doi:10.1631/FITEE.1400377

题目：基于 URL 聚类的快速无障碍检测抽样方法

目的：大多数残疾人士上网都会遇到各种障碍。为减少上网障碍，对网站进行无障碍检测评估是十分必要的。鉴于大部分网站具有海量网页且某些网页需涉及人工检测，通常利用抽样算法对网站进行无障碍检测评估。已有的分层抽样算法 I/O 开销和计算代价大。为解决这一问题，本文提出一种基于 URL 聚类的抽样算法。仅利用 URL 信息进行聚类，然后抽样，最终实现快速的无障碍检测和评估。

创新点：大部分网站的网页内容和 URL 信息都是由有限数量的模板生成的。因此这些网站的无障碍问题都可以追溯到模板。鉴于同一模板生成的网页具有相似结构和 URL 模式，可基于 URL 相似性对网页进行聚类，将同一模板的 URL 聚到一类中。本文所提抽样算法仅利用网页 URL 模式信息，无需存储全部网页内容，从而减少 I/O 开销和计算代价，实现快速的无障碍检测和评估。

方法：利用模板生成的网页具有相似 URL 模式，将 URL 进行聚类以实现同一模板生成的网页聚在一类中。具体过程：首先，解析爬取到的 URL 以获取候选 URL 分词和模板 URL 分词；然后利用最小长度描述原则进行 URL 聚类（算法 1）；最后在每类中按照抽样比例进行抽样。

结论：不同于现有的分层抽样算法，本文提出的抽样算法仅利用 URL 模式信息将网页进行聚类，可减少大量 I/O 开销和计算代价。

关键词：网页抽样；URL 聚类；无障碍检测