Preservation Explorer and Vault for Web 2.0 User-Generated Content

http://han.ddns.comp.nus.edu.sg/prev/

Anqi Cui¹, Liner Yang¹, Dejun Hou², Min-Yen Kan², Yiqun Liu¹, Min Zhang¹, Shaoping Ma¹
¹Tsinghua University, National University of Singapore²
PrEV and NExT

- *Preserve* the past and today’s Web 2.0 User-Generated Content (UGC) as a *Vault*, to help future researchers –
- *Explore* what was happening in our lives
Motivation: The Web

Web 1.0
"the mostly read-only Web"

250,000 sites

1996

45 million global users

Web 2.0
"the wildly read-write Web"

80,000,000 sites

2006

1 billion+ global users

> 600,000,000 sites

2.2 billion global users 2012
Reliable Services?

- The famous Twitter’s Fail Whale:

Below: The Fail Whale failed on July 27th 4am UTC
Almost 30% of recorded history shared over social media has disappeared. (SalahEldeen 2012)
Preservation: Web 1.0 => 2.0

Above: Internet Archive
Right: Country-wide:
  Norway’s Paradigma
  China’s Infomall

(Kahle 1997)

(Albertsen 2003)

(Yan 2004)
Piecemeal => A Unified Manner

(Campbell 2009, Hockx-Yu 2011)
Scenario 1: Summarizing the Data

- Ryan’s course project in 2022: Singaporean hawker center (food court) history
Scenario 1: Query-based View

Query based Topic “hawker”

Higher level: Overview

Lower level: Resource Distribution

Word cloud of popular words

Individual Resource View
Scenario 2: Historical Data Archive

- Blueberry Inc.: Develops “Follow Me” iPhone travel guide app for BJ & SG
- Review existing competitor apps’ descriptions and reviews
Scenario 2: API for Large-scale Access
System Architecture

• Three layers, loosely coupled
  – Preservation
  – Indexing
  – Interface
1. Preservation Layer

- Incoming Data Detection
- Data Format Recognition
- Record Storage
- Backup
# 1. Collected Resources (as of May 2012)

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Resource</th>
<th>No. of Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microblog messages</td>
<td>Twitter</td>
<td>229 M</td>
</tr>
<tr>
<td></td>
<td>Weibo</td>
<td>139 M</td>
</tr>
<tr>
<td>Photos with texts</td>
<td>Flickr, Panoramio</td>
<td>2 M</td>
</tr>
<tr>
<td>Food forum restaurants</td>
<td>27 Singapore sites</td>
<td>6 M (pages)</td>
</tr>
<tr>
<td></td>
<td>Fantong, Dianping (Chinese)</td>
<td>78 K</td>
</tr>
<tr>
<td>Public forum posts</td>
<td>4 Chinese forums</td>
<td>1 M (approx.)</td>
</tr>
<tr>
<td>Product review products</td>
<td>7 e-commerce sites</td>
<td>70 K</td>
</tr>
<tr>
<td>Wiki articles</td>
<td>Hudong (Chinese)</td>
<td>1 M</td>
</tr>
<tr>
<td>Traffic records</td>
<td>Singapore</td>
<td>24 K</td>
</tr>
<tr>
<td></td>
<td>Beijing</td>
<td>19 K</td>
</tr>
<tr>
<td>Question Answering articles</td>
<td>Baidu Zhidao (Chinese)</td>
<td>33 K</td>
</tr>
<tr>
<td></td>
<td>Yahoo! Answers, WikiAnswers</td>
<td>52 K</td>
</tr>
<tr>
<td>Mobile Apps</td>
<td>US App Store</td>
<td>617 K</td>
</tr>
<tr>
<td></td>
<td>Android Market</td>
<td>345 K</td>
</tr>
<tr>
<td></td>
<td>Blackberry, Windows</td>
<td>162 K</td>
</tr>
</tbody>
</table>

*Updates are of Sep 2012. Please refer to the website for an up-to-date statistics.*
2. Indexing Layer

Indexing

Preservation

Inter-
face

Indexing

Extract Content

Generate Solr Input

Solr

Metadata

Response

Query

DB

FS

Raw Data

Extracted Fields

Extracted Solr Input

Extracted Solr Input
3. Interface Layer

- Web frontend: Calendar view, word cloud, trends view, individual view
- API Frontend: Authentication, rate limiting (sandbox provided)
NExT Live UGC Web Portal

http://137.132.145.151:8080/ugcp/
What is PrEV?
- PrEV: City-centric archiving system
- Archive & unify multilingual Web 2.0 data

Whom is it for?
- Individual users: Discover the old good days
- Enterprise-level use: Programmatically access a large amount of data for business and scientific research

How is PrEV built?
- Preservation layer: Collect data from different groups
- Indexing layer: Faceted search
- Interface layer: Flexible for different needs

Future: System performance, user interfaces (visualization)
- StrmWrd: The visualization tool https://github.com/THUNUS/StrmWrd
Questions?

• Website:

http://han.ddns.comp.nus.edu.sg/prev/ or
http://tinyurl.com/prevweb
(mobile devices supported)

• Contact:
  – Anqi Cui (Google+), @CAQ9 (Twitter)
• Crawling & mining UGCs in SG and BJ in:
  – Location-oriented: shared photos and check-in venues;
  – Topic-oriented: forums, question-answering, tweets;
  – Application-oriented: mobile applications and associated information and discussions; and
  – structured: factual, cultural and historical information.

• Carrying out research projects in the areas of: Extreme Database, Live Event Capturing and Sharing, Live Media Processing, **Live Text Search**, Live City

Extreme search in text: real-time search + faceted search
  – Interesting research directions
  – Commercial and industrial applications for the next generation web

Topics to be explored includes:
  – Twitter sentiment analysis
  – Mobile app ranking
  – Social and differential news analysis