Content Mining for the Political Blogosphere

The Web contains a large amount of content that is useful when studied from the right perspective. Blogs are an important subset of this content, and political blogs are particularly useful during elections. Content mining in political blogs provides a structure to answer questions such as the following:

• Who are the top bloggers in left-wing politics?
• What are the “hot topics” in the political blogosphere?
• Show me popular blogs related to Homeland Security.

Overview

A Few Goals

Generate a dynamic “text cloud” that shows what the current “hot topics” are in the political blogosphere:

Abortion Environment Gun Control
Immigration Iraq Healthcare Jobs
Welfare Homeland Security

To provide a natural language search that returns relevant blogs based on content rather than conventional categorization:

Blog Content Search
Which candidate has the best healthcare plan?

Generate a dynamic public opinion graph based on blog content:

Public Opinion of Obama on National Security Over the Last Week

Component Descriptions

Blog Crawler
Crawls the Web to find RSS and Atom feeds for blogs. It pushes these feeds into a set of packed data files known as the feed repository.

Feed Parser
Bloggers may publish in any of several XML-related formats. The parser identifies a post's format, then extracts and maps data fields into the page repository database.

Content Downloader & Extractor
Extracts content blocks from the HTML and writes this content to the content warehouse (an OLAP database).

Content Indexer
Takes the content output by the extractor and indexes it by keywords (or tokens). Indexing is performed by Apache Lucene.