Developing an improved focused crawler for the IDEAL project

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

Refresher - 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

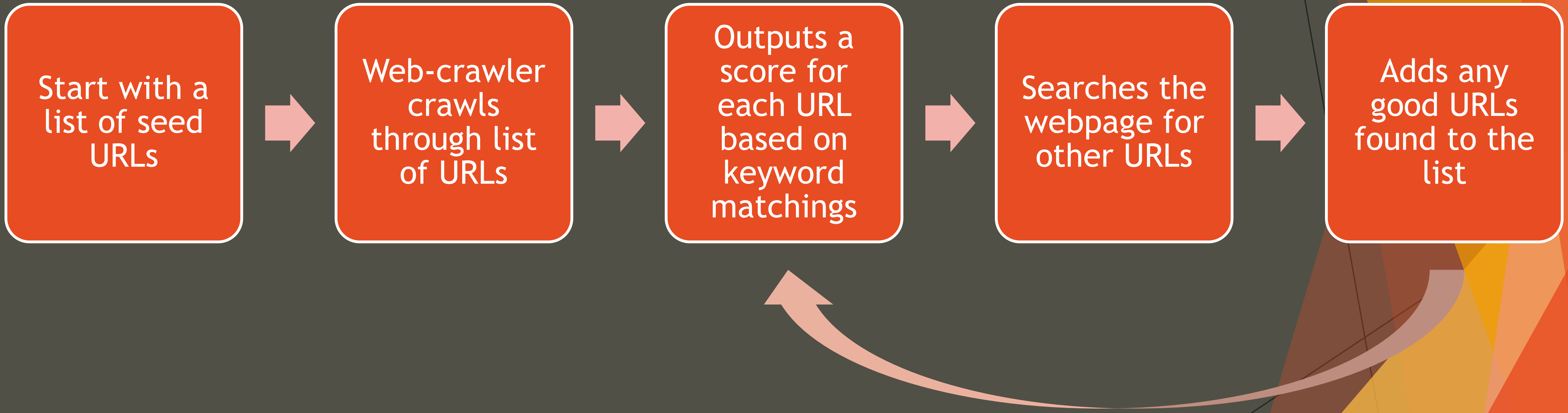
Refresher - Enhanced focus crawler

- 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 this process accessible via a web application

Refresher - Project components

1. Tree construction and visual representation
2. Event representation (i.e. key words and key concepts) versus actual event (i.e. webpage)
3. Integrating updated modules into the existing focused crawler

Refresher - Original Implementation



Start with a list of seed URLs

Web-crawler crawls through list of URLs

Outputs a score for each URL based on keyword matchings

Searches the webpage for other URLs

Adds any good URLs found to the list

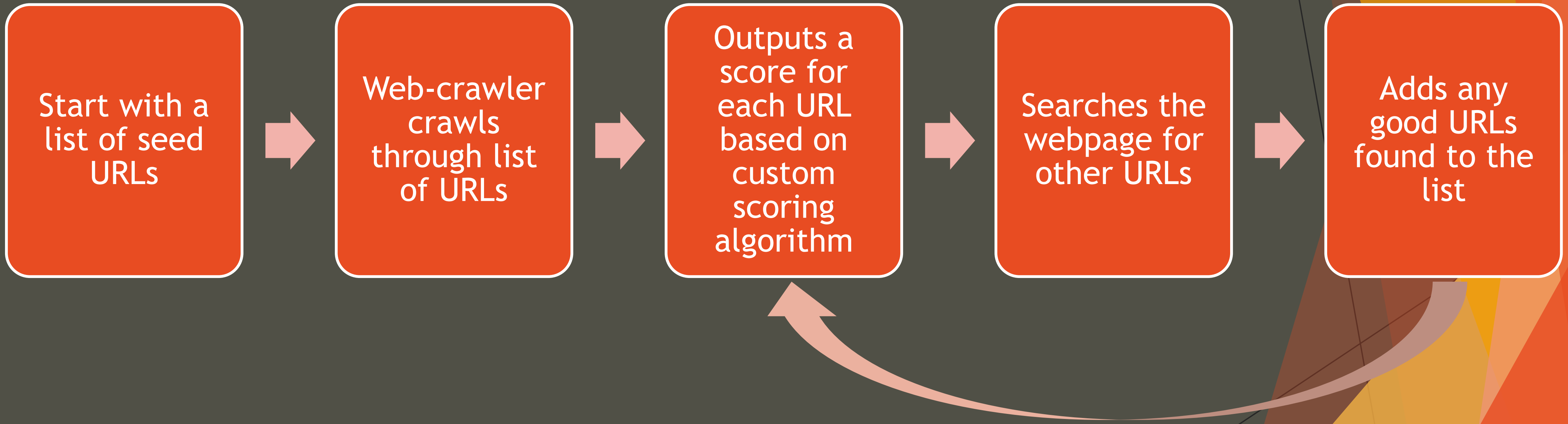
Final Results

- Front-End
  - User can enter multiple seed URLS and other important attributes (date, location, type of disaster, etc.)
  - Constructs a visual tree representation of the articles found via the web application
  - Loads back-end results using JavaScript (previously PHP)
- Back-end
  - Constructs a full event tree from query created by the user
  - Intelligently extracts location, date, type of disaster, etc. using natural language processing as opposed to keyword analysis

Scoring Algorithm

- Date: 20%
  - Year (365) + Month (30) + Day (1)
  - i.e. Year has 365 times the weight of a Day
- Location, Type of Event, Event Name, etc.: 80%
  - Replaces the previously used tree-edit distance
- New scoring algorithm provides more accurate articles that using only tree-edit distance did not find

Final Implementation



Start with a list of seed URLs

Web-crawler crawls through list of URLs

Outputs a score for each URL based on custom scoring algorithm

Searches the webpage for other URLs

Adds any good URLs found to the list

Current Back End Example Data

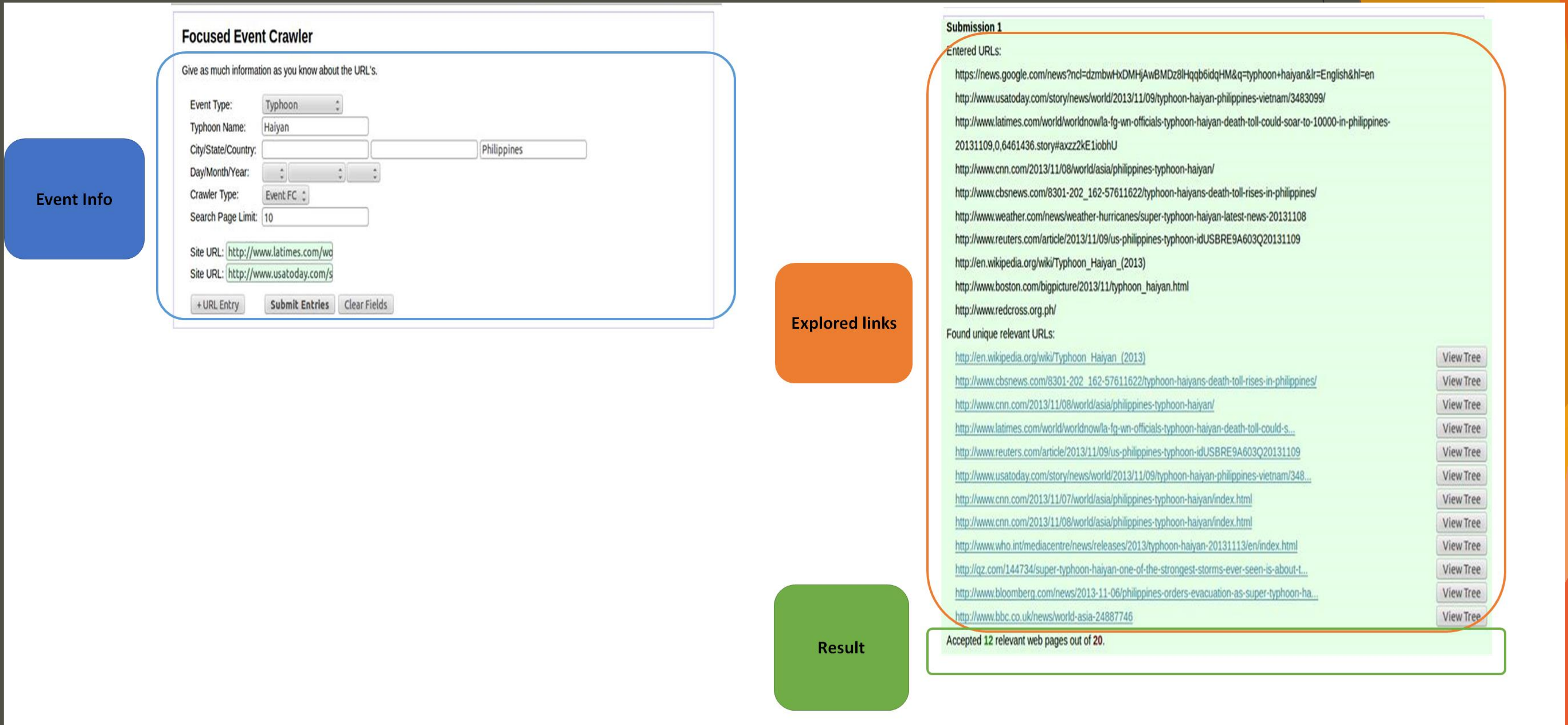
- **Base Focused Crawler**
- 0.479449956362|0.555185525105|http://www.tmz.com/2014/04/27/donald-sterling-photos-magic-johnson-matt-kemp-v-stiviano-racist-audio/2/0.479449956362|0.555185525105|http://www.tmz.com/2014/04/27/donald-sterling-photos-magic-johnson-matt-kemp-v-stiviano-racist-audio/1/0.479449956362|0.555185525105|http://www.tmz.com/2014/04/27/donald-sterling-photos-magic-johnson-matt-kemp-v-stiviano-racist-audio/20/0.479449956362|0.596680123725|http://www.tmz.com/2014/04/27/donald-sterling-photos-magic-johnson-matt-kemp-v-stiviano-racist-audio/14/0.479449956362|0.614986314544|http://www.tmz.com/2014/04/27/donald-sterling-photos-magic-johnson-matt-kemp-v-stiviano-racist-audio/11/0.479449956362|0.603562290076|http://www.tmz.com/2014/04/27/donald-sterling-photos-magic-johnson-matt-kemp-v-stiviano-racist-audio/10/0.479449956362|0.555185525105|http://www.tmz.com/2014/04/27/donald-sterling-photos-magic-johnson-matt-kemp-v-stiviano-racist-audio/12/0.479449956362|0.682182274472|http://www.tmz.com/2014/04/27/donald-sterling-photos-magic-johnson-matt-kemp-v-stiviano-racist-audio/13/0.479449956362|0.621091837905|http://www.tmz.com/2014/04/27/donald-sterling-photos-magic-johnson-matt-kemp-v-stiviano-racist-audio/15/0.479449956362|0.588410106633|http://www.tmz.com/2014/04/27/donald-sterling-photos-magic-johnson-matt-kemp-v-stiviano-racist-audio/16/
- Visited: 100Accepted: 81

Extended Focused Crawler

0.469920800878|0.52|http://www.dailymail.co.uk/tvshowbiz/article-2511646/Kim-Kardashian-eBays-old-clothes-help-Philippines-typhoon-victims--donates-just-10-proceeds-charity.html0.469920800878|1.0|http://www.fao.org/emergencies/crisis/philippines-typhoon-haiyan/en/0.469920800878|0.901299136822|http://www.fao.org/emergencies/crisis/philippines-typhoon-haiyan/crop-damages-map/en/0.469920800878|0.919740571108|http://www.fao.org/emergencies/crisis/philippines-typhoon-haiyan/impact-assessment-map/en/0.469920800878|0.855014256832|http://www.fao.org/emergencies/crisis/philippines-typhoon-haiyan/input-distribution-map/en/0.469920800878|1.0|http://www.fao.org/emergencies/crisis/philippines-typhoon-haiyan/seed-distribution-map/en/

Visited: 100Accepted: 73

Current Front-End Example



Focused Event Crawler

Give as much information as you know about the URL's.

Event Type: Typhoon
Typhoon Name: Haiyan
City/State/Country: Philippines
Day/Month/Year:
Crawler Type: Event FC
Search Page Limit: 10

Site URL: http://www.latimes.com/wo
Site URL: http://www.usatoday.com/s

Submission 1

Entered URLs:

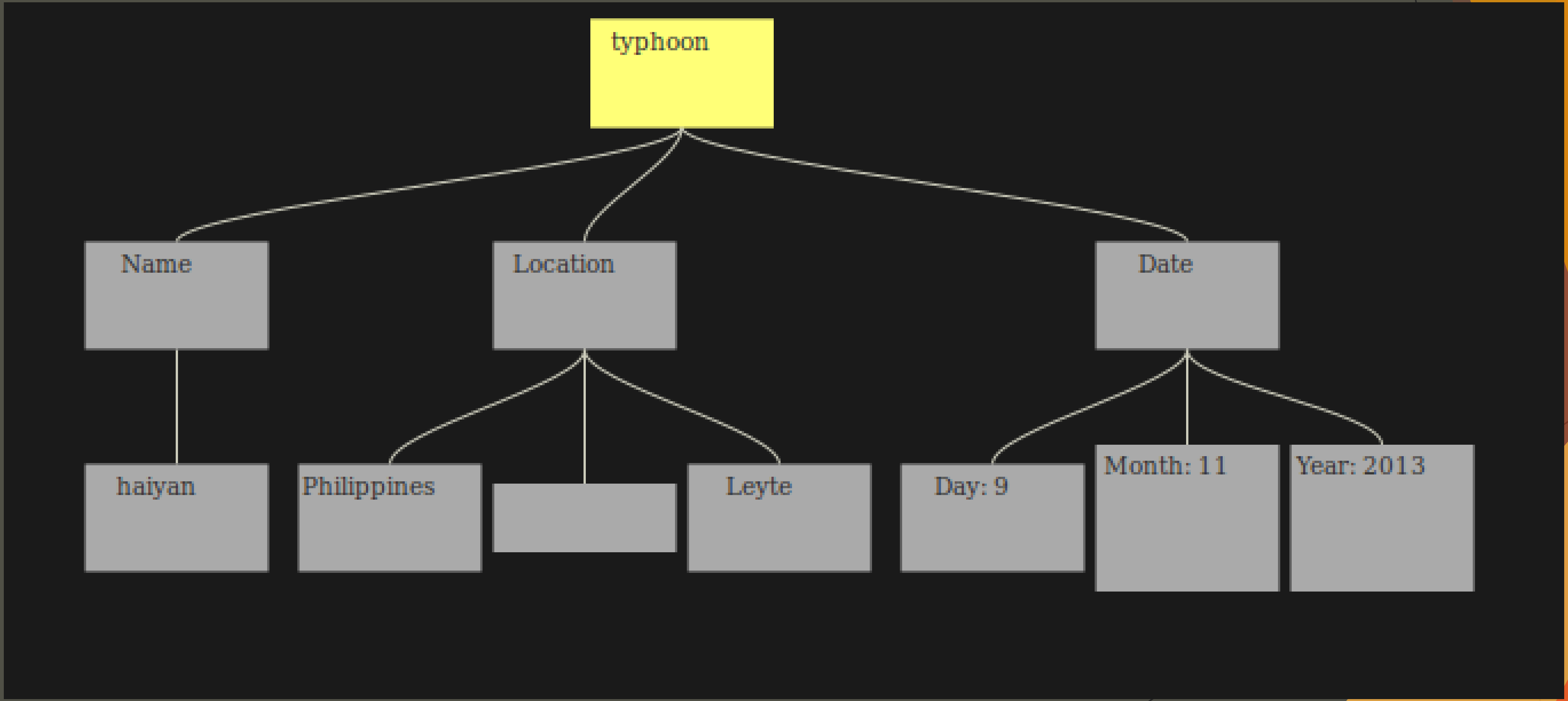
https://news.google.com/news?ncl=dzmbwHxDMHjAwBMDz8lHqqb6idqHM&q=typhoon+haiyan&lr=English&hl=en
http://www.usatoday.com/story/news/world/2013/11/09/typhoon-haiyan-philippines-vietnam/3483099/
http://www.latimes.com/world/worldnow/la-fg-wn-officials-typhoon-haiyan-death-toll-could-soar-to-10000-in-philippines-20131109,0,6461436.story#axzz2kE1iobhU
http://www.cnn.com/2013/11/08/world/asia/philippines-typhoon-haiyan/
http://www.cbsnews.com/8301-202_162-57611622/typhoon-haiyans-death-toll-rises-in-philippines/
http://www.weather.com/news/weather-hurricanes/super-typhoon-haiyan-latest-news-20131108
http://www.reuters.com/article/2013/11/09/us-philippines-typhoon-idUSBRE9A603Q20131109
http://en.wikipedia.org/wiki/Typhoon_Haiyan_(2013)
http://www.boston.com/bigpicture/2013/11/typhoon_haiyan.html
http://www.redcross.org.ph/

Found unique relevant URLs:

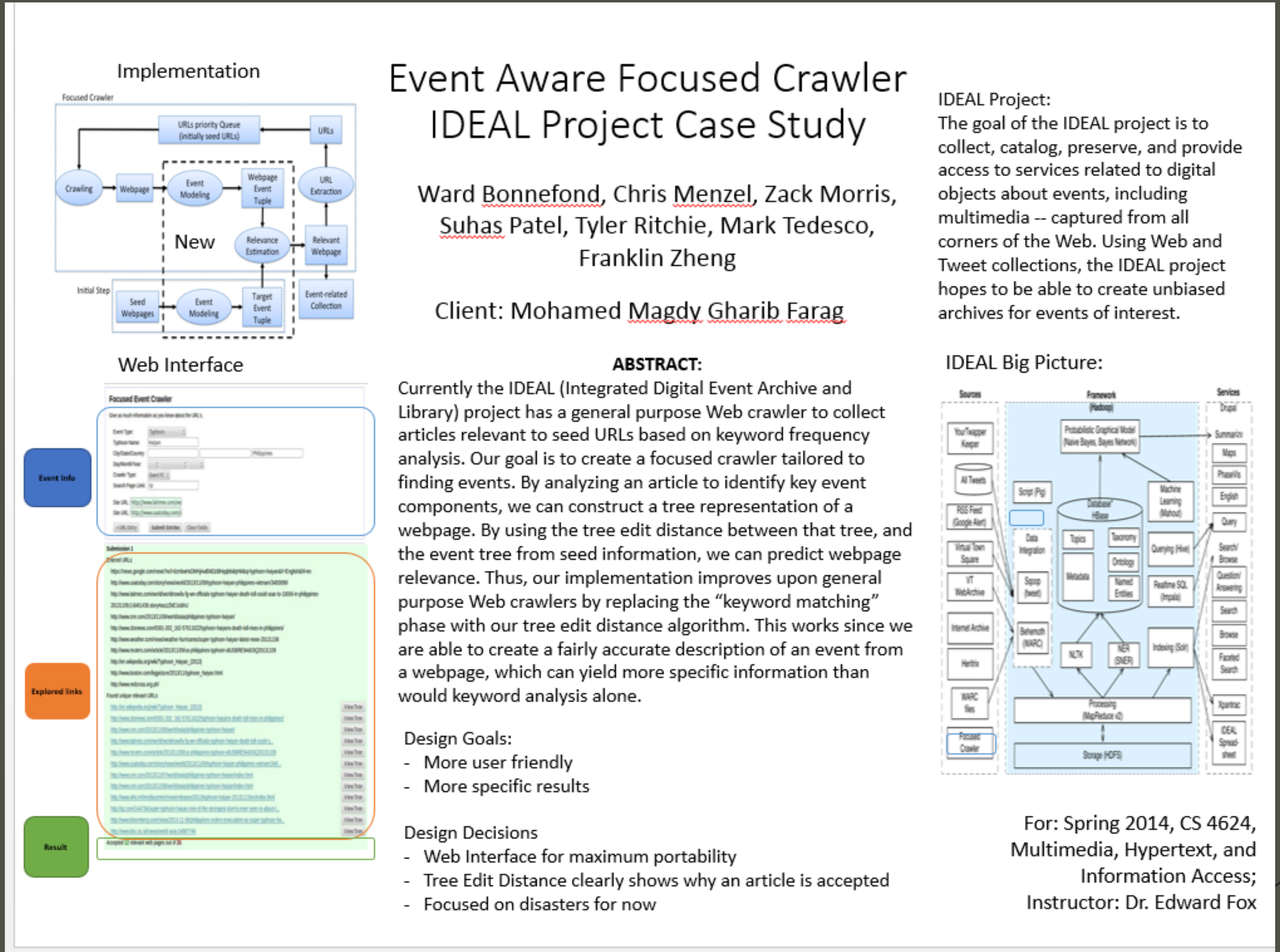
http://en.wikipedia.org/wiki/Typhoon_Haiyan_(2013) View Tree
http://www.cbsnews.com/8301-202_162-57611622/typhoon-haiyans-death-toll-rises-in-philippines/ View Tree
http://www.cnn.com/2013/11/08/world/asia/philippines-typhoon-haiyan/ View Tree
http://www.latimes.com/world/worldnow/la-fg-wn-officials-typhoon-haiyan-death-toll-could-s... View Tree
http://www.reuters.com/article/2013/11/09/us-philippines-typhoon-idUSBRE9A603Q20131109 View Tree
http://www.usatoday.com/story/news/world/2013/11/09/typhoon-haiyan-philippines-vietnam/348... View Tree
http://www.cnn.com/2013/11/07/world/asia/philippines-typhoon-haiyan/index.html View Tree
http://www.cnn.com/2013/11/08/world/asia/philippines-typhoon-haiyan/index.html View Tree
http://www.who.int/mediacentre/news/releases/2013/typhoon-haiyan-20131113/en/index.html View Tree
http://qz.com/144734/super-typhoon-haiyan-one-of-the-strongest-storms-ever-seen-is-about-t... View Tree
http://www.bloomberg.com/news/2013-11-06/philippines-orders-evacuation-as-super-typhoon-ha... View Tree
http://www.bbc.co.uk/news/world-asia-24887746 View Tree

Accepted 12 relevant web pages out of 20.

Current Front End – Tree View



Extras – Poster



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

Implementation

Focused Crawler
URLs priority Queue (initially seed URLs)
URLs
Crawling
Webpage
Event Modeling
Webpage Event Tuple
URL Extraction
New
Relevance Estimation
Relevant Webpage
Initial Step
Seed Webpages
Event Modeling
Target Event Tuple
Event-related Collection

Web Interface

Focused Event Crawler
Event Info
Explored links
Result

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

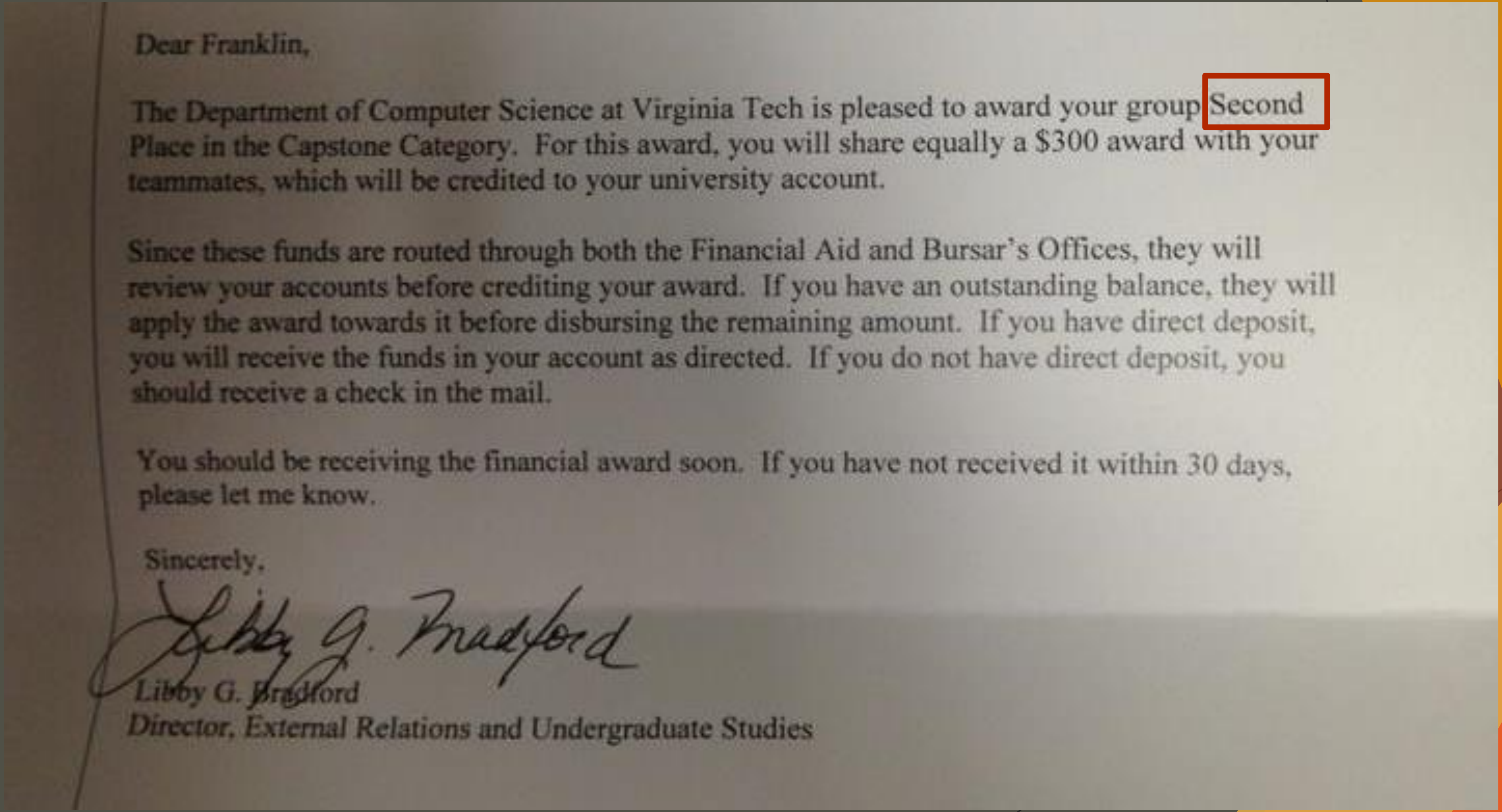
- Web Interface for maximum portability
- Tree Edit Distance clearly shows why an article is accepted
- Focused on disasters for now

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.

IDEAL Big Picture:

For: Spring 2014, CS 4624, Multimedia, Hypertext, and Information Access;
Instructor: Dr. Edward Fox

Extras – Second Place Letter



Dear Franklin,

The Department of Computer Science at Virginia Tech is pleased to award your group Second Place in the Capstone Category. For this award, you will share equally a $300 award with your teammates, which will be credited to your university account.

Since these funds are routed through both the Financial Aid and Bursar's Offices, they will review your accounts before crediting your award. If you have an outstanding balance, they will apply the award towards it before disbursing the remaining amount. If you have direct deposit, you will receive the funds in your account as directed. If you do not have direct deposit, you should receive a check in the mail.

You should be receiving the financial award soon. If you have not received it within 30 days, please let me know.

Sincerely,

Libby G. Bradford

Libby G. Bradford
Director, External Relations and Undergraduate Studies

Questions?