New-Web search with microblog annotations

Tom Rowlands, Schol. of Computer Science, ANU and CSIRO ICT Centre, David Hawking, Funnelback Pty. Ltd., and Ramesh Sankaranarayana, SoCS, ANU

Web search engines collect documents by crawling, but this is resource intensive and biased toward documents that are older.

We describe a novel system for ‘new-Web’ search using ‘microblog annotations’. Twitter messages containing a URL are used as a description of the URL’s target. As Twitter is often used for the discussion of recent events, these messages may offer unique and timely evidence.

Our system is uncomplicated and may offer a significant advantage in revealing new Web documents that would otherwise be hidden from searchers.

Pages that have been on the Web for some time are likely to have links pointing to them, guiding Web crawlers and increasing query-independent evidence for their relevance. New pages, which we assume to be of particular interest for recent-event searches, are likely to have few, if any, incoming links, making them harder or impossible for a Web crawler to find.

Microblogging evidence

Twitter is a microblogging service where many people publicly discuss ‘What’s happening now?’ from their perspective. Millions of messages—‘tweets’ in Twitter parlance—each day contain information about what people are doing, where they are and their views on the topics of the moment. It is not uncommon for the message to be a comment on a new document, pointed to by a URL. Therefore, tweets contain both links to otherwise hard-to-find pages and descriptions of those pages. For example, in Figure 1, the user is commenting on a recently transcribed speech.

Re-tweets and social network characteristics of Twitter may provide further evidence, but these are not demonstrated or investigated in this work.

Static investigation

For our initial investigation, we collected a large number of tweets containing the text ‘http’. URLs were extracted using a regular expression and downloaded using a standard Web crawler (with a crawl depth of zero).

This investigation revealed that URLs specified in tweets are dominated by redirections. Users can use ‘shortening services’ to make more efficient use of the one hundred and forty permitted characters in a tweet. Short URLs may also offer less chance of corruption.

Live investigation

We have constructed a simple system as a demonstration.

The Twitter Streaming ‘filter’ API is used to gather those tweets containing the text ‘http’. These are recorded on disc and the associated URLs queued for a dissemination process. The disseminator distributes URLs to queues by hostname. Each host’s queue is serviced by a separate process, which downloads URLs in turn. In the event of a redirection, which is quite common, the new URL is inserted back in the disseminators queue, if it is not a duplicate. The tweets are indexed constantly; as soon as one indexing process finishes, it is declared the live index and a new indexing process begins. Older events are less interesting, so only a small window of time needs to be considered for each indexing run.

Future research will include a robust evaluation for recent document search and the investigation of other options for microblog evidence in Web search.

Figure 1: The content of the tweet is a description of the Web page and allows us to discover the Web page more quickly than crawling.