Web Crawlers and Privacy: The Need to Reboot Robots.txt

Arvind Narayanan, Stanford University
Pete Warden, Mailana Inc.

Privacy norms, rules and expectations in the real world go far beyond the "public/private" dichotomy. Yet in the realm of web crawler access control, we are tied to this binary model via the robots.txt allow/deny rules. This position paper describes some of the resulting problems and argues that it is time for a more sophisticated standard.

The problem: privacy of public data. The first author has argued that individuals often expect privacy constraints on data that is publicly accessible on the web. Some examples of such constraints relevant to the web-crawler context are:

- Data should not be archived beyond a certain period (or at all).
- Crawling a small number of pages is allowed, but large-scale aggregation is not.
- "Linkage" of personal information to other databases is prohibited.

Currently there is no way to specify such restrictions in a machine-readable form. As as result, sites resort to hacks such as identifying and blocking crawlers whose behavior they don't like, without clearly defining acceptable behavior. Other sites specify restrictions in the Terms of Service and bring legal action against violators. This is clearly not a viable solution — for operators of web-scale crawlers, manually interpreting and encoding the ToS restrictions of every site is prohibitively expensive.

There are two reasons why the problem has become pressing: first, there is an ever-increasing quantity of behavioral data about users that is valuable to marketers — in fact, there is even a black market for this data — and second, crawlers have become very cheap to set up and operate.

The desire for control over web content is by no means limited to user privacy concerns. Publishers concerned about copyright are equally in search of a better mechanism for specifying fine-grained restrictions on the collection, storage and dissemination of web content. Many site owners would also like to limit the acceptable uses of data for competitive reasons.

The solution space. Broadly, there are three levels at which access/usage rules may be specified: site-level, page-level and DOM element-level. Robots.txt is an example of a site-level mechanism, and one possible solution is to extend robots.txt. A disadvantage of this approach, however, is that the file may grow too large, especially in sites with user-generated content what may wish to specify per-user policies.

A page-level mechanism thus sounds much more suitable. While there is already a "robots" attribute to the META tag, it is part of the robots.txt specification and has the same limitations on functionality. A different META tag is probably an ideal place for a new standard.

Taking it one step further, tagging at the DOM element-level using microformats to delineate personal information has also been proposed. A possible disadvantage of this approach is the overhead of parsing pages that crawlers will have to incur in order to be compliant.

Conclusion. While the need to move beyond the current robots.txt model is apparent, it is not yet clear what should replace it. The challenge in developing a new standard lies in accommodating the diverse requirements of website operators and precisely defining the semantics of each type of constraint without making it too cumbersome to write a compliant crawler. In parallel with this effort, the development of legal doctrine under which the standard is more easily enforceable is likely to prove invaluable.