Alexandria Demonstrations

The Alexandria Team
HistDiv – History by Diversity: Helping Historians Search News Archives

Jaspreet Singh
Exploring News Archives

Complex Information need

Pan & Zoom

Focus & Context

Why not support both?

Figure 2: Iran's Total Exports and Imports, 1986-2006

Annual export growth rate 1987-2006: 8.6%

Annual import growth rate 1987-2006: 7.0%
The iCrawl Wizard
Supporting Interactive Focused Crawl Specification

Gerhard Gossen, Elena Demidova, Thomas Risse
Focused Web Crawling for Web Science

Creation of large Web datasets through Focused Crawling:

- Automatic collection through Web Crawler
- Collector specifies **topic of crawl** through keywords and entities
- Only documents related to topic are collected

*iCrawl*: Help scientist create Web datasets **easily**

- Integrated crawling of Social Media
- Better monitoring & analysis of running crawls
- **Simplified creation of (focused) crawls:** Crawl Specification Wizard

Gossen et al., “iCrawl: Improving the freshness of web collections by integrating social web and focused web crawling,” JCDL ’15
Demo: http://icrawl.l3s.uni-hannover.de:8090/
Conclusions and Future Work

Implementation of Wizard for Focused Crawling

- Easy to use
- Extensible
- Puts crawling into the hands of researchers

Current Work

- Re-crawling of Web archives to create topical sub-collections (submitted)
- Additional sources
  - Social Media
  - Directories (e.g. German newspapers)
Temporal Analysis of Social Networks in Archives

Miroslav Shaltev
Temporal Analysis of Social Networks in Archives

Approach
- Background: computationally intensive cluster process.
- Foreground: server / client solution.

Example
Temporal evolution of a graph constructed with seeds Barack Obama, John McCain from a small portion of the British archive data in the period October 2008 – October 2009.
Analysing Language-specific Differences in Multilingual Wikipedia

Simon Gottschalk
On June 17, 1958, as a crane stretched from the north side of the new bridge to join the two chords of the unfinished arch, several spans collapsed. Seventy-nine workers plunged 30 metres (100 ft) into the water. Eighteen were killed either instantly or shortly thereafter, possibly drowned by their heavy tool belts. A diver searching for bodies drowned later, bringing the total fatalities for the collapse to 19. In a subsequent Royal Commission inquiry, the bridge collapse was attributed to miscalculation by bridge engineers. A temporary arm, holding the fifth anchor span, was deemed too light to bear the weight. [2]
Tempas – Temporal Archive Search Based on Delicious and Twitter

Helge Holzmann
Temporal Archive Search

- Search a Web archive based on external (temporal) data sources (e.g., social networks)
- Custom index structure allows for temporal query suggestions and captures temporal relevance
- Links to Internet Archive’s Wayback Machine
  - No need to index the entire archive
  - Naturally filters less important content
  - Search how a website is described in contrast to how it ‘describes’ itself

Tags / keywords + URLs + times extracted from Delicious posts and Tweets
Suggesting News Articles for Enriching and Expanding Wikipedia Entity Pages

Besnik Fetahu
Automated News Suggestion to Wikipedia Entity Pages

Task #1

Feature extraction

News article

As It Happened: Cyclone Reaches Orissa

Some half a million people were evacuated from the southeastern Indian coast as Cyclone Phailin, a tropical storm from the Bay of Bengal, bore down on India. The states of Orissa and Andhra Pradesh, both of which have large coastal populations, were on high alert ahead of the storm’s expected arrival.

Task #2

One classifier per entity type

Article section placement

[state]: geography

[city]: climate

Entity page

Odisha

Bay of Bengal

Phailin

Sections
ArchiveWeb: Working with Web Archive Collections

Trevor Fernando Zeon
ArchiveWeb: working with web archive collections

- Collaborative search and sharing platform
- Building workable collections of resources
- Sharing of collections between a group of people
- User annotations for resources in the form of tags and comments
- Flexibility to build your own personal archive of a resource
Archive-It Interface

Human Rights Documentation Initiative

Collected by: University of Texas at Austin Libraries, Human Rights Documentation Initiative

Archived since: Apr. 2007

Description: The University of Texas Libraries Human Rights Documentation Initiative Collection features fragile websites containing human rights documentation and related content from human rights organizations and advocates around the globe. Browse collection by topic, geographic area, language or do a keyword search at http://www.library.utexas.edu/archives.

Subject: Society & Culture, Politics & Government, Law & Legal Studies

Narrow Your Results

Sites for this collection are listed below. Narrow your results as left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Enter search terms here

Resource

Title: La Strada Moldova
URL: http://lastrada.md/

Description: Moldovan and some English. Focuses on women's rights, domestic violence, and human trafficking. Government continues to allow and in some cases condone discrimination based on minority, gender, and religion; in some regions NGOs are routinely threatened and harassed.

Captured 32 times between Apr 29, 2007 and Apr 9, 2015

Subject: Governance, rule of law, and human rights

Title: Ligne des Droits de la Personne dans la Region des Grands Lacs / League for the Defense of Human Rights of the Great Lakes
URL: http://ldgl.org/

Description: French and some English. Human rights umbrella group with members from multiple countries including Burundi, Tanzania, Malawi, and Rwanda.

Collector: UT Libraries
ArchiveWeb Interface

Collections

Resource Preview

User Annotations

Web Science – Investigating the Future of Information and Communication

Ivana Marenzi, Trevor Fernando Zeon
WikipEvent: How we get events from the crowd?

Binh Giang Tran
Our Target and Solution

- **Tracking the evolution of event from Wikipedia**
  - How Michael Phelps becomes the holder of the record of most medals ever?
  - How was “the death of Osama Bin Laden”

- **Finding similar events by event similarity measurement**
  - Events are collected from Wikipedia revisions and WCEP
  - Represent an event (text, entities, references) as a high-dimensional vector
  - Vector similarity measurement: Cosine / Jaccard / …

- **Improvement**
  - Would like to use hashing technique for a fast and effective implementation (Mitzenmacher et al. 2014)
Information overload problem

- A thousand of daily reported events
- Tons of daily news articles
- A lot of information sources:
  - Mainstreams
  - Social Media
  - Blogs
  - ...

According to Chartbeat, over 92,000 articles are posted to the web every 24 hours.
Source: A Day in the Life of Web Content
A Bing-inspired Ranking for Entity-centric Web Archive Search

Tu Nguyen
• Search Engines provide a good quality ranking for general web search.

• Web archive search and general web search have similar information needs for some types of queries.

• There is no state-of-the-art ranking model for web archive in literature.

  • Lack of user feedbacks, what are the real user intents?