Abstract

Mohamed Magdy

Intelligent Event Focused Crawling

There is need for an integrated event focused crawling system to collect Web data about key events. When an event occurs, many users try to locate the most up-to-date information about that event. Yet, there is little systematic collecting and archiving anywhere of information about events. We propose intelligent event focused crawling for automatic event tracking and archiving, as well as effective access. We extend the traditional focused (topical) crawling techniques in two directions, modeling and representing: events and webpage source importance. We developed an event model that can capture key event information (topical, spatial, and temporal). We incorporated that model into the focused crawler algorithm. For the focused crawler to leverage the event model in predicting a webpage's relevance, we developed a function that measures the similarity between two event representations, based on textual content. Although the textual content provides a rich set of features, we proposed an additional source of evidence that allows the focused crawler to better estimate the importance of a webpage by considering its website. We estimated webpage source importance by the ratio of number of relevant webpages to non-relevant webpages found during crawling a website. We combined the textual content information and source importance into a single relevance score. For the focused crawler to work well, it needs a diverse set of high quality seed URLs (URLs of relevant webpages that link to other relevant webpages). Although manual curation of seed URLs guarantees quality, it requires exhaustive manual labor. We proposed an automated approach for curating seed URLs using social media content. We leveraged the richness of social media content about events to extract URLs that can be used as seed URLs for further focused crawling. We evaluated our system through four series of experiments, using recent events: Orlando shooting, Ecuador earthquake, Panama papers, California shooting, Brussels attack, Paris attack, and Oregon shooting. In the first experiment series our proposed event model representation, used to predict webpage relevance, outperformed the topic-only approach, showing better results in precision, recall, and F1-score. In the second series, using harvest ratio to measure ability to collect relevant webpages, our event model-based focused crawler outperformed the state-of-the-art focused crawler (best-first search). The third series evaluated the effectiveness of our proposed webpage source importance for collecting more relevant webpages. The focused crawler with webpage source importance managed to collect roughly the same number of relevant webpages as the focused crawler without webpage source importance, but from a smaller set of sources. The fourth series provides guidance to archivists regarding the effectiveness of curating seed URLs from social media content (tweets) using different methods of Selection.