Meng-ni Zhang, Can Wang, Jia-jun Bu, Zhi Yu, Yu Zhou, Chun Chen, 2015. A sampling method based on URL clustering for fast web accessibility evaluation. *Frontiers of Information Technology & Electronic Engineering*, **16**(6):449-456. [doi:10.1631/FITEE.1400377]

A sampling method based on URL clustering for fast web accessibility evaluation

Key words: Page sampling, URL clustering, Web accessibility evaluation

Contact: Can Wang
E-mail: wcan@zju.edu.cn

ORCID: http://orcid.org/0000-0002-5890-4307

Background

- Millions of people with disabilities are affected when surfing the Web. To make the website more accessible for the disabled, it is important to evaluate the accessibility level of the whole website.

- Because of the sheer size of the modern websites and possible involvement of human judgment in evaluation, we must rely on sampling methods to reduce the cost of evaluation.

Motivation

- Most pages in a website nowadays are generated from a limited number of scripts.
  - Accessibility problems in these pages can be traced back to defects in scripts.
  - Stratified sampling methods can better locate accessibility problems in a site by clustering the pages according to their underlying templates and then drawing from each cluster.
- However, in existing stratified sampling methods, all the pages in a website need to be downloaded and analyzed for clustering, causing huge I/O and computation costs.

Main idea

- To reduce the cost in stratified sampling, we propose a novel page sampling method based on URL clustering for web accessibility evaluation, namely URLSamp.
  - By exploiting similarities in URL patterns, URLSamp clusters pages by their generating scripts and can thus effectively detect accessibility problems from webpage templates.
  - Using only the URL information in page clustering, URLSamp can efficiently scale to large websites.

The URLSamp method

- The URLSamp method consists of the following two steps (the details of the clustering algorithm are given in Algorithm 1):
  - URL parsing: parse the URLs and obtain candidate terms.
  - URL clustering and sampling: use a greedy clustering method to obtain the optimal partition, and then randomly sample from each cluster for web accessibility evaluation.

Experimental results (1)

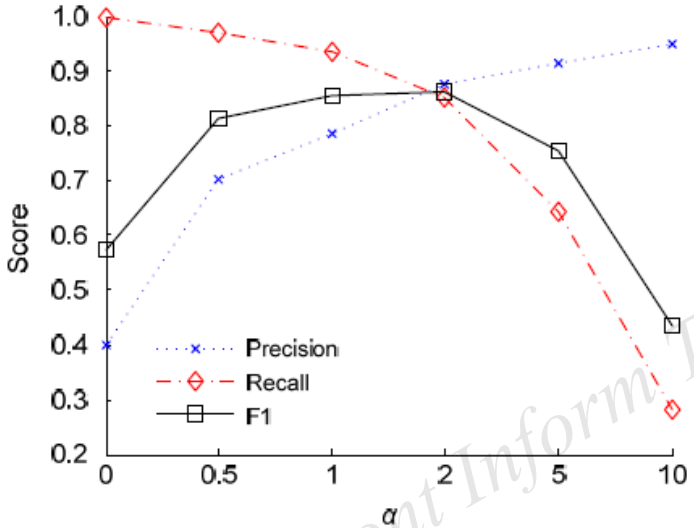

Fig. 1 Relationship between α and the clustering results
The vertical axis represents the average scores of precision and recall over the eight websites

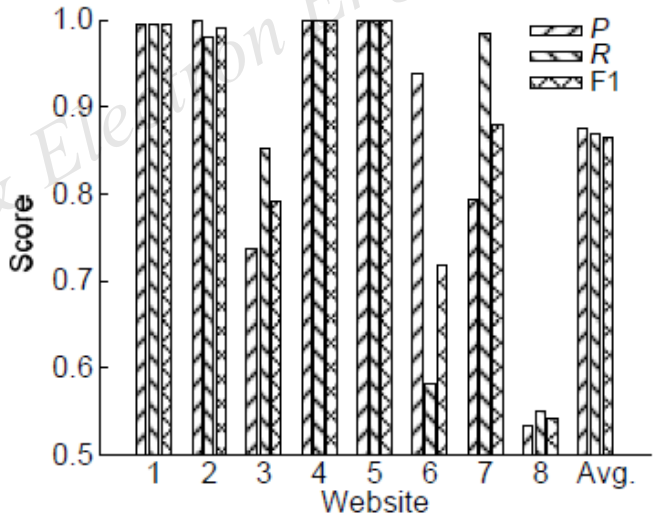

Fig. 2 Evaluation performance of clustering on the eight websites

Experimental results (2)

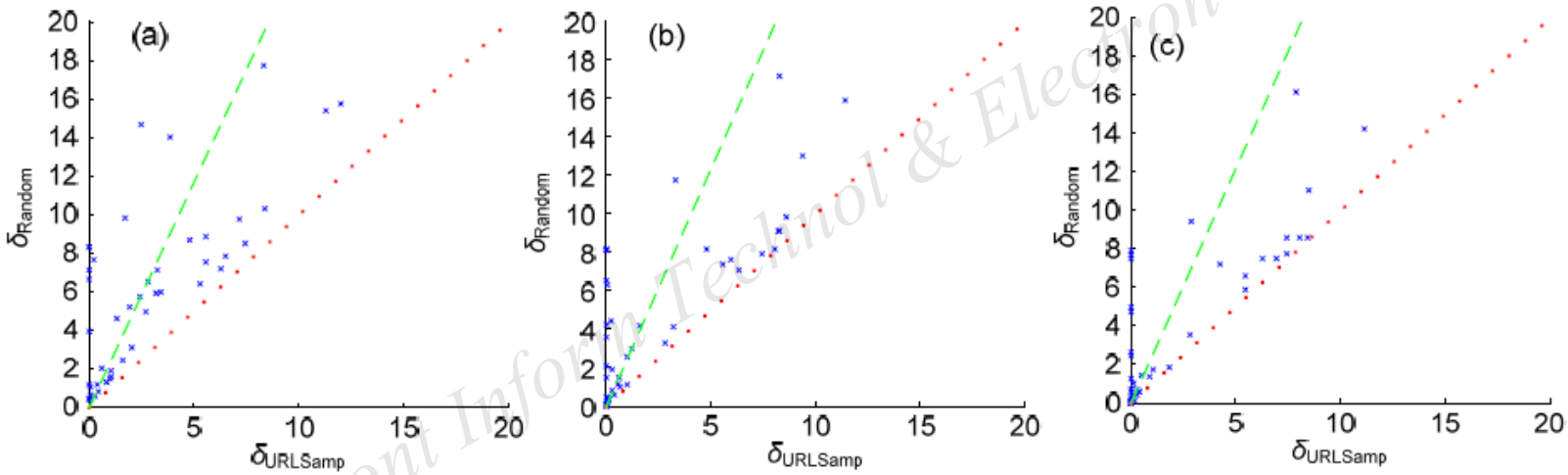

Fig. 3 Comparisons of our URLSamp algorithm and the uniform random sampling algorithm under different sampling ratios: (a) γ=0.01; (b) γ=0.05; (c) γ=0.10

The dotted bisecting line denotes $\delta_{URLSamp}=\delta_{Random}$. Each cross represents a sample error pair from the two algorithms running on a specific website. The dashed line shows the comparison of the average errors between the two sampling methods

Experimental results (3)

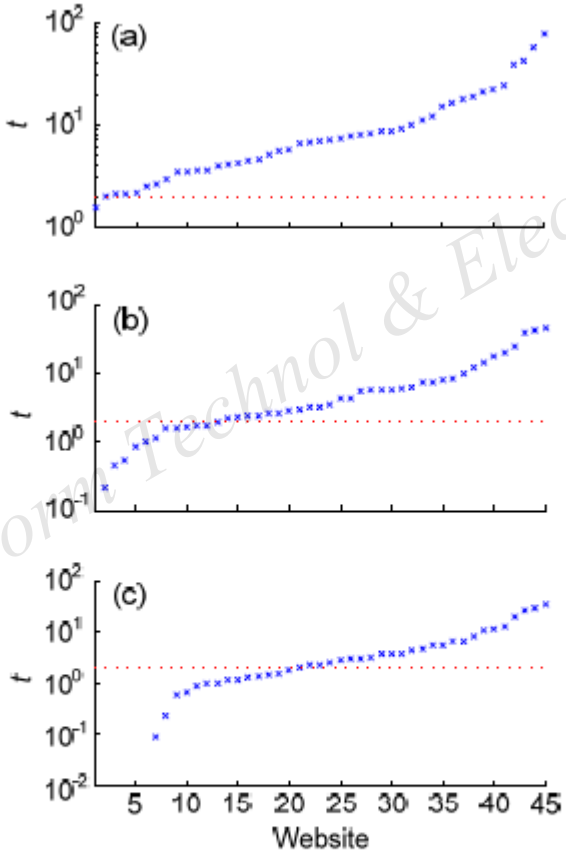

Fig. 4 A two-tailed t-test with η=0.05 under different sampling ratios: (a) γ=0.01; (b) γ=0.05; (c) γ=0.10
The dotted line corresponds to t=1.972

Conclusions

- URLSamp exploits only similarity in URL patterns to cluster webpages, thus avoiding the high cost in analyzing the huge volume of web pages.
- Meanwhile, by clustering pages according to their generating scripts, URLSamp effectively detects accessibility problems from webpage templates.
- Experimental results on real world datasets show the effectiveness of our URLSamp algorithm.