Exploring history through newspaper archives

Jasna Škrbec, IJS
Marko Grobelnik, Blaž Fortuna,Boštjan Pajntar

October 10th 2011

ailab.ijs.si

Outline

- Introduction
  - Archives
- Motivation
- Architecture
  - Preprocessing
- Components
  - Searchpoint
  - Graphs
  - Document Atlas
  - Timeline
- Future work
- Demo

Introduction

- News publishers collected archives of news
  - Our goal is to build a system to make such archives usable through text mining & visualization
- Archive characteristics:
  - Large corpora (up-to few M articles)
  - Rich meta data (specific for each archive)
  - Different input formats (xml structure)
  - Poor search interfaces (not specialized for archives)

New York Times archive

- 1987 – 2007
- over 1.5M articles
- Almost 20GB
- Meta data
- Covering news all over the world

Example of an article

Flooded Midwest Braces for More Storms

By Gretchen Ruethling, January 5th, 2005

Five Midwestern states where flooding has killed 11 people and forced thousands from their homes were bracing for worse this weekend, as the storm that caused mudslides in California continued its march east on Friday.

Roads were closed and residents evacuated in scattered spots from West Virginia to California, where more than 1,000 fled their homes near Corona after an earthen dam began to seep water.

In the Midwest, the hardest-hit areas were in Ohio and Indiana, whose governors declared states of emergency in the flooded areas.

Joe Heim, a meteorologist with the Ohio River Forecast Center of the National Weather Service, said the Maumee River in northwest Ohio, the Wabash River on the western border of Indiana and the Ohio River downstream of Evansville, at Indiana's southwest tip, were still rising and posed threats.

A woman and her 22-year-old son were electrocuted on Thursday in Shirley in central Illinois when flash-floods sent a foot of water into their basement.

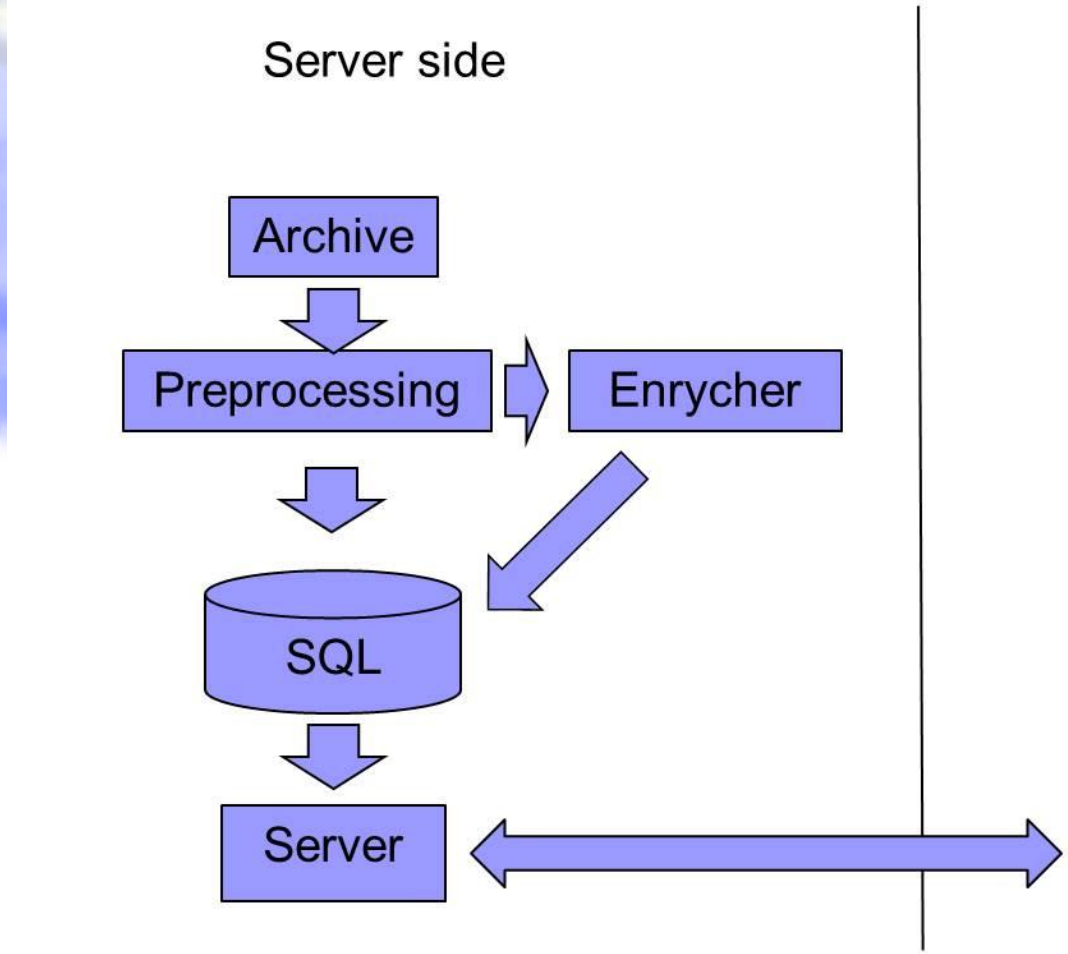
…

- Enrycher keywords
  - Natural Disasters and Hazards
  - United States
  - North America
  - Science and Environment
- Enrycher categories
  - Science/Earth Sciences/Natural Disasters and Hazards/Floods/Warnings and Forecasts
- Meta data keywords
  - Weather
  - Mudslides
  - Rain
  - Floods
- Meta data classifiers
  - Top/News/U.S./Midwest
  - Top/Features/Travel/Guides/Destinations/North America

Motivation

- handling large data structures
- helping user search and browse through archives
- helping user read more about related topics
- visualizing how things are connected in time, place, etc.
- getting user's attention and interest in other related issues
- showing context
- recognizing stories through articles through time

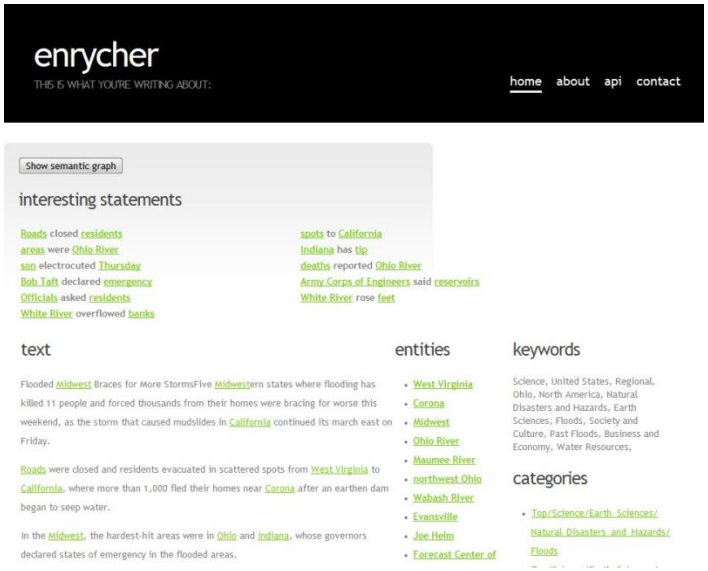
Architecture



Client side

Preprocessing

- Extracting content from xml files
  - Title, text, author, date
  - Next step is to extract meta data specific for each type of archive
- Extracting context with Enrycher
  - Extraction of entities
    - people
    - organizations
    - locations
  - Classification
    - Dmoz topic ontology
  - Extraction of keywords



Exploring Archive

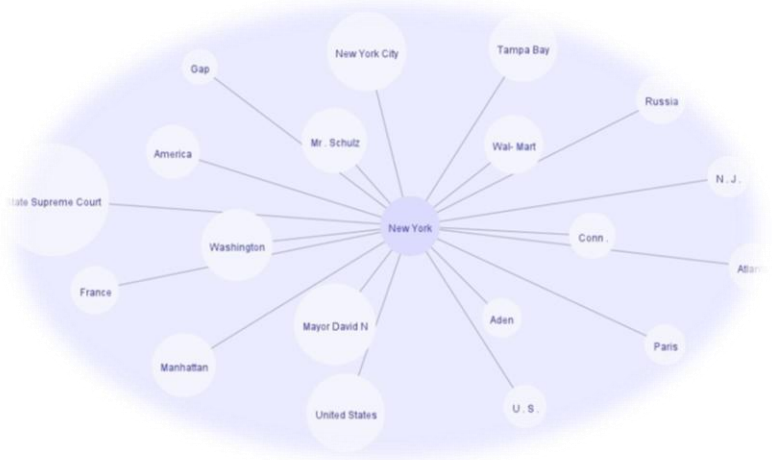
- Faceted Search interface
  - search by entities, keywords, categories, authors, dates
- Directory interface
  - Top categories
  - Lists of authors, keywords, entities, years

Searchpoint

- Visualization of search results
- Dynamic ranking
- Multidimensional
  - Person
  - Location
  - Organization

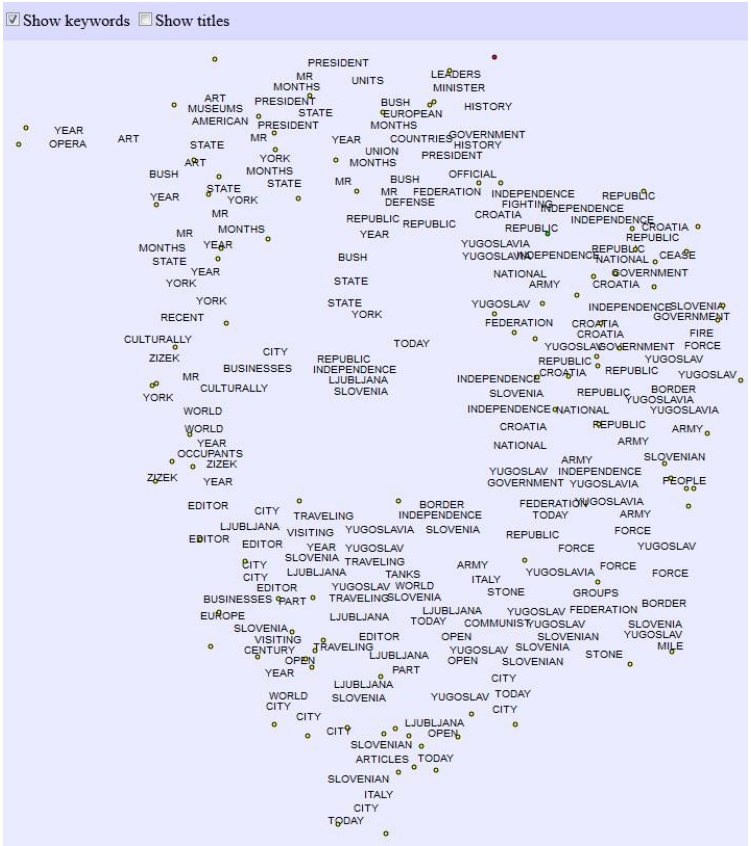
Network of Entities

- Connection between entities
- Width of the connection corresponds to the strength
- Size of the entity corresponds to the intensity in articles



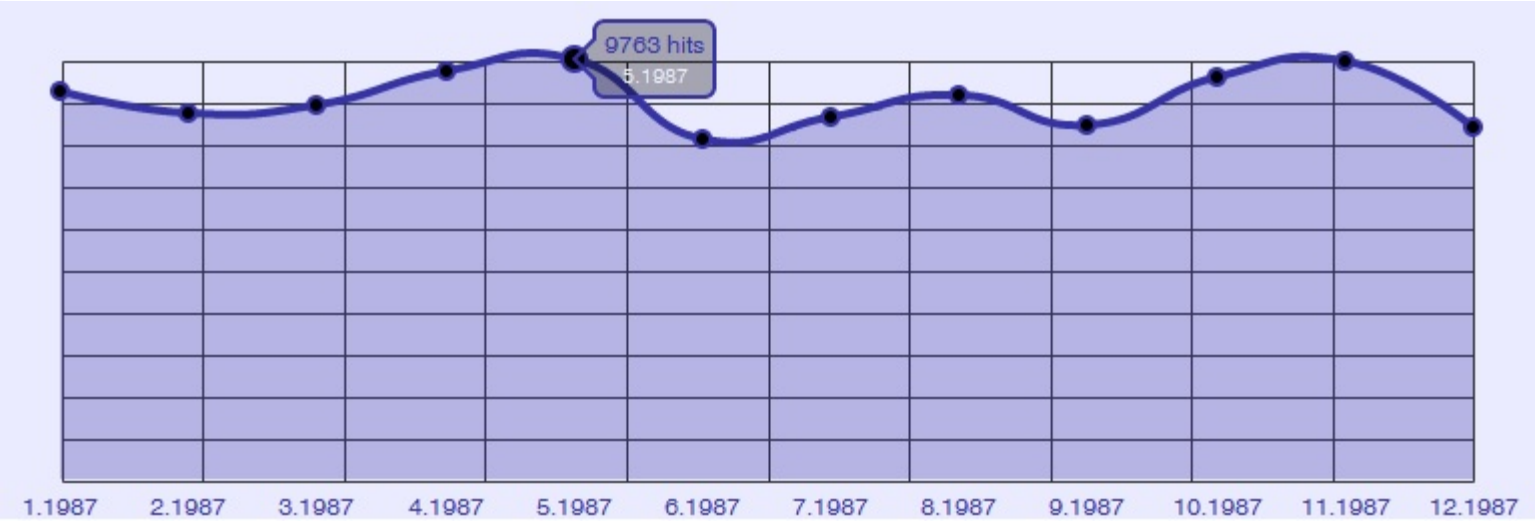
Document Atlas

- Visualization of search results
  - Based on similarity between articles
  - Articles of same topic or same story are closer together
- Keywords
  - Extracted from nearby articles



Timeline

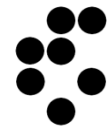
- Time component is important in archives
- Number of articles during a year
- Instance of an entity over the years



Plans for the future

- Improve search
  - narrowing criteria
  - suggestions
- Adding more new visualizations and tools developed in AiLab to improve search and presentation of content in time, space and other contexts
- Adding links to similar content (stories)
- Adding links to outside resources (like dbpedia) or bring this resources inside this application
- Improve usability & appearance of user interface
- Search for more new things and ideas…

DEMO



?