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

Virginia Tech – Blacksburg
CS4624
March 3, 2014

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

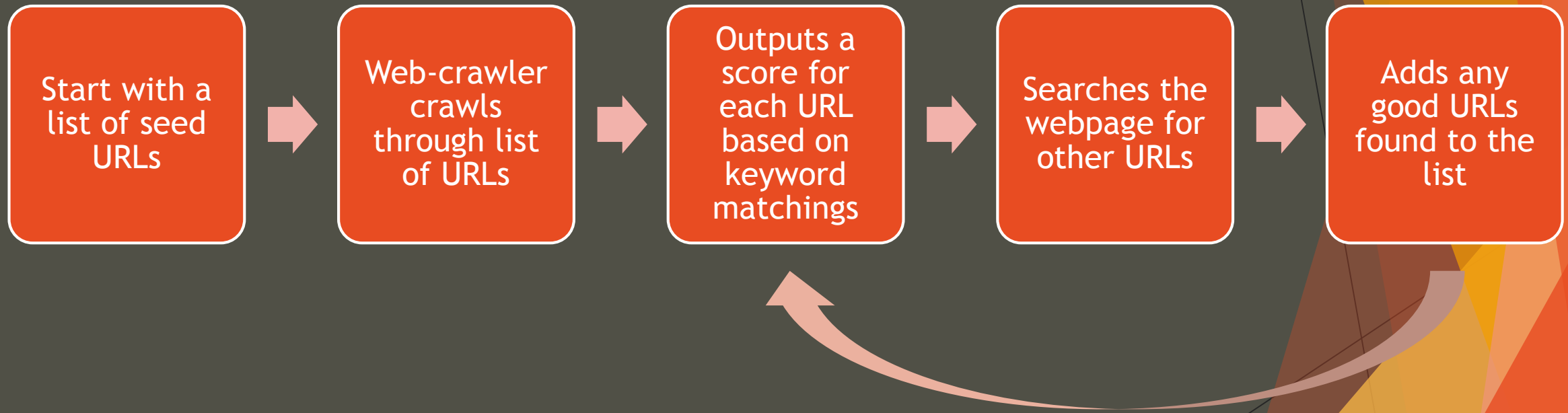
Enhanced focus crawler

- 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



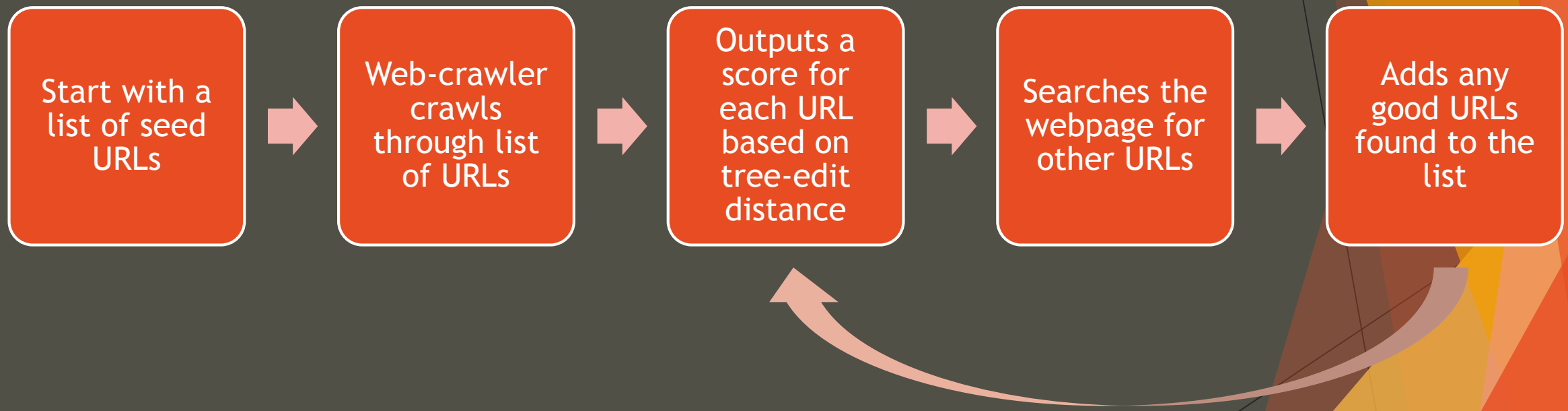
Current Progress

- 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



Current Back-End Example

```
[u'typhoon', u'philippin', u'haiyan', u'storm', u'2013', u'peopl', u'citi', u'ho
me', u'tacloban', u'said']
1,0.692901327128, http://en.wikipedia.org/wiki/Typhoon_Haiyan_(2013)
1,0.852749065569, http://www.boston.com/bigpicture/2013/11/typhoon_haiyan.html
1,0.853248253833, http://www.cbsnews.com/8301-202_162-57611622/typhoon-haiyans-d
eath-toll-rises-in-philippines/
1,0.79137585252, http://www.cnn.com/2013/11/08/world/asia/philippines-typhoon-ha
iyan/
1,0.881304262777, http://www.latimes.com/world/worldnow/la-fg-wn-officials-typho
on-haiyan-death-toll-could-soar-to-10000-in-philippines-20131109,0,6461436.story
#axzz2kE1iobhU
1,0.55161008187, http://www.redcross.org.ph
1,0.943095638363, http://www.reuters.com/article/2013/11/09/us-philippines-typho
on-idUSBRE9A603Q20131109
1,0.854781232935, http://www.usatoday.com/story/news/world/2013/11/09/typhoon-ha
iyan-philippines-vietnam/3483099/
1,0.587093476619, http://www.weather.com/news/weather-hurricanes/super-typhoon-h
aiyan-latest-news-20131108
1,0.612578121633, https://news.google.com/news?ncl=dzmbwHxDMHjAwBMDz8lHqqb6idqHM
&q=typhoon+haiyan&lr=English&hl=en
10
10
```

Keywords

Webpages that were crawled through (and their ranked scores)

Number of Pages Requested and Number of Pages Found

Current Front-End Example

Questions?