：

所述控制服务器在接收到所述爬虫服务器发送的第一网页数据时，对所述第一网页数据进行特征提取，得到第一文本特征向量；

将所述第一文本特征向量输入预先训练好的指纹生成模型，得到第一网页指纹；

计算所述第一网页指纹与预设存储数据库中的存储网页指纹之间的相似度值，并判断所述相似度值中是否存在大于预设阈值的相似度值；

若所述相似度值中不存在大于预设阈值的相似度值，则将所述第一网页数据存入所述预设存储数据库中；

所述计算机可读指令被处理器执行时，还实现如下步骤：所述控制服务器在接收到所述爬虫服务器发送的第一网页数据时，对所述第一网页数据进行分词处理，得到第一分词集；根据预设停用词表过滤所述第一分词集中的停用词，得到第二分词集；根据预设规则计算所述第二分词集中各分词的权重，并根据所述第二分词集中各分词及其权重得到第一文本特征向量。

18、如权利要求17所述的计算机可读存储介质，其特征在于，所述计算

机可读指令被处理器执行时，还实现如下步骤：

获取训练样本，所述训练样本包括多个经标注的网页数据组合；

对所述训练样本进行特征提取，得到第二文本特征向量；

将所述第二文本特征向量输入初始神经网络模型，得到对应的第二网页指纹；

根据所述第二网页指纹、所述训练样本的真实相似度值，通过预设算法计算得到损失值；

根据所述损失值通过梯度下降算法对所述初始神经网络模型的参数进行更新，并对所述训练样本进行迭代训练，得到训练好的指纹生成模型。

19、如权利要求 18 所述的计算机可读存储介质，其特征在于，所述计算机可读指令被处理器执行时，还实现如下步骤：

根据所述第二网页指纹计算所述训练样本中的网页数据之间的相似度值，记作预测相似度值；

根据所述预测相似度值、所述训练样本的真实相似度值，通过预设损失函数计算得到损失值；

其中，所述预设损失函数为： $L(\hat{c}, c) = \|\hat{c} - c\|^2$

其中， $\hat{c}$ 为所述预测相似度值， c 为所述训练样本的真实相似度值。

20、如权利要求 17 所述的计算机可读存储介质，其特征在于，所述计算机可读指令被处理器执行时，还实现如下步骤：

计算所述第二分词集中各分词的词频和逆向文件频率；

根据所述各分词的词频和逆向文件频率计算得到各分词的词频-逆向文件频率，并将所述词频-逆向文件频率作为权重；

根据所述权重的大小选取预设数量的权重，并根据选取的权重及与所述选取的权重对应的分词，生成第一文本特征向量。

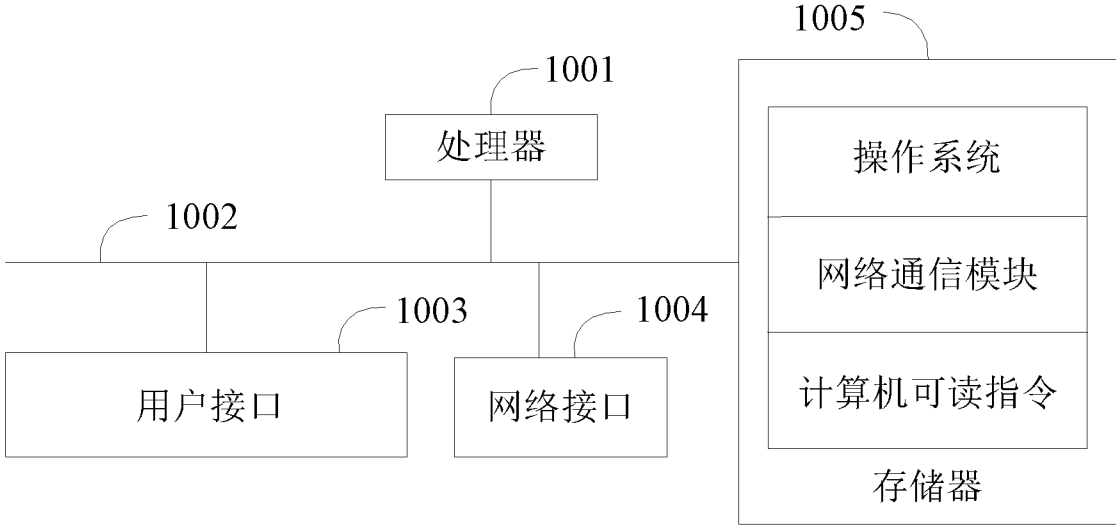


图 1

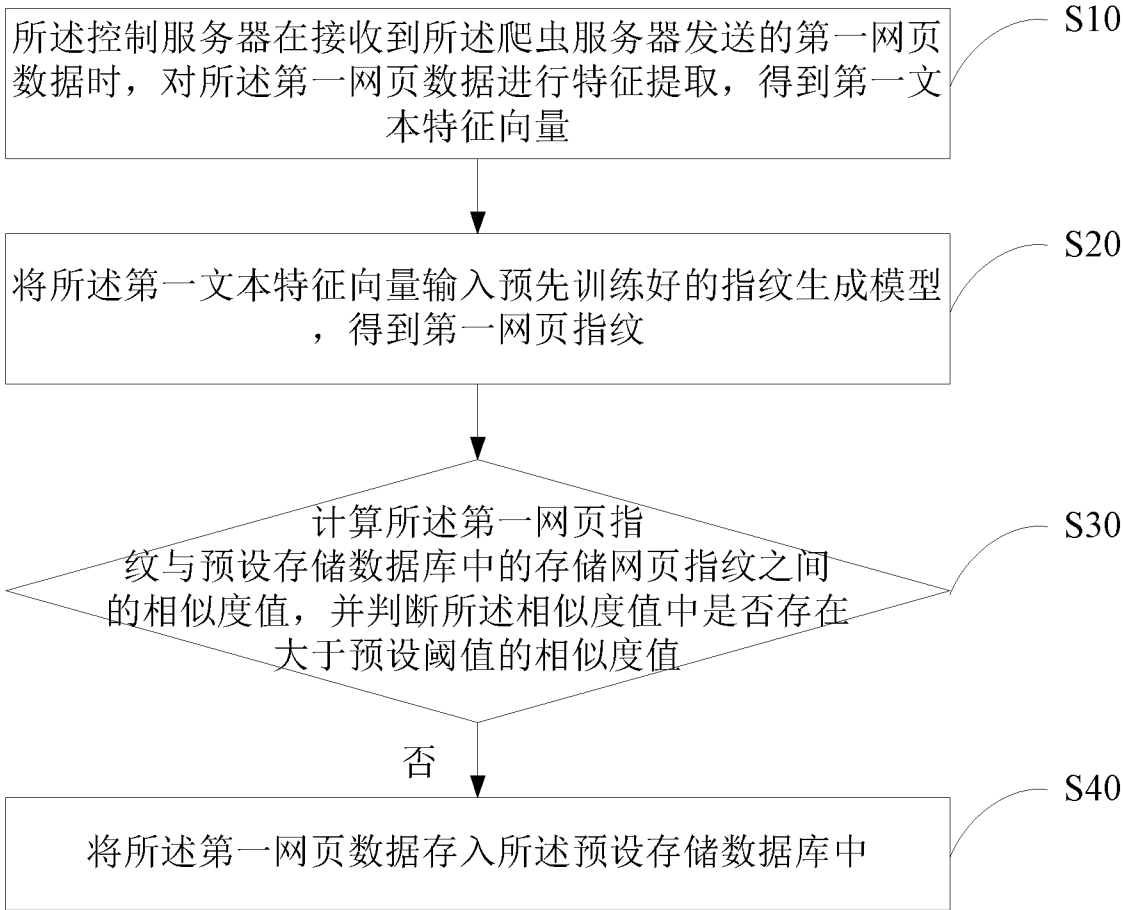


图 2

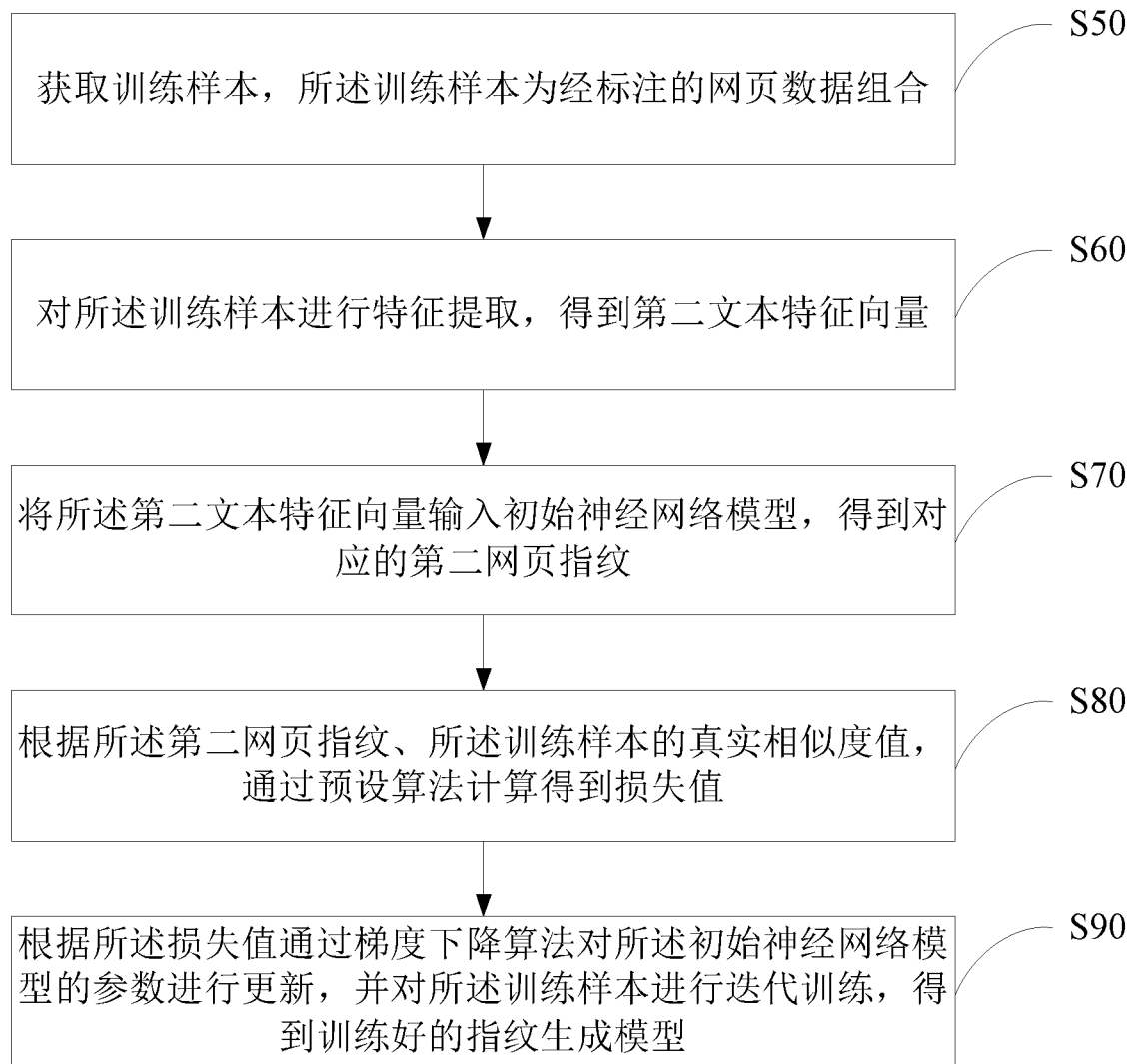


图 3

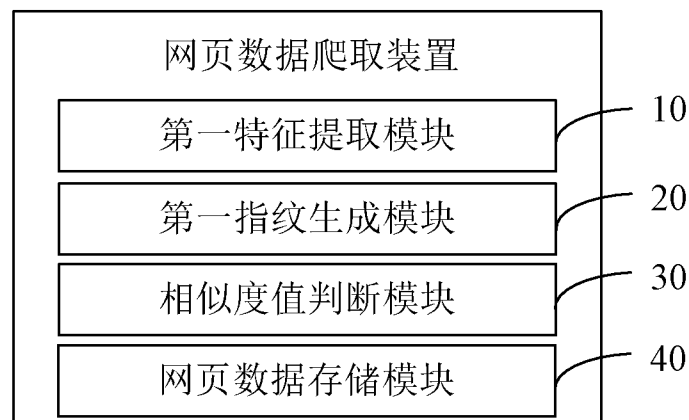


图 4

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/118144

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/951(2019.01)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)

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)

CNKI, CNPAT, WPI, EPODOC: 网页, 爬虫, 去重, 重复, 神经网络, 词频, 指纹, 哈希, webpage, crawl, duplicate, fingerprint, neural, network, hash

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 101645082 A (HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY) 10 February 2010 (2010-02-10) description, page 5, line 4 to page 9, line 14	1-20
PX	CN 109918554 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 21 June 2019 (2019-06-21) description, paragraph [0058]-[0084] and claims 1-10	1-20
A	CN 108647263 A (HUIYIN INSTITUTE OF TECHNOLOGY) 12 October 2018 (2018-10-12) entire document	1-20
A	CN 106126512 A (BEIJING TOPSEC NETWORK SECURITY TECHNOLOGY CO., LTD. et al.) 16 November 2016 (2016-11-16) entire document	1-20



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

07 January 2020

Date of mailing of the international search report

27 February 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/118144

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	101645082	A	10 February 2010	None	
CN	109918554	A	21 June 2019	None	
CN	108647263	A	12 October 2018	None	
CN	106126512	A	16 November 2016	None	

国际检索报告

国际申请号

PCT/CN2019/118144

A. 主题的分类 G06F 16/951 (2019.01) i 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) 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)) CNKI, CNPAT, WPI, EPDOC: 网页, 爬虫, 去重, 重复, 神经网络, 词频, 指纹, 哈希, webpage, crawl, duplicate, fingerprint, neural, network, hash																	
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>X</td> <td>CN 101645082 A (华中科技大学) 2010年 2月 10日 (2010 - 02 - 10) 说明书第5页第4行到第9页第14行</td> <td>1-20</td> </tr> <tr> <td>PX</td> <td>CN 109918554 A (平安科技深圳有限公司) 2019年 6月 21日 (2019 - 06 - 21) 说明书第[0058]-[0084]段, 权利要求1-10</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 108647263 A (淮阴工学院) 2018年 10月 12日 (2018 - 10 - 12) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 106126512 A (北京天融信网络安全技术有限公司 等) 2016年 11月 16日 (2016 - 11 - 16) 全文</td> <td>1-20</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	X	CN 101645082 A (华中科技大学) 2010年 2月 10日 (2010 - 02 - 10) 说明书第5页第4行到第9页第14行	1-20	PX	CN 109918554 A (平安科技深圳有限公司) 2019年 6月 21日 (2019 - 06 - 21) 说明书第[0058]-[0084]段, 权利要求1-10	1-20	A	CN 108647263 A (淮阴工学院) 2018年 10月 12日 (2018 - 10 - 12) 全文	1-20	A	CN 106126512 A (北京天融信网络安全技术有限公司 等) 2016年 11月 16日 (2016 - 11 - 16) 全文	1-20
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求															
X	CN 101645082 A (华中科技大学) 2010年 2月 10日 (2010 - 02 - 10) 说明书第5页第4行到第9页第14行	1-20															
PX	CN 109918554 A (平安科技深圳有限公司) 2019年 6月 21日 (2019 - 06 - 21) 说明书第[0058]-[0084]段, 权利要求1-10	1-20															
A	CN 108647263 A (淮阴工学院) 2018年 10月 12日 (2018 - 10 - 12) 全文	1-20															
A	CN 106126512 A (北京天融信网络安全技术有限公司 等) 2016年 11月 16日 (2016 - 11 - 16) 全文	1-20															
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																	
<table border="0"> <tr> <td> * 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 </td> <td> “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 </td> </tr> </table>			* 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													
* 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															
2020年 1月 7日		2020年 2月 27日															
ISA/CN的名称和邮寄地址		授权官员															
中国国家知识产权局 (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		赵晓春 电话号码 86-(10)-53961295															

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/118144

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	101645082	A	2010年 2月 10日	无	
CN	109918554	A	2019年 6月 21日	无	
CN	108647263	A	2018年 10月 12日	无	
CN	106126512	A	2016年 11月 16日	无	