



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)

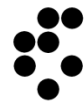




New York Times archive

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

The screenshot shows the New York Times website interface. At the top, there's a navigation bar with the United logo, the newspaper's name 'The New York Times', and the date 'Friday, October 7, 2011'. Below this is a banner for 'Get a Full Times Experience' and 'BECOME A DIGITAL SUBSCRIBER'. The main content area features a large article titled 'Some Jobless Find Fault in Extending Aid to Unemployed' by Shaila Dewan. To the right of this article is a photo of three women, with the caption 'Three Women Share Nobel Peace Prize'. Below the main article is a section titled 'Central Banks Take Different Tacks on Europe's Economy'. The left sidebar contains a 'Switch to Global Edition' dropdown and a list of categories like 'JOBS', 'POLITICS', 'TECHNOLOGY', etc. The right sidebar includes a 'Log In With Facebook' button, 'WHAT'S POPULAR NOW' section, and a 'MARKETS' table showing stock prices for various countries. At the bottom right, there's a large advertisement for digital subscriptions with the text 'WHERE THE STORY COMES FIRST. BECOME A DIGITAL SUBSCRIBER. JUST 99¢ FOR 4 WEEKS'.





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



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

enrycher
THIS IS WHAT YOU'RE WRITING ABOUT: [home](#) [about](#) [api](#) [contact](#)

Show semantic graph

interesting statements

Roads closed residents areas were Ohio River son electrocuted Thursday Bob Taft declared emergency Officials asked residents White River overflowed banks spots to California Indiana has tip deaths reported Ohio River Army Corps of Engineers said reservoirs White River rose feet

text

Flooded [Midwest](#) Braces for More Storms 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.

entities

- [West Virginia](#)
- [Corona](#)
- [Midwest](#)
- [Ohio River](#)
- [Maumee River](#)
- [northwest Ohio](#)
- [Wabash River](#)
- [Evansville](#)
- [Joe Helm](#)
- [Forecast Center of](#)

keywords

Science, United States, Regional, Ohio, North America, Natural Disasters and Hazards, Earth Sciences, Floods, Society and Culture, Past Floods, Business and Economy, Water Resources,

categories

- [Top/Science/Earth_Sciences/Natural_Disasters_and_Hazards/Floods](#)



Exploring Archive

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

Source: The New York Times The New York Times archive Nature Reuters

Search all fields:
Search person:
Search organization:
Search location:
Search entity:
Search keyword:
Search author:
Search date:
Search category:

Main categories:

- Arts
- Computers
- Regional
- Shopping
- Reference
- Business
- Recreation
- Science
- Society
- Sports
- Kids_and_Teens
- Health
- Games
- Home
- News
- Adult
- _Dragons_Games
- _Dragons
- _Conquer_Games
- M_University
- _Clank_Series

Entities

- Dina Ibrahim
- Ms. Kalf
- Israel Street
- PO Box 28
- 10 Hawthorne Lane
- Istip Latin
- 68 Wheeler Road
- Culber Jazz Ensemble
- Washington Buddy
- Beach Dickey
- Jefferson Bob
- European and Eurasian Affairs
- Antony Maanum
- Rachel Fricker
- Chris Forestieri
- Torre Hustling Home
- ROBERT BIRNBAUM-Robert
- Joseph L. Terlizze
- Mr. Schmapp
- Keith A. Robinson

Keywords

- Arts
- Web Applications
- On the Web
- Internet
- Computers
- Writers Resources
- Contests
- Weblogs
- Photo Sharing
- Crafts
- Switzerland
- Regional
- Europe
- Business and Economy
- Video
- Shopping
- Recordings
- Entertainment
- DVD
- Reference
- Knowledge Management
- Knowledge Flow
- Design
- Stocks and Bonds
- Resources
- Regions
- Misano Adriatico
- Localities
- Italy
- Emilia-Romagna

Years

- 1987-1988
- 1989-1990
- 1991-1992
- 1993-1994
- 1995-1996
- 1997-1998
- 1999-2000
- 2001-2002
- 2003-2004
- 2005-2007

Authors

- Elaine Louie
- DARYLN BREWER
- Bernard Gladstone
- Angeline Goreau
- AP
- Reuters
- ROBERT HANLEY
- Special to the New York Times
- JOHN CORRY
- JOHN S. WILSON
- JOHN A. KINCH
- PAULA DEITZ
- MEL GUSSOW
- Daphne Angles
- HOWARD G. GOLDBERG
- The Associated Press
- FRANK J. PRIAL
- DANIEL F. CUFF
- H. J. MAIDENBERG
- THOMAS MORGAN
- PATRICIA LEIGH BROWN
- William Safire
- C. Gerald Fraser
- WILL CRUTCHFIELD
- KENNETH N. GILPIN
- ALAN TRUSCOTT
- THOMAS W. ENNIS
- Dave Anderson
- WILLIAM R. GREER
- Jon Pareles



Searchpoint

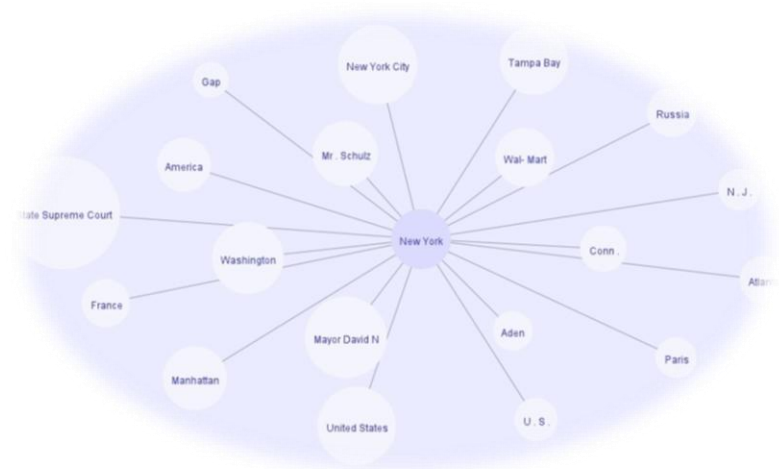
- Visualization of search results
- Dynamic ranking
- Multidimensional
 - Person
 - Location
 - Organization





Network of Entities

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





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



?