Event Aware Focused Crawler

IDEAL Project Case Study

Ward Bonnefond, Chris Menzel, Zack Morris, Suhas Patel, Tyler Ritchie, Mark Tedesco, Franklin Zheng

Client: Mohamed Magdy Gharib Farag

ABSTRACT:
Currently the IDEAL (Integrated Digital Event Archive and Library) project has a general purpose Web crawler to collect articles relevant to seed URLs based on keyword frequency analysis. Our goal is to create a focused crawler tailored to finding events. By analyzing an article to identify key event components, we can construct a tree representation of a webpage. By using the tree edit distance between that tree, and the event tree from seed information, we can predict webpage relevance. Thus, our implementation improves upon general purpose Web crawlers by replacing the “keyword matching” phase with our tree edit distance algorithm. This works since we are able to create a fairly accurate description of an event from a webpage, which can yield more specific information than would keyword analysis alone.

Design Goals:
- More user friendly
- More specific results

Design Decisions
- Web Interface for maximum portability
- Tree Edit Distance clearly shows why an article is accepted
- Focused on disasters for now

IDEAL Project:
The goal of the IDEAL project is to collect, catalog, preserve, and provide access to services related to digital objects about events, including multimedia -- captured from all corners of the Web. Using Web and Tweet collections, the IDEAL project hopes to be able to create unbiased archives for events of interest.