CSE 403 Project Proposal
Cassius Butcher (cbutcher)
Harshad Petwe (hsp3)
Nolan Cash (nolanac)
Tyler Oshiro (tyoshiro)

3/30/2012

KNOW Project News Article Web Crawler and Categorizer

The KNOW Project at the University of Washington has the potential to be a revolutionary new tool in the research and analysis of international events. However, the project currently lacks a systematic way of adding new article annotations as well as an effective method of searching through the annotations that already exist. Our team proposes to solve both of these problems by creating a periodic web crawler that inspects news websites on a daily basis, analyzes metadata of new articles to create annotations, populates a database with those annotations, and allows users to make sophisticated queries on all annotations. This product will allow the students of the Jackson school and any other users of the KNOW Project to more effectively find and analyze relevant international news articles.

The primary modules of our system will be an initializer, a crawler, an archiver, and retriever. The initializer will be responsible for triggering the crawler to begin crawling each day and will specify which news sources to retrieve data from. Further research will need to be conducted to decide the optimal time of day (likely around 3 a.m. PST) to crawl as well as the maximum number of news sources our database framework will be able to support. The crawler will navigate to the specified news sites and scan through the new articles for that day. Metadata will be retrieved from the HTML tags of these articles to provide relevant annotation information such as dates, topics, authors, etc. The archiver will then be responsible for receiving the metadata from the crawler and constructing a properly formatted annotation. This annotation will be stored in the KNOW Project database in an efficient way so that it may be queried later. Finally, the retriever will be the tool utilized by the user to make sophisticated queries of the annotation database to allow for comparison of articles based on any of the metadata previously retrieved.

Our concept design will be implemented with existing technologies capable of accomplishing our objectives. The exact system design will be easier to develop and more accurate once we take a look at the web and database framework that the KNOW Project already has and how we the developers can interface with it. However, we already have a good idea of how the main features will be implemented. The web crawling could be done in a few languages. Python has libraries dedicated to interfacing with html pages and tags, so this may be the best language option to use for reading html tags. Converting tags to annotations should be fairly straightforward in any language as long as we choose and adhere to one data format convention. The other challenge would be properly interfacing with the database. We will need to examine what kind of database framework is being used by the KNOW Project and we will need to determine whether we can find existing libraries to interface with the database or whether we need to write our own libraries to do so. Another possibility is that we may need to use two different programming languages: one to web crawl and the other to interface with the database; in that case our main challenge would be to figure out how to get the two languages to properly communicate without losing correctness or efficiency. Our primary development environment will likely be Eclipse since it has good plugins for collaboration tools like SVN.

We have the foundations of our developing framework in place and are prepared to meet the challenges we are likely to encounter. The single most serious challenge that we see in developing the product in a timely manner is trying to integrate our web crawler with the already existing database that the Jackson School has provided. More discussion with the Jackson School will be needed to thoroughly determine their specific requirements and preferences. Communication will be the primary method of minimizing the risk in creating a program that doesn't meet our client's needs. Other challenges include a possible blocking of our web crawler due to too many page requests over a brief span of time. It is possible that we will have to use some sort of proxy to mask our page requests to crawl the news articles without being detected.

We are proposing an idea that will have exciting implications for both our clients as students of international issues and ourselves as developers. These features will give students the ability to compare and contrast news stories and track their changes across both geographic regions and time. This will allow students to analyze the effects of cultural and societal biases on the publication of news articles. Furthermore, this project presents interesting technical challenges to us as developers. We are interested in the process of developing a web crawler, including learning which languages are the most efficient to implement a crawler and what libraries are best to utilize. We are also interested in seeing how our program will interface with the KNOW databases and exploring the best database design to keep our features time and space efficient.

Our proposal is truly a novel concept. The closest thing we have to competitors are other services that compile news sources like Google News, Flipboard, and RSS readers. However these products don't allow researchers to track and compare articles based on geographic locations or time. It is these core features of our proposal that will make the KNOW Project a truly revolutionary research tool in the field of international studies.