Which webpage should we crawl first? Social media-based webpage source importance guidance

Mohamed Farag
Virginia Tech, Blacksburg, VA 24061
Arab Academy for Science and Technology, Alexandria, Egypt
mmagdy@vt.edu, mmagdy@aast.edu

Edward A. Fox
Virginia Tech
Blacksburg, VA 24061
fox@vt.edu

Social media has proven to be an important and rich asset for collecting webpages about events. Ensuring full Web archive coverage of an event is not an easy task, for several reasons. First, events differ in impact and importance. Big events tend to last for a long time, impact multiple places, and even spark a range of debates about diverse topics. Second, to build a Web collection that fully covers an event requires sampling an unbiased set of webpages from the WWW (which is huge, heterogeneous, and dynamically changing). The size of the WWW makes difficult finding and unbiased set of webpages by manual techniques for collecting, curating, and sampling. Fortunately, focused crawlers have proven effective in automating and accelerating the process of collecting webpages, starting from a set of seed URLs. However, the ability of the focused crawler to find relevant and diverse webpages depends on the quality (content quality and linking structure quality) and the broad coverage (seed URLs from different webpage sources and publishing venues/genres) of the seed URLs.

In IDEAL project, we have been researching building webpage/tweet collections about events. Our three main approaches have been: 1) the Internet Archive's Archive-It service for archiving webpages, 2) a pair of archiving tools for collecting tweets, and 3) event model-based focused crawling of webpages.

We propose a hybrid approach for building unbiased collections of webpages with high coverage using seed URLs generated from social media content (tweets), together with event model-based focused crawlers. The tweet collection processes ensure a large sample of seed URLs with broad and heterogeneous genres of webpages (horizontal/exploring aspect) while the event model-based focused crawler ensures high quality and relevant webpages (vertical/exploiting aspect). Although tweet collections about events are very rich source of seed URLs, they contain a lot of noise (porn, job marketing, other spam, and varied other types of non-relevant tweets or URLs).

One important step required before using URLs extracted from tweets is an importance analysis of each URL/webpage source, e.g., considering the domain name of the URL.

For example, the source of the URL http://www.cnn.com /philadelphia-amtrak-derailment/ is www.cnn.com.

The importance of a source can be estimated by how many relevant webpages about the event are from that source. The webpage source analysis has two benefits: first it helps order the sources, and therefore the URLs from these sources, according to their importance to the event. This ordering of URLs helps the focused crawler focus on the part of the Web graph that is expected to have more relevant webpages. Second, this ordering helps in filtering out sources of noisy / non-relevant content.

Table 1 Harvest ratio for event focused crawler using two methods of seed selection with different numbers of seeds

	K = 10	K = 50	K = 100
Top K Frequent unique websites	0.685	0.752	0.817
Top K Frequent websites with Redundancy	0.645	0.763	0.775

We ran our experiments about Brussels Attack event. We ran our event focused crawler to collect 1000 webpages starting from different sets of seed URLs. The sets of seed URLs tested in our experiments differ in two aspects: the number of URLs and the uniqueness of the URLs. We selected the seed URLs from a pool of URLs extracted from a set of tweets collect using twitter streaming API. We used the harvest ratio measure to evaluate the output of the focused crawlers. Table 1 summarizes the results of the different settings of the seed URLs. As results shows, as we increase the number of seeds, the more relevant webpages (high harvest ratio) the focused crawler can find. Also distributing the seed URLs across several websites increases the ability of the focused crawler to find more relevant webpages.

ACKNOWLEDGMENTS

This material is based upon work supported by the US National Science Foundation under Grant No. IIS-1319578.

1. REFERENCES

[1] Heritrix, http://www.crawler.archive.org/index.html accessed on 12/26/2015

[2] Archive-It, https://archive-it.org/, accessed on 12/26/2015

[3] Edward A. Fox, Donald Shoemaker, Andrea Kavanaugh, Steven Sheetz, Jefferson Bailey, Mohamed Farag, Sunshin Lee. Integrated Digital Events Archiving and Library (IDEAL), http://eventsarchive.org, accessed on 1/24/2016

[4] Sklearn, http://scikit-learn.org/stable/, accessed on 10/12/2015

[5] Python, https://www.python.org/, accessed on 10/12/2015.

[6] Readability, https://github.com/buriy/python-readability, accessed on 03/28/2016.