Developing an improved focused crawler for the IDEAL project

Client: Mohamed Magdy Gharib Farag

Presenters: Ward Bonnefond, Chris Menzel, Zack Morris, Suhas Patel, Tyler Ritchie, Mark Tedesco, Franklin Zheng
IDEAL project

- Integrating Digital Event Archiving and Library
- Finding webpages related to an event (i.e., natural disaster)
- Store found webpages locally for parsing and analysis
Enhanced focus crawler

- Extract key words and key concepts (i.e., date, location, type of disaster)
- Construct trees based on these words and concepts
- Develop algorithm to compare different trees and their relationships
- Make the above process accessible via a PHP web application, so users can visually see the results generated by the focused crawler
Project components

1. Tree construction and visual representation
2. Represent a webpage (an “event”) in terms of keywords and key concepts (an “event representation”)
3. Integrating updated modules into the existing focused crawler
Original Implementation

1. Start with a list of seed URLs
2. Web-crawler crawls through list of URLs
3. Outputs a score for each URL based on keyword matchings
4. Searches the webpage for other URLs
5. Adds any good URLs found to the list
Current Progress

- Front-End
  - User can enter multiple seed URLs into a textbox and submit them to Python bundle
  - Python bundle returns scored webpages, which are then displayed on the front-end webpage

- Back-end
  - Halfway through creating an event tree from online articles
  - Type of storm can be retrieved from the title of an article
Future Work

- Finish producing the event-tree
- Compare it with the tree provided by user to estimate the relevancy of an article
- Make the GUI for displaying the event-tree for a specific event
- Finish the UI for the webpage, to make it easier to show the back-end data to users
Projected Implementation

1. Start with a list of seed URLs
2. Web-crawler crawls through list of URLs
3. Outputs a score for each URL based on tree-edit distance
4. Searches the webpage for other URLs
5. Adds any good URLs found to the list
Current Back-End Example

[u'typhoon', u'philippin', u'haiyan', u'storm', u'2013', u'people', u'city', u'home', u'tacloban', u'said']

1.0.55161008187, http://www.redcross.org.ph
1.0.612578121633, https://news.google.com/news?nc1=dzmbwHxDMHjAwBMDz81Hqqb6idqHM&aq=typhoon+haiyan&lr=English&hl=en
10
10
Current Front-End Example

Focused Event Crawler

Site URL:  
Site URL:  

+ URL Entry  Submit Entries  Clear Fields

Results

There have been no current submissions.
Questions?