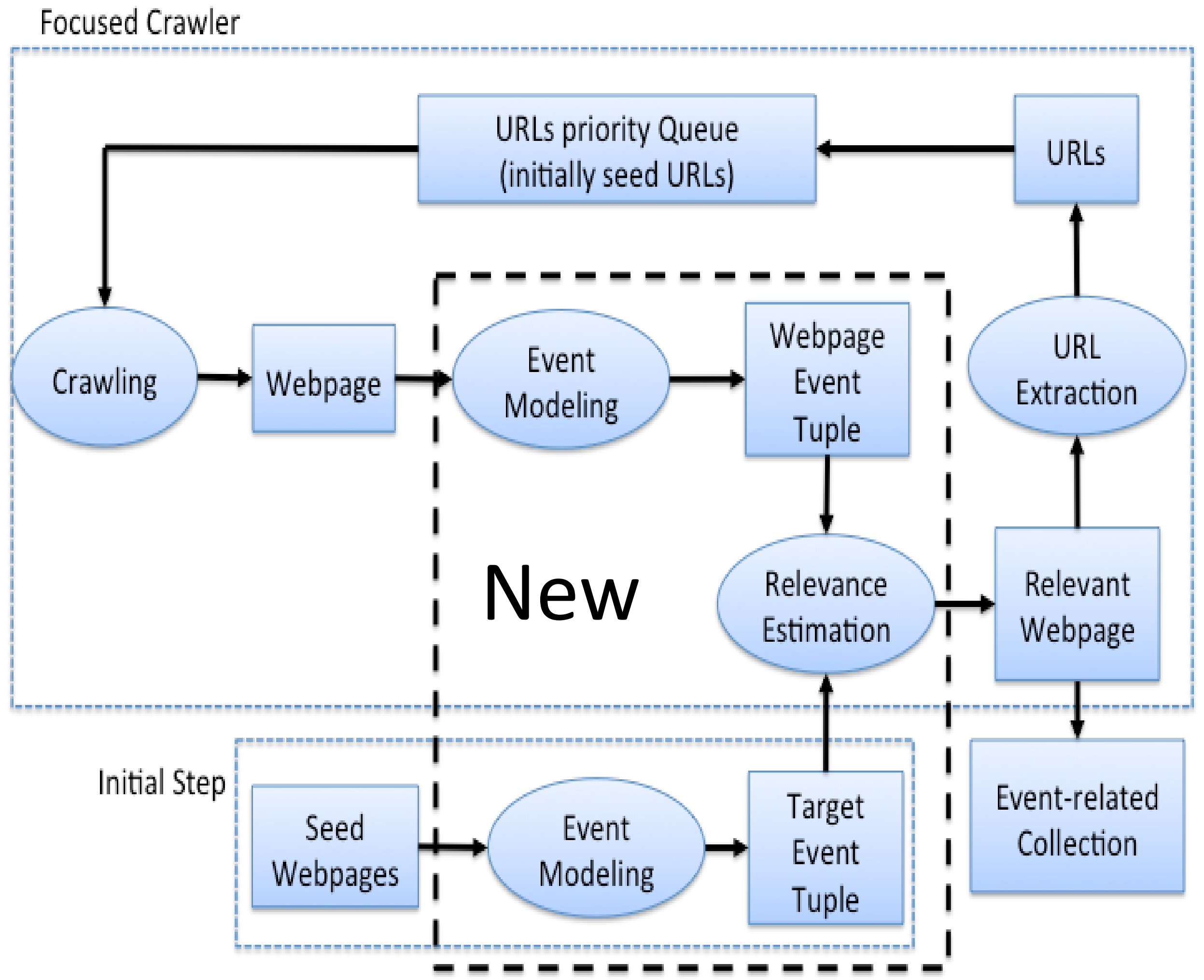
Event Aware Focused Crawler
IDEAL Project Case Study

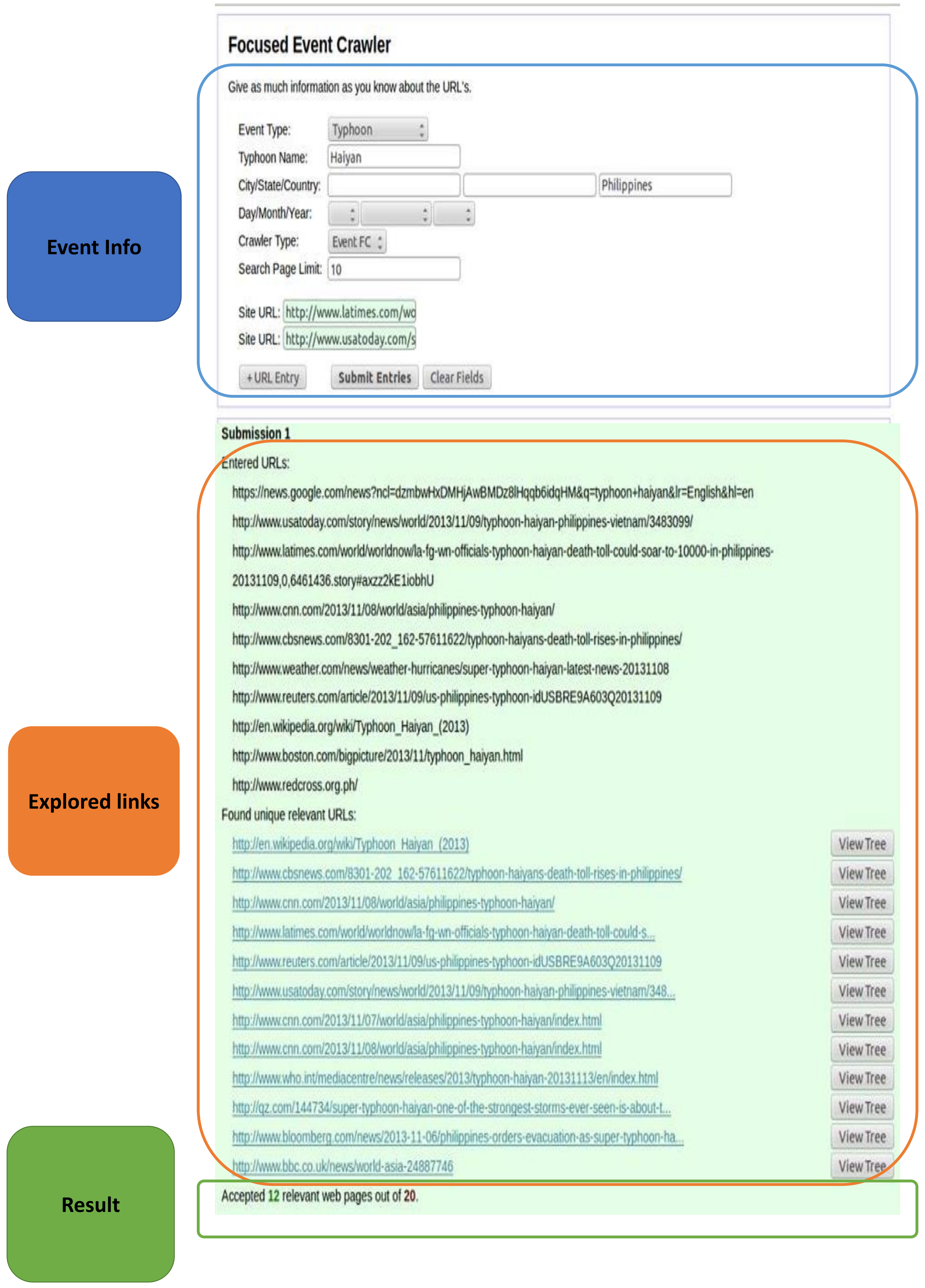
Ward Bonnefond, Chris Menzel, Zack Morris, Suhas Patel, Tyler Ritchie, Mark Tedesco, Franklin Zheng

Client: Mohamed Magdy Gharib Farag

Implementation



Web Interface



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.

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:

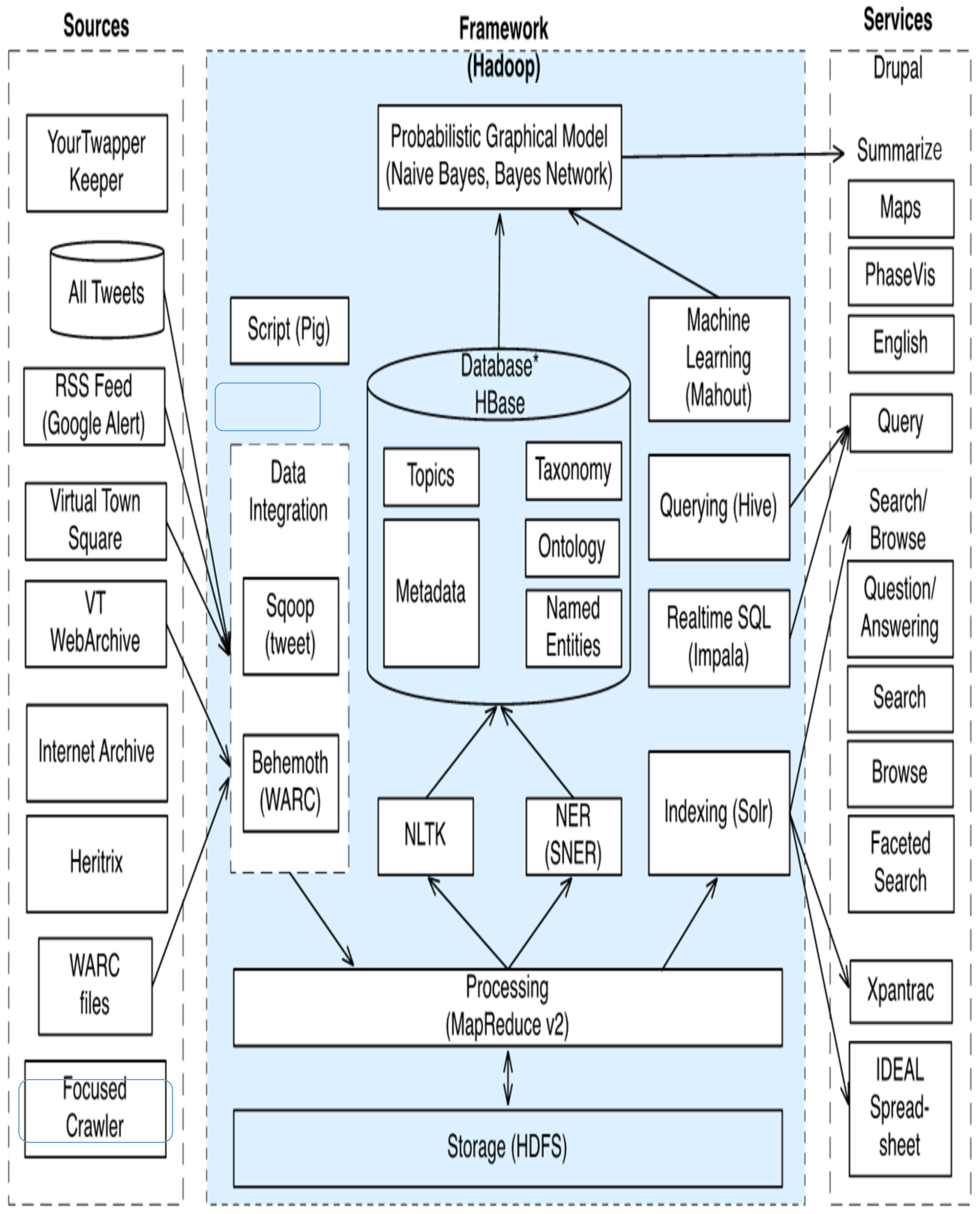
- More user friendly
- More specific results

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