A Politics Mash-Up
Inki Hong
Faculty Advisor: Ben Taskar

Abstract:
- A mashup in the worldwide web is an application that collects data from multiple sources of information and presents them in a unique way. This paper describes a mashup on politics that integrates both geographical and chronological dimensions regarding the 2008 presidential election of the United States of America.

Overview:
- The system is composed of three major subsystems.
  - The backend subsystem is responsible for collecting data from multiple sources of information, analyzing the assembled data, and extracting useful information from them.
  - The middle subsystem is responsible for permanently storing constructed information in a structured way so that it serves as a link between the front and back.
  - Finally, the frontend subsystem is responsible for presenting these pieces of information in a geographically and chronologically meaningful way.

Frontend:
- Various Web technologies are used to implement the frontend subsystem.
  - JavaScript with Google Maps API/AJAX provides client-side interactive functionality.
  - JSP/Tomcat is used to implement server-side scripting.

2-Pass Crawling:
- Due to the magnitude of the World Wide Web, some information are difficult to collect by automated programs, and those information such as past news articles are fetched through more complicated methods such as two-pass crawling.
  - Phase 1: Web robot searches news sites with keywords related to the upcoming election and then fetches the search results. At this point if the site gives all the necessary information, the robot is done. If not, it proceeds to next step.
  - Phase 2: The robot visits popular web search sites where past news articles are already indexed, query for a new set of keywords constructed from previous phase before fetching its results.
  - Phase 3: Among search results from previous phase, the robot downloads only the results that match the original election-related keywords with high score.

Geocode Index:
- Geocoding is the process of converting street addresses and zip codes to latitude-longitude coordinates. Coordinate conversions at runtime can be a costly operation. The geocoder module builds this mapping data based on zip codes in advance so that faster lookup is possible. This data is also maintained by Lucene along with collected data.

Lucene Search Engine:
- This system leverages Lucene, an open-source search engine system, over traditional RDBMS for permanent storage.
- It is 100% Java, thus easier to integrate with other parts of the system, which are also based on Java.
- It is easy to use and to maintain indexes (tables), one of the reasons being that they are intuitively implemented as directories on file systems.
- Search engines maintain inverted tables very efficiently so that text-based searching is faster than traditional databases.

Conclusions:
- This system downloads pieces of information related to the 2008 Presidential Election from multiple websites.
- It organizes them by discarding unnecessary parts and compiling them in a structured format through indexing so that searching over the data is fast and accurate.
- Finally, it presents meaningful information using web application with chronological and geographic dimensions.
- For further improvements, both back and front subsystem can be extended.
- More web robots can be incorporated to give a richer set of data into the system. And robots themselves can be automated into server processes.
- Web interface can be improved according to potential user's needs, such as automatic search word suggestion based on popular searches.