Putting people to work on the Web

Sihem Amer-Yahia
DR CNRS
Laboratoire d’Informatique de Grenoble
Sihem.Amer-Yahia@cnrs.fr
Laboratoire d’Informatique de Grenoble

24 teams
5 research axes
Sihem Amer-Yahia

- Ph.D. in CS, 1999, Univ. of Paris-Orsay & INRIA

- 1999-2006: Research Scientist, at&t labs
- 2006-2011: Senior Research Scientist, Yahoo! Labs

- 2011-2012: Research Director, QCRI

- DR1 CNRS@LIG
  - Head of the SLIDE team
The Web Today

The Web was initially a technology platform. End-users acted as consumers, think *Google Search* in its early days.

In the past 10 years, the Web became *a social milieu*. End-users act as both consumers and content providers, think *Facebook*, *Twitter*, *IMDb*.

Crowdsourcing, an online destination to find work. End-users act as workers.
*Tasks* often easier to complete by humans than by machines.
Ameryahia Mouloud & Sihem Amer–Yahia

on n'arrive pas à te suivre. Tu es où ? Ici tout va bien.

Sihem Amer–Yahia

July 26 at 12:49am · Like

Write a comment...
Outline

1. Users as content generators
   – Mostly passive content generation
   – An endless source of data for data scientists
Europe's next nightmare

The political consequences of the Eurozone crisis include increasing xenophobia and nationalism among politicians.

Dani Rodrik  Last Modified: 12 Nov 2011 10:24
User Comments on one AJE Article

#comments: 1558
Tweets

source: SocialFlow

#Al-Jazeera

#Fox News
A group of women in Morocco make a living as prostitutes in a culture that is very unforgiving toward women in that profession.

<table>
<thead>
<tr>
<th>Votes</th>
<th>Percentage</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>95</td>
<td>13.6%</td>
<td>10</td>
</tr>
<tr>
<td>32</td>
<td>4.6%</td>
<td>9</td>
</tr>
<tr>
<td>80</td>
<td>11.4%</td>
<td>8</td>
</tr>
<tr>
<td>136</td>
<td>19.4%</td>
<td>7</td>
</tr>
<tr>
<td>90</td>
<td>12.8%</td>
<td>6</td>
</tr>
<tr>
<td>38</td>
<td>5.4%</td>
<td>5</td>
</tr>
<tr>
<td>30</td>
<td>4.3%</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>1.3%</td>
<td>3</td>
</tr>
<tr>
<td>17</td>
<td>2.4%</td>
<td>2</td>
</tr>
<tr>
<td>174</td>
<td>24.8%</td>
<td>1</td>
</tr>
</tbody>
</table>

Reviews: 6 users, 2 reviews.
Grenoble on flickr

Photos de tout le monde

Voir les 237 264
Grenoble on Flickr
• Identify photos of a given city
• Sort them by creation time

• Extract Candidate POIs
  o Lonely Planet/Y! Travel/Wikipedia
  o Yahoo! Maps API

• Tag & geo-based POI association

<POI, start time, end time>
<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:00</td>
<td>Start from ground zero.</td>
</tr>
<tr>
<td>09:00</td>
<td>Spend 27 minutes at ground zero.</td>
</tr>
<tr>
<td>09:27</td>
<td>Transit to empire state building.</td>
</tr>
<tr>
<td>10:19</td>
<td>Spend 1 hour and 13 minutes at empire state building.</td>
</tr>
<tr>
<td>11:32</td>
<td>Transit to new york public library.</td>
</tr>
<tr>
<td>11:47</td>
<td>Spend 29 minutes at new york public library.</td>
</tr>
<tr>
<td>12:16</td>
<td>Transit to radio city music hall.</td>
</tr>
<tr>
<td>12:43</td>
<td>Spend 51 minutes at radio city music hall.</td>
</tr>
<tr>
<td>13:34</td>
<td>Transit to central park.</td>
</tr>
<tr>
<td>13:57</td>
<td>Spend 40 minutes at central park.</td>
</tr>
<tr>
<td>14:37</td>
<td>Transit to rockefeller center.</td>
</tr>
<tr>
<td>15:10</td>
<td>Spend 37 minutes at rockefeller center.</td>
</tr>
<tr>
<td>15:47</td>
<td>Transit to grand central terminal.</td>
</tr>
<tr>
<td>16:09</td>
<td>Spend 27 minutes at grand central terminal.</td>
</tr>
<tr>
<td>16:36</td>
<td>Transit to chrysler building.</td>
</tr>
<tr>
<td>16:42</td>
<td>Spend 31 minutes at chrysler building.</td>
</tr>
<tr>
<td>17:13</td>
<td>Transit to brooklyn bridge.</td>
</tr>
<tr>
<td>17:45</td>
<td>Spend 36 minutes at brooklyn bridge.</td>
</tr>
<tr>
<td>18:21</td>
<td>Transit to statue of liberty.</td>
</tr>
<tr>
<td>18:42</td>
<td>Spend 42 minutes at statue of liberty.</td>
</tr>
<tr>
<td>19:24</td>
<td>Transit to little korea.</td>
</tr>
<tr>
<td>19:50</td>
<td>Spend 31 minutes at little korea.</td>
</tr>
<tr>
<td>20:21</td>
<td>Transit to ground zero.</td>
</tr>
</tbody>
</table>
Outline

1. Users as content generators
   – Mostly passive content generation
   – An endless source of data for data scientists

2. Users as workers
   – Active content generation
   – A new opportunity to generate useful data
Crowdsourcing

Workers

Crowdsourcing Platform

Tasks by requesters

pybossa, foulefactory, CrowdFlower, Amazon Mechanical Turk, Crowd4U
Goal: training a learning algorithm
Receipt Transcription on Amazon MT

Classify Receipt
Hit Reward: $0.02

Real readable original receipt  Not a receipt or not readable

The following details can often be found at the top or bottom of the receipt. Enter as much information as you can find.

Find and enter the business phone number:
Phone
Example: (888) 555-1234 or 8885551234

Find and enter the business address:
Address
City
State  Postal code
Example: 321 Fake Street, Los Angeles, CA, 90210

Next
Goal: disaster management
CrowdMap from Ushahidi
Goal: Galaxy classification

Galaxy Zoo

Note: Please always classify the galaxy in the centre of the image.

SHAPE
Is the galaxy simply smooth and rounded, with no sign of a disk?

Smooth  Features or disk  Star or artifact
Goal: Social experiment (small scale)
Goal: Itinerary evaluation (large scale)

Q1: Overall, would you rate the proposed itinerary as:
— Not at all useful to a tourist
— Not so useful to a tourist
— Somewhat useful to a tourist
— Very useful to a tourist

Q2: How would you rate the set of points of interest included in the itinerary?
— Make no sense
— Mostly inappropriate
— Somewhat appropriate
— Mostly appropriate

Q3: How would you rate the visit times at the landmarks?
— Not accurate at all
— Somewhat accurate
— Mostly accurate
— Completely accurate
If you picked choices 3 or 4, did you find the visit times too short or too long?

Q4: How would you rate the transit times between the landmarks?
— Not accurate at all
— Somewhat accurate
— Mostly accurate
— Completely accurate
If you picked choices 3 or 4, did you find the transit times too short or too long?
"On the Internet, nobody knows you're a dog."
Beware of Spammers

QUALIFICATION EVALUATION

Please choose the most suitable name of the point of interest based on your experience. This would judge your fitness to take the travel itinerary evaluation task in the next section.

- Empire State Building
- Rockefeller Center
- Chrysler Building
- Flatiron Building
- Saint Patrick’s Cathedral
- Trinity Church
- Herald Square
- Washington Sq Park
- Lincoln Center
Users as content generators

• Goal: Transform user-generated content into
  – insights for data scientists, social scientists
  – useful information to end-users

• Challenges
  – Scale: (on-the-fly) analysis of Big Social Data
  – Relevance: information utility for different users
Users as workers

• Goal: Generate data for various applications
  – to train algorithms, to manage disasters, run social experiments

• Challenges
  – Profiling: detect spammers, learn workers’ skills
  – Recommendations: help requesters find workers, help workers find tasks
Users remain outside the Web

Passive and Active Content Generator  Content Consumer
Toward a user-centric Web

*Privacy*: Where is my data stored? Who is it shared with?

*Transparency*: Why am I receiving this recommendation?

*Fairness*: Am I being discriminated against?
If we knew what it was we were doing, it would not be called research, would it?

(Albert Einstein)