Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

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

Meng-ni ZHANG†, Can WANG†‡, Jia-jun BU, Zhi YU, Yu ZHOU, Chun CHEN

(*College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*)

†E-mail: mengnier@zju.edu.cn; wcan@zju.edu.cn

Received Nov. 2, 2014; Revision accepted Apr. 21, 2015; Crosschecked May 18, 2015

Abstract: When evaluating the accessibility of a large website, we rely on sampling methods to reduce the cost of evaluation. This may lead to a biased evaluation when the distribution of checkpoint violations in a website is skewed and the selected samples do not provide a good representation of the entire website. To improve sampling quality, stratified sampling methods first cluster web pages in a site and then draw samples from each cluster. In existing stratified sampling methods, however, all the pages in a website need to be analyzed for clustering, causing huge I/O and computation costs. To address this issue, we propose a novel page sampling method based on URL clustering for web accessibility evaluation, namely URLSamp. Using only the URL information for stratified page sampling, URLSamp can efficiently scale to large websites. Meanwhile, by exploiting similarities in URL patterns, URLSamp cluster pages by their generating scripts and can thus effectively detect accessibility problems from web page templates. We use a data set of 45 web sites to validate our method. Experimental results show that our URLSamp method is both effective and efficient for web accessibility evaluation.

Key words: Page sampling, URL clustering, Web accessibility evaluation
doi:10.1631/FITEE.1400377 **Document code:** A **CLC number:** TP391.3

1 Introduction

Website accessibility aims to help people with disabilities to perceive, understand, navigate, and interact with the web pages (Abou-Zahra, 2008). The importance of accessibility is highlighted in the guidelines published by W3C: WCAG 1.0 in 1991 and WCAG 2.0 in 2008, in which four web accessibility design principles are defined, namely being perceivable, operable, understandable, and robust. Lawsuits against violations in web accessibility followed, including the ones against America Online and Southwest Airlines in the United States, and the 2000 Sydney Olympics in Australia, etc. and promoted people's awareness about accessibility (Astbrink, 2001). With all these developments, it might be expected that accessibility for disabled users is sufficiently appreciated.

However, recent studies reported that most websites still had numerous problems which made them either difficult or nearly impossible to access for disabled users from countries worldwide (Sullivan and Matson, 2000; Disability Rights Commission, 2004a; Ellison, 2004; Marincu and McMullin, 2004; Hong *et al.*, 2008; Hanson and Richards, 2013).

Many methods have been proposed to evaluate the accessibility level of websites for disabled users, such as user testing, barrier walkthrough, conformance testing, and subjective evaluations (Disability Rights Commission, 2004b; Brajnik, 2006; 2008; Velleman *et al.*, 2006). These methods can be regarded as functions mapping the accessibility features of the website to the values assumed to determine the

‡ Corresponding author
* Project supported by the National Natural Science Foundation of China (Nos. 61173185 and 61173186) and the Natural Science Foundation of Zhejiang Province, China (No. LZ13F020001)
ORCID: Meng-ni ZHANG, http://orcid.org/0000-0002-7547-0168; Can WANG, http://orcid.org/0000-0002-5890-4307

accessibility level of pages or websites (Vigo and Brajnik, 2011).

In general, the process of measuring web accessibility consists of the following steps (Brajnik and Lomuscio, 2007): (1) collecting web pages by a web crawler; (2) evaluating the site with some accessibility evaluation methods, possibly involving human judgment; (3) combining the metrics obtained to score the site's accessibility level. However, in practical evaluations, we frequently encounter websites with a huge number of pages. Directly running evaluation on all these pages is prohibitively expensive and we have to rely on sampling methods to reduce the cost in accessibility evaluation.

Many page sampling methods have been proposed and used in web accessibility evaluation, including ad hoc sampling methods (Velleman *et al.*, 2006), random-walk based methods (Henzinger *et al.*, 2000), and stratified sampling methods. As the distribution of checkpoint violations in a website is usually highly skewed, both ad hoc sampling methods and uniform random sampling methods may lead to high bias in evaluation, if problematic pages are not detected. In contrast, stratified sampling methods usually obtain better sampling results by first clustering web pages in a site, and then draw samples from each cluster, since most accessibility problems in a page can be traced back to its generating script or template. Brajnik *et al.* (2007) compared the results of 13 sampling methods for website accessibility evaluation and found that the best sampling method is a stratified approach.

However, in the stratified sampling methods used in Brajnik *et al.* (2007), all the pages in a website have to be analyzed to compute their error profiles. This process is costly because it needs to scan the entire website, download and mirror all the Hyper Text Markup Language (HTML), Cascading Style Sheets (CSS), JavaScript, and image files, and analyze them with a web accessibility evaluation tool.

To address this issue, we propose in this paper a novel stratified page sampling method based on URL clustering, namely URLSamp. Instead of analyzing all the pages in a website, we cluster the pages by their uniform resource locators (URLs). Most pages in a website nowadays are generated from a limited number of scripts, leading to a site consisting of clusters of pages, each generated by the same script (Blanco *et al.*, 2011). Therefore, most of the accessibility problems in these sites can be traced back to defects in scripts, which form a good basis for sampling and evaluating pages. Since a page and its URL are generated by the same script simultaneously, pages with similar structures will have URLs with similar patterns and we can exploit similarities in URLs for clustering pages.

2 Related work

For websites containing a large collection of pages, it is impractical to inspect every page to evaluate the site's accessibility level. In these cases, sampling methods are the most effective measure to accelerate accessibility evaluation (Abou-Zahra, 2008). In this section, we will briefly review the various web page sampling and web accessibility evaluation methods.

The simplest among all the web page sampling methods is the ad hoc sampling suggested in Unified Web Evaluation Methodology (UWEM) (Velleman *et al.*, 2006), which selects some pre-defined pages in a site, such as the homepage and contact page. However, ad hoc sampling often requires human intervention in choosing the representative pages, which raises the cost and subjectivity in sampling.

Uniform random sampling methods (Rusmevichientong *et al.*, 2001) randomly choose pages from a website and generally ensure that each page has the same probability of being chosen at any stage during the sampling process. Although these methods seem statistically fair, they may lead to suboptimal results for accessibility evaluation. For instance, a page with image CAPTCHA will cause serious accessibility issues for visually impaired users if no alternative verification form is presented. However, a page with image CAPTCHA is unlikely to be selected by uniform random sampling methods since image CAPTCHA exists only in a very small portion of pages. Another drawback of uniform random sampling methods is that they need to collect all pages in a website for sampling, incurring a high cost in page crawling.

Henzinger *et al.* (2000) proposed sampling methods based on random walks of the web graph. Their methods consist of two phases: (1) the walk

phase, starting from the homepage and following outgoing links with a certain probability; (2) the sampling phase, pages visited during the walk phase being randomly selected with the same probability. In the EU Internet Information Accessibility (EIAO) Project (Ulltveit-Moe *et al.*, 2006), another random-walk based sampling method was adopted, in which all the links in a page on the walking path are extracted and selected with a certain probability for further crawling. Thus, all link pathways within a website are treated equally. Random-walk based methods may bias the sampling in that problematic pages far away from the seeding URL will have a small probability to be chosen.

In comparison, stratified methods can better cover pages with checkpoint violations when the distribution of violations in the website is skewed. By first clustering pages into groups and drawing pages from each group, stratified methods more efficiently select pages with accessibility issues. King *et al.* (2005) described a stratified sampling method which samples pages by their error profiles. The error profiles are represented as a vector of n components, each consisting of the number of violations of a list of n checkpoints. King *et al.* (2005) first clustered all these pages in a website. From each cluster they randomly selected pages until they obtained their sample size. Brajnik *et al.* (2007) compared the effects of 13 sampling methods and developed 9 variations of such error profile based sampling methods. The error profiles were generated by an accessibility testing tool after the testing tool tests all the pages in a website. Specifically, this process needs to download HTML, CSS, JavaScript, and image files, and mirror them on a temporary web server before the data can be analyzed by the testing tool.

There have been many surveys of website accessibility and many evaluation, validation, and repairing tools to check a site's accessibility against the best practices and guidelines. The common methods currently used to evaluate web accessibility include automated tools, design guidelines, and user studies, or combinations of each (Mankoff *et al.*, 2005). An important study on web accessibility was carried out in Sullivan and Matson (2000). They studied the content accessibility of the 50 most popular websites, and compared manually analyzed content accessibility and overall automated usability as measured by LIFT (Sullivan and Matson, 2000). Products like Bobby and LIFT are publicly automated checking tools available to validate website accessibility, and extensive studies on how to run them have been conducted (Pernice and Nielsen, 2001a; 2001b). Accessibility Commons (Kawanaka *et al.*, 2008) is a common infrastructure to integrate, store, and share metadata designed to improve web accessibility. Hanson and Richards (2004) developed a prototype system which was used to adapt web pages to meet the needs of older adult intermediaries by a proxy server.

3 Sampling based on URL clustering

In this section, we introduce our sampling method URLSamp for web accessibility evaluation. Pages and their URLs in a website nowadays are generated by a limited number of scripts, leading to a site consisting of clusters of pages, each generated by the same script (Blanco *et al.*, 2011). This has a two-fold effect in web accessibility evaluation: (1) Most accessibility issues in a website can be traced back to scripts generating these pages. Consequently, the distribution of checkpoints and the corresponding violations in a website are highly related to these scripts. (2) Since a page and its URL are simultaneously generated by the same script, pages with similar structures will have URLs with similar patterns. Therefore, it is useful to exploit the similarity in URLs to cluster pages and then sample pages from each cluster for web accessibility evaluations, which is the main idea underlying our URLSamp method. The URLSamp method consists of the following two steps:

1. URL parsing: parse the URLs and obtain candidate terms.

2. 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.

We will describe each step in detail.

3.1 URL parsing

Considering a website with pages and their URLs, we define $U=\{u_1, u_2, \ldots, u_n\}$ as a set of n URLs in this website, with u_i being the ith URL. We

follow the approach in Blanco *et al.* (2011) to split a URL into multiple tokens by character '/', with some tokens useless for clustering eliminated, such as the domains of URLs. We define $T(U)=\{t_1, t_2, \ldots, t_m\}$ as a candidate set of m terms selected from tokens obtained. Terms for URL u_i is $T(u_i)$ and obviously we have $T(u_i)\subset T(U)$.

Given a subset of URLs $U_a\subset U$ containing n_a URLs, we have

$$T(U_a)=\bigcup_{u_i\in U_a} T(u_i). \tag{1}$$

Following the approach in Blanco *et al.* (2011), we define $\mathrm{ST}(U_a)$ as the set of script terms derived from U_a, and $\mathrm{DT}(U_a)$ the set of data terms. That is,

$$\mathrm{ST}(U_a)=\bigcap_{u_i\in U_a} T(u_i), \tag{2}$$

$$\mathrm{DT}(U_a)=T(U_a)-\mathrm{ST}(U_a). \tag{3}$$

For example, the URLs in Table 1 contain eight candidate terms: $T(U)$={Bike, BikeView, 1208, 1166, News, NewsContent, 377ClassId=97, 378ClassId=97} (the domain name {www.giant.com.cn} is eliminated). For URL set $\{u_1, u_2\}$, the script terms are $\mathrm{ST}(\{u_1, u_2\})$={Bike, BikeView} and the data terms are $\mathrm{DT}(\{u_1, u_2\})$={1208, 1166}.

Table 1 URL examples

ID	URL
1	www.giant.com.cn/Bike/BikeView/1208
2	www.giant.com.cn/Bike/BikeView/1166
3	www.giant.com.cn/News/NewsContent/377ClassId=97
4	www.giant.com.cn/News/NewsContent/378ClassId=97

3.2 URL clustering and sampling

The purpose of URL clustering is to group URLs generated by the same script into the same cluster using the script terms and data terms extracted in the previous step.

Given a URL set U, a partition $\mathrm{Pa}(U)$ splits U into a limited number (k) of disjoint and nonempty subsets. That is,

$$\mathrm{Pa}(U)=\{P_1, P_2, \ldots, P_k\} \tag{4}$$

$$\text{s.t. } \forall 1\le i, j\le k,\ i\ne j,\ P_i\ne\varnothing,\ P_j\ne\varnothing,$$

$$\bigcup_{P_i\in \mathrm{Pa}(U)} P_i=U, \tag{5}$$

$$P_i\cap P_j=\varnothing. \tag{6}$$

The clustering process proceeds by iteratively partitioning U into disjoint sets X. Many criteria can be used to guide the partition, including N-cut, information entropy, and minimum description length. The minimum description length used in the URL clustering algorithm is defined as

$$\mathrm{mdl}(\mathrm{Pa}(U))=ck+\sum_i n_i\ln\frac{n}{n_i}+\alpha\sum_{P_i\in X}\mathrm{num}(\mathrm{DT}(P_i)), \tag{7}$$

where k is the number of clusters, n_i the number of URLs in cluster P_i, n the size of URL set U, $\mathrm{num}(\mathrm{DT}(P_i))$ the number of data terms in P_i, and c and α the two preset parameters. The optimal partition is achieved with the smallest mdl. The details of the clustering algorithm are given in Algorithm 1.

Algorithm 1 URL clustering

Input: URL set U; maximum number of clusters $k_{\max}$; preset parameters c, α.

Output: Clustering results $\mathrm{Pa}_{\mathrm{opt}}(U)$.

Parse all the URLs in U, and obtain the candidate term set $T(U)$;

Initial partition $\mathrm{Pa}_{\mathrm{opt}}(U)=\varnothing$, $\mathrm{Pa}_{\mathrm{tmp}}(U)=\varnothing$, and mdl-score $\mathrm{mdl}_{\mathrm{opt}}=+\infty$, $\mathrm{mdl}_{\mathrm{tmp}}=+\infty$;

for each $t_i\in T(U)$ **do**
  Find URL set $U(t_i)=\{u_j|t_i\in T(u_j)\}$;
  $\mathrm{Pa}_{\mathrm{tmp}}(U)=\{U(t_i), U-U(t_i)\}$;
  Calculate $\mathrm{mdl}_{\mathrm{tmp}}$ for $\mathrm{Pa}_{\mathrm{tmp}}(U)$ as Eq. (7);
  if $\mathrm{mdl}_{\mathrm{tmp}}<\mathrm{mdl}_{\mathrm{opt}}$ **then**
    $\mathrm{Pa}_{\mathrm{opt}}(U)=\mathrm{Pa}_{\mathrm{tmp}}(U)$;
    $\mathrm{mdl}_{\mathrm{opt}}=\mathrm{mdl}_{\mathrm{tmp}}$;
  end if
end for
Find the most frequent term t_1 in U;
for $2\le k<k_{\max}$ **do**
  Find the most frequent term t_k in $U-\bigcup_{i=1}^{k-1}U(t_i)$;
  $\mathrm{Pa}_{\mathrm{tmp}}(U)=\{U(t_1), U(t_2), \ldots, U(t_k), U-\bigcup_{i=1}^{k}U(t_i)\}$;
  Calculate $\mathrm{mdl}_{\mathrm{tmp}}$ for $\mathrm{Pa}_{\mathrm{tmp}}(U)$ as Eq. (7);
  if $\mathrm{mdl}_{\mathrm{tmp}}<\mathrm{mdl}_{\mathrm{opt}}$ **then**
    $\mathrm{Pa}_{\mathrm{opt}}(U)=\mathrm{Pa}_{\mathrm{tmp}}(U)$;
    $\mathrm{mdl}_{\mathrm{opt}}=\mathrm{mdl}_{\mathrm{tmp}}$;
  end if
end for
return $\mathrm{Pa}_{\mathrm{opt}}(U)$;

After clustering, each cluster contains the web pages generated by the similar scripts. We then draw samples from each cluster, with the number of samples from each cluster being proportional to its size. The sampling ratio γ is defined as the number of samples divided by the number of all pages.

4 Experiments and results

In this section, we evaluated the performance of our algorithms on a real world dataset. Both clustering experiments and sampling experiments were performed, with the first evaluating the effectiveness of the URL clustering algorithm and the second comparing the performances of the uniform random sampling algorithm and our URLSamp algorithm for web accessibility evaluation.

4.1 URL clustering experiments

4.1.1 Datasets

We collected data from eight websites with a total of 32279 web pages during Dec. 16–18, 2013. The detailed website information is shown in Table 2. For each website, we manually annotated the pages with cluster labels. These labeled pages are then used as the ground truth in our evaluation.

4.1.2 Evaluation metrics

In the clustering experiments, we used three metrics, precision (P), recall (R), and F1-score (F1), to quantitatively evaluate the experimental performance. The computation of the three metrics depends on the true positive (TP) decision, which assigns two similar URLs to the same cluster; the false positive (FP) decision, which assigns two dissimilar URLs to the same cluster, and the false negative (FN) decision, which assigns two similar URLs to different clusters. These three metrics are defined as

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \tag{8}$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{9}$$

$$\mathrm{F1}=\frac{2PR}{P+R}. \tag{10}$$

4.1.3 Clustering results

The parameter α in Eq. (7) controls the granularity of the partition in clustering. Good clustering results depend on the appropriate value of α. The following experiment shows the influences of different α on the clustering results. Similar to Blanco *et al.* (2011), we set the values of the other two parameters c, $k_{\max}$ to 2.0, 30, respectively, in this experiment. Fig. 1 shows the change of precision and recall with the change of α. We take the average scores over the eight websites. As expected, as α increases, precision increases monotonically, while recall decreases monotonically. The optimal value for F1 is reached when α=2.

Fig. 2 shows the precision, recall, and F1 scores of the clustering results for the eight websites. Table 2 shows more information including the number of URLs and the running time of the clustering algorithm. Excellent clustering results were seen in four

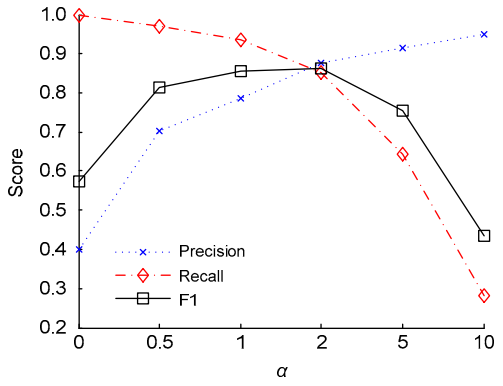

Fig. 1 Relationship between α and the clustering results
The vertical axis represents the average scores of precision and recall over the eight websites

Table 2 Evaluation performance of clustering on the eight websites

ID	Website	URL count	P	R	F1	Time (s)
1	http://www.giant.com.cn/	816	0.997	0.996	0.996	0.130
2	http://store.sonkwo.com/	663	0.999	0.982	0.990	0.110
3	http://www.scdpf.org.cn/	1453	0.738	0.851	0.790	0.480
4	http://xikou.fhnews.com.cn/	453	1	1	1	0.450
5	http://www.xiaomi.com/	6463	1	1	1	0.374
6	http://www.mca.gov.cn/	4949	0.939	0.583	0.719	1.140
7	http://www.walmartchina.com/	173	0.794	0.985	0.879	0.030
8	http://www.ituring.com.cn/	17309	0.534	0.551	0.542	3.740
	Average	4035	0.875	0.869	0.865	0.807

websites (Nos. 1, 2, 4, and 5). However, the clustering result for website 8 seems to be relatively poor. After manually checking the websites, we discovered that the pages in websites 1, 2, 4, and 5 use only a limited number of templates and consequently only a limited number of script terms are contained in the URLs. In contrast, we observed a complicated hierarchical structure in website 8, which results in many different script terms. While we set k=30, the problem can be alleviated using a larger k_{max}, which, however, will greatly deteriorate the algorithm performance.

The running time of the clustering algorithm is nearly linear in the number of URLs (Blanco *et al.*, 2011). So, it is very efficient and can scale well to a large data set.

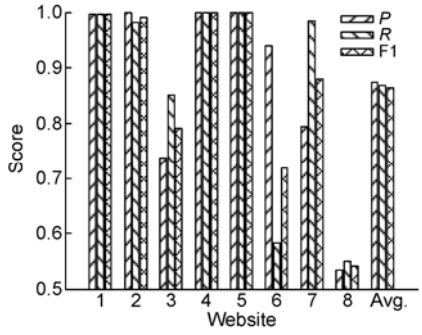

Fig. 2 Evaluation performance of clustering on the eight websites

4.2 Sampling experiments

The second experiment was to evaluate the sampling performance of our URLSamp algorithm. We start with a description of the dataset.

4.2.1 Datasets

The dataset contains 45 websites for ministries, commissions, and provincial governments in China. The dataset is the result of the China Government Website Accessibility Evaluation Campaign in 2013, which was undertaken by the China Research Center of Information and Accessibility Technology for Disability from Sept. to Oct., 2013. The web accessibility evaluation follows the official web accessibility standard in China, namely YD/T1761 2012 by the Ministry of Industry and Information Technology of the People's Republic of China. The standard is quite similar to WCAG 2.0 except that there are five conformance levels in China's standard and three conformance levels in WCAG 2.0. Three of the five conformance levels were applied in the accessibility evaluation. Both automatic evaluation and human judgment are involved in the evaluation.

4.2.2 Evaluation metrics

We adopted the evaluation scheme used in Brajnik *et al.* (2007) for our sampling experiment. Define θ as the value calculated from the entire website and θ_s the value calculated from the sample set. The sample error is described as the absolute difference between θ_s and θ:

$$\delta=|\theta_s-\theta|. \tag{11}$$

A good sampling algorithm will lead to a small δ value. For simplicity, we use $\delta_{URLSamp}$ to denote the sample error of our URLSamp algorithm and δ_{Random} the sample error of the uniform random sampling algorithm.

4.2.3 Sampling experiment results

As previously described, the three parameters in the clustering algorithm were set as α=2.0, c=2.0, and k_{max}=30. The sample errors for the two sampling algorithms were averaged over 100 replications. The two sampling algorithms were compared under different sampling ratios, γ=0.01, 0.05, and 0.10.

The results are shown in Fig. 3. Obviously, all the 45 crosses are located above the dotted bisecting line; i.e., $\delta_{URLSamp}$ is smaller than δ_{Random}. The slope of the dashed line is the ratio between the average errors for the uniform random sampling and URLSamp. It can be seen that our URLSamp algorithm significantly outperforms the uniform random sampling algorithm in web accessibility evaluation.

To explore the relationship between sampling ratio γ and sample error δ, we used $\overline{\delta}_{URLSamp}$ to denote the average sample error in 45 websites for our URLSamp algorithm and $\overline{\delta}_{Random}$ the average for the uniform random sampling algorithm.

The average sample errors in 45 websites for two algorithms under different sampling ratios γ are shown in Table 3. URLSamp has much smaller average sample errors than uniform random sampling under different γ values. However, the difference

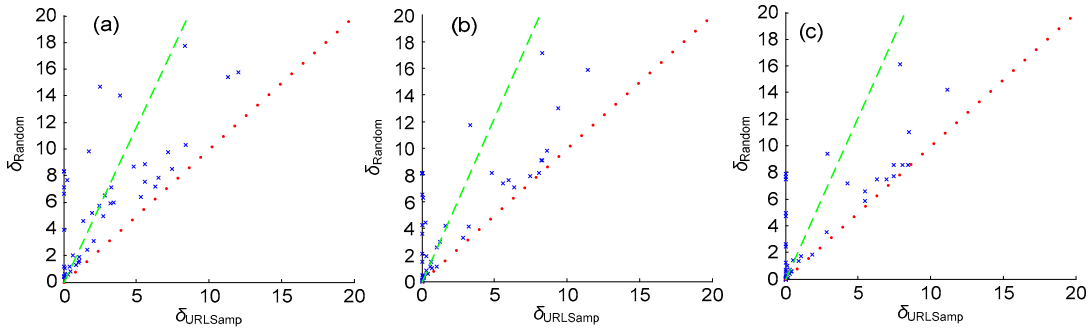

Fig. 3 Comparisons of our URLSamp algorithm and the uniform random sampling algorithm under different sampling ratios: (a) γ=0.01; (b) γ=0.05; (c) γ=0.10

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

Table 3 Average errors of URLSamp and uniform random sampling

γ	$\bar{\delta}_{URLSamp}$	$\bar{\delta}_{Random}$	$\bar{\delta}_{Random}-\bar{\delta}_{URLSamp}$
0.01	2.852	6.020	3.168
0.05	2.455	4.717	2.262
0.10	2.208	3.841	1.633

between the two algorithms decreases steadily as the sampling ratio γ increases.

4.2.4 *t*-test

We used a two-tailed *t*-test with η=0.05 (significance level) to verify that URLSamp is statistically superior to uniform random sampling (Fig. 4). The significance level (*t*-value) is defined in Eq. (12). All results were averaged over 100 test runs.

$$t=(\bar{\delta}_{Random}-\bar{\delta}_{URLSamp})\Big/\sqrt{\frac{s_{Random}^2}{n_{Random}}+\frac{s_{URLSamp}^2}{n_{URLSamp}}}, \quad (12)$$

where $\bar{\delta}_{URLSamp}$ and $\bar{\delta}_{Random}$ denote the average sample errors of the two algorithms in a website of over 100 test runs, s_{Random} and $s_{URLSamp}$ denote the unbiased estimators of the variance of the average sample error, and $n_{Random}=n_{URLSamp}=100$. Thus, when $t>1.972$ and η=0.05, the improvement of our algorithm was regarded as statistically significant. When the sampling ratio γ=0.01, almost all the crosses were above the dotted line, indicating that our algorithm achieved a significant improvement in most of the websites. When γ=0.10, statistically significant

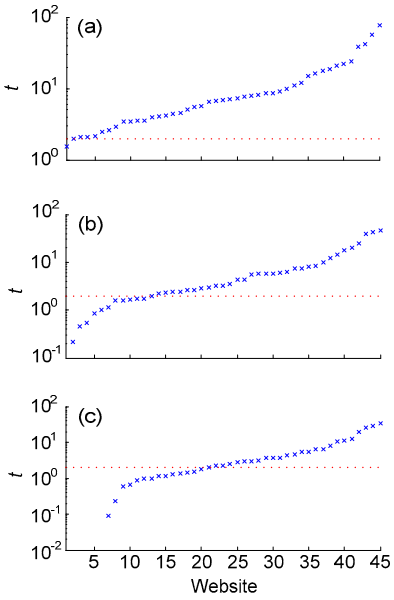

Fig. 4 A two-tailed *t*-test with η=0.05 under different sampling ratios: (a) γ=0.01; (b) γ=0.05; (c) γ=0.10

The dotted line corresponds to t=1.972

improvements were witnessed in about half of the websites.

5 Conclusions and future work

In this paper, we present a novel web page sampling method based on URL clustering, URLSamp, for web accessibility evaluation. Different from existing stratified sampling methods, URLSamp exploits only similarity in URL patterns to cluster web pages, thus avoiding the high cost in analyzing the

huge volume of web pages. Experimental results on real world datasets show the effectiveness of our URLSamp algorithm.

There are several interesting problems to be investigated in our future work: (1) Our sampling method shows relatively poor performance in websites with complicated structures. It would be interesting to try more sophisticated URL clustering algorithms that can better handle complicated structures. (2) Content in a website nowadays is highly dynamic in nature. Sampling from the incremental content instead of the whole website will significantly improve the efficiency of accessibility evaluation. We will explore incremental sampling for accessibility evaluation in the future.

References

Abou-Zahra, S., 2008. Web accessibility evaluation. *In*: Harper, S., Yesilada, Y. (Eds.), Web Accessibility: a Foundation for Research. Springer, London, p.79-106.

Astbrink, G., 2001. The legislative impact in Australia on universal access in telecommunications. Proc. Universal Access in Human-Computer Interaction Conf., p.1042-1046.

Blanco, L., Dalvi, N., Machanavajjhala, A., 2011. Highly efficient algorithms for structural clustering of large websites. Proc. 20th Int. Conf. on World Wide Web, p.437-446. [doi:10.1145/1963405.1963468]

Brajnik, G., 2006. An Accessibility Evaluation Method Based on Barrier Walkthrough. Available from http://www.dimi.uniud.it/giorgio/projects/bw.

Brajnik, G., 2008. A comparative test of web accessibility evaluation methods. Proc. 10th Int. ACM SIGACCESS Conf. on Computers and Accessibility, p.113-120. [doi:10.1145/1414471.1414494]

Brajnik, G., Lomuscio, R., 2007. SAMBA: a semi-automatic method for measuring barriers of accessibility. Proc. 9th Int. ACM SIGACCESS Conf. on Computers and Accessibility, p.43-50. [doi:10.1145/1296843.1296853]

Brajnik, G., Mulas, A., Pitton, C., 2007. Effects of sampling methods on web accessibility evaluations. Proc. 9th Int. ACM SIGACCESS Conf. on Computers and Accessibility, p.59-66. [doi:10.1145/1296843.1296855]

Disability Rights Commission, 2004a. Formal Investigation Report: Web Accessibility.

Disability Rights Commission, 2004b. The Web: Access and Inclusion for Disabled People—a Formal Investigation. The Stationery Office, UK.

Ellison, J., 2004. Assessing the accessibility of fifty United States government web pages: using Bobby to check on Uncle Sam. *First Monday*, **9**(7). [doi:10.5210/fm.v9i7.1161]

Hanson, V.L., Richards, J.T., 2004. A web accessibility service: update and findings. Proc. 6th Int. ACM SIGACCESS Conf. on Computers and Accessibility, p.169-176. [doi:10.1145/1028630.1028661]

Hanson, V.L., Richards, J.T., 2013. Progress on website accessibility. *ACM Trans. Web*, **7**(1), Article 2. [doi:10.1145/2435215.2435217]

Henzinger, M.R., Heydon, A., Mitzenmacher, M., *et al.*, 2000. On near-uniform URL sampling. *Comput. Netw.*, **33**(1-6):295-308. [doi:10.1016/S1389-1286(00)00055-4]

Hong, S., Katerattanakul, P., Joo, S.J., 2008. Evaluating government website accessibility: a comparative study. *Int. J. Inform. Technol. Dec. Mak.*, **7**(3):491-515. [doi:10.1142/S0219622008003058]

Kawanaka, S., Borodin, Y., Bigham, J.P., *et al.*, 2008. Accessibility commons: a metadata infrastructure for web accessibility. Proc. 10th Int. ACM SIGACCESS Conf. on Computers and Accessibility, p.153-160. [doi:10.1145/1414471.1414500]

King, M., Thatcher, J.W., Bronstad, P.M., *et al.*, 2005. Managing usability for people with disabilities in a large web presence. *IBM Syst. J.*, **44**(3):519-535. [doi:10.1147/sj.443.0519]

Mankoff, J., Fait, H., Tran, T., 2005. Is your web page accessible?: a comparative study of methods for assessing web page accessibility for the blind. Proc. SIGCHI Conf. on Human Factors in Computing Systems, p.41-50. [doi:10.1145/1054972.1054979]

Marincu, C., McMullin, B., 2004. A comparative assessment of web accessibility and technical standards conformance in four EU states. *First Monday*, **9**(7). [doi:10.5210/fm.v9i7.1160]

Pernice, K., Nielsen, J., 2001a. Beyond ALT Text: Making the Web Easy to Use for Users with Disabilities. Technical Report, Nielsen Norman Group, USA.

Pernice, K., Nielsen, J., 2001b. How to Conduct Usability Studies for Accessibility. Technical Report, Nielsen Norman Group, USA.

Rusmevichientong, P., Pennock, D.M., Lawrence, S., *et al.*, 2001. Methods for sampling pages uniformly from the World Wide Web. Proc. AAAI Fall Symp. on Using Uncertainty within Computation, p.121-128.

Sullivan, T., Matson, R., 2000. Barriers to use: usability and content accessibility on the web's most popular sites. Proc. ACM Conf. on Universal Usability, p.139-144. [doi:10.1145/355460.355549]

Ulltveit-Moe, N., Snaprud, M., Nietzio, A., *et al.*, 2006. Early Results from Automatic Accessibility Benchmarking of Public European Web Sites from the European Internet Accessibility Observatory (EIAO). Available from http://mortengoodwin.net/publicationfiles/dfa2006.pdf.

Velleman, E., Velasco, C., Snaprud, M., *et al.*, 2006. D-WAB4 Unified Web Evaluation Methodology (UWEM 1.0). Technical Report, WAB Cluster.

Vigo, M., Brajnik, G., 2011. Automatic web accessibility metrics: where we are and where we can go. *Interact. Comput.*, **23**(2):137-155. [doi: 10.1016/j.intcom.2011.01.001]