Longitudinal Analytics of Web Archive data
Duration: Sep 2010 - Aug 2013

Zsolt Fekete, zsfekete@ilab.sztaki.hu
Andras Benczur, benczur@ilab.sztaki.hu
LAWA Partners

Archives
- Internet Memory, France
- Hanzo Archive, UK

Research
- HUJI - Scott Kirkpatrick
- MPI Saarbrücken - Marc Spaniol
- U Patras, Greece

FI Conference 16-17 May, Budapest
LAWA players/nodes and roles

Users
- Pose queries to archived data

Super Users
- Pose management queries to monitor

Archiving Agents
- Init crawls
- Collect results
- Process results:
  - Annotate with timestamps
  - Rank

Indexing Agents
- Keyword search
- Range queries

Catalogue Agents
- Where is data, indices etc – for query routing

Data Store Agents
- Data
- Metadata
- Indices
- Catalogues
  - ...

Query Processing Agents
- utilize catalogues, indices, etc to route queries, process results, monitor the Virtual Web Observatory and perform system tasks such as load balancing, resource management, etc.
Web Content Analytics: Challenges

- Web Archive Data is
  - BIG - serioiusly big!
  - Noisy, made by „mad people”
- Content boundary is less and less clear
  - URL no longer determines content - JavaScript, stateful browsing (see Facebook)
  - Content, host, site and server - orthogonal notions
- Brute force is not the answer
  - Scaling, distribution, smart reuse
  - Moving data is not viable
  - Debugging is a pain: distributed, and doesn’t matter if things break
LAWA Goals

- Provide LAWA participants with access to:
  - Large data sets
  - Map-Reduce (Hadoop) clusters

- Use these facilities to:
  - Develop new/advanced algorithms for data analysis - Virtual Web Observatory
  - Enhance the capabilities of Hadoop:
    - Network aware scheduling
    - Operations with multiple Hadoop file systems
    - Access to Internet Memory data sets
Map-Reduce??
Map-Reduce

Input

| Record | Record | Record | Record | Record | Record | Record | Record | Record | Record |

Map Task (Key, value)

Shuffle/Sort

Map Task (Key, value)

Shuffle/Sort

Map Task (Key, value)

Shuffle/Sort

Reduce

Reduce

Reduce

Output

| Record | Record | Record | Record | Record | Record | Record | Record | Record | Record |

Zsolt Fekete, András Benczúr  Web Analytics  Future Internet Platform 2011 Május 3
Example: frequent patterns

• Frequent URLs or words in Web pages, click logs, etc.
  twitter.com, Android, iPad, ...

• Frequent n-grams in Web pages, query logs, etc.
  Google street view, street view objection, BP oil spill, yes we can, ...

• Co-occurrence counts for URLs, words, n-grams, etc.
  Sandra Bullock Golden Raspberry, Sandra Bullock Academy Award

• Frequent phrases over time, interesting phrases as discriminators

• Entity-level analytics (over time):
  people, places, organizations, events, etc.
  meetings of Obama with companies and places that he has visited
Map-Reduce: Frequent n-grams

FUNCTION map($i, P_i$)
   List $\mathcal{N} \leftarrow$ generateNgrams($P_i$)
   FOR $v_i \in \mathcal{N}$ DO
      emit($n_i, 1$)

FUNCTION reduce($n_i, [v_1, v_2, v_3, ...]$)
   $support \leftarrow 0$
   FOR $v_i \in [v_1, v_2, v_3, ...]$ DO
      $support \leftarrow support + v_i$
   IF $support \geq MINSUPPORT$
      emit($n_i, support$)
The SZTAKI group: Research Highlights

**Recommenders: KDD Cup 2007 Task 1 First Prize**
Predict the probability that a user rated a movie in 2006, based on 2005 training data

**Spam filtering: Web Spam Challenge 1 first place, ECML/PKDD Discovery Challenge organization**

**Churn prediction: method presented at KDD Cup 2009 Workshop**
(just after 3rd place in 2008)

**Participation in ImageCLEF: Ad Hoc set of visual object and concept classification**

**TREC (ClueWeb09 - six TBs!) first participation in 2010**
IR and Web Processing

- Home developed search engine
  Hungarian Telecom 2002-, T-Mobile 2010-, AEGON Hungary 2006-
- Web Spam (LiWA FP7 ICT 2008-10)
- Multimedia IR in judicial domain (FP7 ICT 2008-10)
Analytics in a Virtual Web Observatory: Sample ideas for the task in LAWA - I

WEB LABELING AND CLASSIFICATION
Alison Hull Public Relations is a young, dynamic PR agency, specialising in the luxury hotel, spa and restaurant industry. Our particular focus is on upmarket and design-led properties.

The AHPR team brings together over 20 years experience of promoting top hotels, restaurants and spas.

4 Copenhagen Street London N1 OJD
T 020 7837 8808  F 020 7837 0559
Introduction

English | Français | Arabic

The European Commission launched a new regional project in the field of justice in the Euro-Mediterranean area, EuroMed Justice II (January 2008 – January 2011) with a budget of 5 million Euro. This project is implemented by a consortium led by European Institute of Public Administration (EIPA) and composed of the International and Ibero-American Foundation of Administration and Public Policies (FIIAPP) and of the Spanish General Council of the Judiciary (CGPJ).

implemented by

Lead firm

FIIAPP
Analytics in a Virtual Web Observatory: Sample ideas for the task in LAWA - II

WEB VISUALIZATION AND NAVIGATION: A WIKIPEDIA EXAMPLE
### Search Result

<table>
<thead>
<tr>
<th>page</th>
<th>ID</th>
<th>category</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business intelligence tools</td>
<td>1725209</td>
<td>Company, Organisation</td>
<td>0.033893</td>
</tr>
<tr>
<td>Information Builders</td>
<td>3322738</td>
<td>Company, Organisation</td>
<td>0.028129</td>
</tr>
<tr>
<td>Dimensional Insight</td>
<td>2756125</td>
<td>Company, Organisation</td>
<td>0.025022</td>
</tr>
<tr>
<td>Spotfire</td>
<td>3520914</td>
<td>Company, Organisation</td>
<td>0.025022</td>
</tr>
<tr>
<td>Talend</td>
<td>6786929</td>
<td>Company, Organisation</td>
<td>0.025022</td>
</tr>
<tr>
<td>BIRT Project</td>
<td>3569413</td>
<td>Work, Software</td>
<td>0.025022</td>
</tr>
<tr>
<td>InetSoft</td>
<td>5312704</td>
<td>Company, Organisation</td>
<td>0.025022</td>
</tr>
<tr>
<td>Palo (OLAP database)</td>
<td>3236587</td>
<td>Work, Software</td>
<td>0.025022</td>
</tr>
<tr>
<td>Prelytis</td>
<td>4280001</td>
<td>Company, Organisation</td>
<td>0.025022</td>
</tr>
<tr>
<td>Panorama Software</td>
<td>3416100</td>
<td>Company, Organisation</td>
<td>0.025022</td>
</tr>
<tr>
<td>LucidEra</td>
<td>4565541</td>
<td>Company, Organisation</td>
<td>0.025022</td>
</tr>
<tr>
<td>Data warehouse</td>
<td>5123</td>
<td></td>
<td>0.024617</td>
</tr>
<tr>
<td>Data analysis</td>
<td>1321436</td>
<td></td>
<td>0.022493</td>
</tr>
<tr>
<td>WebFOCUS</td>
<td>1128377</td>
<td></td>
<td>0.020766</td>
</tr>
<tr>
<td>Accounting intelligence</td>
<td>3081812</td>
<td></td>
<td>0.020563</td>
</tr>
</tbody>
</table>
2 Visualize, navigate
Kérdések?

Fekete Zsolt
zsfekete@ilab.sztaki.hu
Benczúr András
benczur@ilab.sztaki.hu

http://datamining.sztaki.hu