基于 URL 和网页类型的网页信息采集研究

作者/张锋，天津工业大学计算机科学与软件学院 天津天狮学院工商管理学院

文章摘要：Internet上的海量数据对人们有效、快速地使用这些资源和信息提出了挑战。网页信息采集更新的方法在很大程度上决定了网页更新的效果。为提高网页信息更新的效果，本文从抓取入口页面开始，采集网页后进行去重操作，并将网页分块提取出超链接URL信息。在此基础上，应用网页更新策略提高网页更新效果。最后，应用基于URL和网页类型的采集更新检测方法来实现网页信息采集。

关键词：页面更新;入口页面;信息采集;更新检测

DOI:10.16589/j.cnki.cn11-3571/tn.2017.02.016

1. 绪论

Internet，也对人们有效、快速地使用这些资源和信息提出了一个挑战。Internet 是一个开放性、动态性和异构性的全球分布式网络，数据的海量性、资源的分散性、管理的不一致性、结构的多样性，使用户寻找自己所需的信息像大海捞针一样困难。随着网上信息资源的不断变化，采集系统需要不断更新它所访问过的网页，来判断要重新访问哪些页面以及用怎样的频率去访问。

网络信息的快速增长和网页动态变化的特性使因信息更新或网址变动造成的搜索引擎信息的缺失日益增加，导致搜索引擎整体性能下降。因此，如何快速有效地对网页进行更新，如何使网页的更新效果最好，成为一个重要的研究课题。

2. URL类型和网页类型

■ 2.1 URL 类型

统一资源定位符（URL，Uniform Resource Locator）是因特网上标准的资源的地址，用于完整地描述 Internet 上网页和其他资源地址的一种标识方法。Internet 上的每一个网页都具有一个唯一的名称标识。在 Internet 上所有资源都有一个唯一的 URL 地址。

网页的 URL 可分为 4 种类型：Root 根形式（通常代表网站首页）、Subroot 次级根形式（一个域名只跟随一个文件目录）、Path 路径形式（域名后跟随两个或多个文件目录）、File 文件形式（以文件名结尾的 URL、多级目录）。

■ 2.2 网页类型分类

网页类型包括：（1）网站中包含少量的变化页面，由大量链接组成，网站新增网页会反应在这类页面中，通常是网站首页、栏目首页。（2）网站中包含大量的稳定页面，这部分页面表示了网页的内容，这类页面不会改变，除非消失。即使发生改变，改变后的价值与原有的网页相比，改变幅度也是很小的。比如新闻报道、产品介绍、论坛帖子等。(3) 部分页面也包含大量链接，但这些链接指向的页面不是网站最新的页面。

■ 2.3 入口页面

优先抓取入口页面，能够取得较好的更新效果。通过排序算法，URL 较短的入口页面比其它页面要容易被找到。我们发现主页和其下级页面 (sub-page) 有特定的关系，并且该相关页面 (主页的下级页面) 的数量级要比主页大得多。

如果满足（1）代表这个页面的 URL 为 root 类型；（2）代表这个页面的 URL 为 subroot 或 path 类型，页面链接信息丰富，新的网页链接数目与旧的网页链接相比，达到某个阈值的特征之一，就是入口页面。

根据与主页相关的子页面在测试集里面的统计结果，以及和主页的关系，PERS 算法可以根据相关子页面查找主页，其一元统计语言模型公式为：

$$P(D|T_1,...,T_n) \subset P(D)\prod_{i=1}^{n}(1-\lambda)P(T_i|C)+P(T_i|D)$$

在相同查询词下，在基于内容匹配条件下，返回子页面要比返回入口页面容易很多，而根据相关性原理，排列在前面的子页面具有最大的相关性，因此提出根据排列在前面的主页相关子页面来找到相应的主页。

3.网页URL去重

■ 3.1 网页去重必要性

在采集过程中我们需要判断待采集的页面是否已经采集过，需要把已采集的网页地址记录下来，组成已采集网页地址集合，当新的采集开始之前，首先判断其地址是否在已采集网页地址集合中。如在其中，表示网页已经采集，否则采集网页，把网页地址放在 URL 库中，从而避免网页的重复采集，浪费资源。为此，可以通过 URL 散列运算可以有效地对同源（同一 URL）网页进行去重，通过对其参数进行合理的调整，可以达到满意的结果。

■ 3.2 Bloom Fliter 算法

Bloom Fliter 是一种基于散列的查找算法，用于查找一个元素是否在集合中，可以对海量数据集进行表示和查找操作。Bloom Fliter 算法的基本思想为：

（1）设数据集合$A=\{a_1,a_2,\cdots,a_n\}$，含有n个元素，为待操作的集合；

（2）Bloom Fliter 用一个长度为m的位向量来表示集合中的元素，位向量V初始化全为 0；

（3）k个具有均匀分布特性的散列函数$h_1,h_2,\cdots,h_n$，

值域均为$\{1,2,\cdots,m\}$；

（4）对于元素的加入操作首先通过k个散列函数产生k个随机数$h_1,h_2,\cdots,h_n$，使位串V相应$h_1,h_2,\cdots,h_n$位均置为 1；同理，元素的查找为判定相应位是否全为 1。

4. 网页分块

需要对去重后的网页进行页面分块并对其语义块进行分析，位于页面上不同位置区域的内容往往具有不同的语义。采用 VIPS（Vision-based Page Segmentation，视觉式版面切割）算法，可以通过计算机运行一个视觉式程序来对网页区域进行分割，利用网页的布局特征，从网页的 DOM 树中抽取合适的节点，然后找出这些节点的分割符号。VIPS 算法流程示意图如图 1 所示。

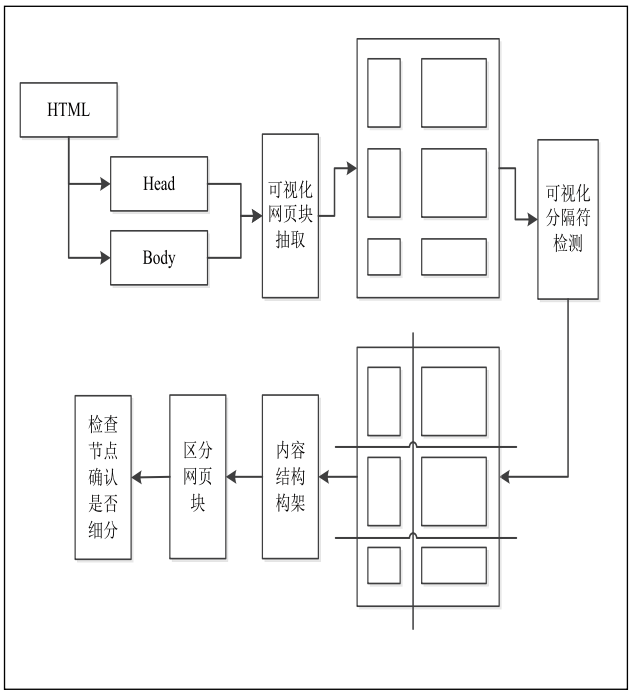

图 1 VIPS 算法流程示意图

5 .网页更新采集

由于网上的信息资源不断变化，机器人需要不断更新它所访问过的网页。因此，机器人必须判断重新访问哪些网页，以及用怎样的频率去访问。网页更新策略在很大程度上决定了网页更新的效果。

（1）对网页改变频率进行估算，计算出平均改变频率 $f'=\left(\frac{1}{n}\right)\sum_{i=1}^{n}f_i$。然后把改变频率大于等于平均改变频率的网页分为$F_1=\{f_i\mid f_i\geq f'\}$，小于平均改变频率的分为另一类$F_2=\{f_i\mid f_i<f'\}$。分别计算出其平均改变频率$F_1$和$F_2$，按照这两个平均改变频率分别访问这两类网页。

（2）假定每隔时间t访问一次网页p，一共访问n次。用$X(i)$表示在i次访问中网页是否变化过，即：$X(i)$=1（网页改变）或$X(i)$=0（网页未改变）。有网页改变总次数为 $X=\sum_{i=1}^{n}X_i$，访问的总时间$T=nt=n/f$（f是访问网页频率）。

（3）通过评测网页改变频率与访问网页频率之比率$b=f'/f$来评测网页改变频率f'。在时间t内，网页不改变的概率为：

$$p=P\left(X\left(t_{i+1}\right)-X\left(t_i\right)=0\right)=f'^{0}e^{-ft}/0!=e^{-f't}=e^{f'/f}=e^{-b}$$

即$X(i)$=0的概率为p（即e^{-b}），$X(i)$=1的概率为$1-p$（即$1-e^{-b}$）。

可得$P(X=i)=C_n^i(1-p)^i p^{n-i}$。

6 .结束语

本文分析 URL 类型和网页类型后得到其对网页信息采集的重要作用，根据两者的关联，提出从入口页面采集网页信息的方法。从抓取入口页面开始，采集网页后进行去重操作，并将网页分块提取出超链接 URL 信息。在此基础上，应用网页更新策略提高网页更新效果。最后，应用基于 URL 和网页类型的采集更新检测方法来实现网页信息采集。

参考文献

＊ [1] 胡越，张源伟，雷军．自定规则的 AJAX 网页信息采集功能的设计 [J]．物联网技术，2016，09：86-87.

＊ [2] 徐春凤，王艳春，翟宏宇．全自动网页信息采集系统 [J]．长春理工大学学报（自然科学版），2015，02：151-154.

＊ [3] 张小集，白清源．可自定规则的 Ajax 网页信息采集框架的开发 [J]．电脑开发与应用，2014，10：29-31.

＊ [4] 王娟，吴金鹏．网络爬虫的设计与实现 [J]．软件导刊，2012，04：136-137.

＊ [5] 张雷，李菁姝，马宇新．利用网页信息采集技术建立医院内网新闻平台的探讨 [J]．教育教学论坛，2013，51：198-199.

＊ [6] 胥小波，赵尔凡，康荣保．基于语义分析的互联网人物信息提取 [J]．信息安全与通信保密，2013，12：103-108.