Building the Influence Graph

Mischa Tuffield PhD @mischatuffield
20130220
What is PeerIndex?
We reward people for their online influence ...

Jorgen Sundberg
Helping Companies with Social Media Strategy at @LinkHumans - Official LinkedIn® EMEA Talent Solutions Partner. Warning: Tweets may contain lame jokes.

Benchmark topics

<table>
<thead>
<tr>
<th>Topic</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology and Internet</td>
<td>75</td>
</tr>
<tr>
<td>Business and Finance</td>
<td>65</td>
</tr>
<tr>
<td>Arts, Media and Entertainment</td>
<td>64</td>
</tr>
<tr>
<td>News and society</td>
<td>44</td>
</tr>
<tr>
<td>Science and Environment</td>
<td>42</td>
</tr>
<tr>
<td>Leisure and Lifestyle</td>
<td>40</td>
</tr>
<tr>
<td>Sports</td>
<td>39</td>
</tr>
<tr>
<td>Health and Medical</td>
<td>36</td>
</tr>
</tbody>
</table>

Top topics

- Social Media
- Publishing
- Social Networking
- Influencer Marketing
- Facebook
- Social Media Marketing
- HR Recruiting
- Careers
- Loyalty programs
- Public Relations
- Web Development
- Direct marketing
We connect Brands with People
We are building the Influence Graph

Influencers

Jay Vyas
Nigel Shedbolt
Ferenc Huszar
Tom Rube
Riazaat Butt
Twitter
Roland Jackson
Andrew Lucas
Ian Davis
Natalie V. Rooke
Taylor Singletary
Dan Brickley

Influencing

Michael Smethurst
Natalie V. Rooke
Davide Palmisano
Ferenc Huszar
Matthew Rowe
Azeem
Andrew Lucas
Hung Lee
Shoreditch Grind
Shoreditch
Paul Gersgthy

PeerIndex
What does that mean

● What? (data)
  ○ ~100m interactions per day from FB, Twitter, LinkedIn
  ○ 3 Billion Social Actions per day
  ○ > 100m user profiles

● So What? (analysis and insight)
  ○ the influence graph
  ○ influence between people. Influence contextualised via topics
  ○ inferred or exact demographic information (gender, loc, age)
And now to explain the how ...
A Gold Standard Data Set
300 Topics Classifiers

Training used John Langfords Vowpal Wabbit (Online Logistic Regression)

Trained and Tested on 300 million URLs

Topic classification implemented in JAVA.

Can you believe that Java's String.getBytes() is not efficient!
Platform - The team are all DevOps

Elastic MapReduce
Dynamo DB
RDS
S3
Architecture

On-demand scaling at all stages and is highly modular, which together allows rapid development of resource intensive real-time analytics plugins.
Some Open Source Software
Storm - Real Time, Streams

1. Parsing
   - (a) Social Action involving URLs
     - Source
     - Parser
     - Parser

2. Crawling
   - Crawler
   - Crawler
   - Crawler

3. Data Augmentation
   - Augmentation
   - Augmentation

4. Storing
   - Storer
   - Storer
   - Storer

S3

(a) Social Action involving URLs
(b) Social Action without URLs
## Our API - March 2013

http://partners.peerindex.com/developers/our-api/

<table>
<thead>
<tr>
<th>Category</th>
<th>Endpoint</th>
<th>Look-up keys</th>
<th>Returns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actor</td>
<td>Basic</td>
<td>peerindex_id, twitter screen_name, twitter id</td>
<td>Score, name, screen_name, profile pic url.</td>
</tr>
<tr>
<td>Actor</td>
<td>Extended</td>
<td>peerindex_id, twitter screen_name, twitter id</td>
<td>Basic info, age, gender, location, bio.</td>
</tr>
<tr>
<td>Actor</td>
<td>Topic</td>
<td>peerindex_id, twitter screen_name, twitter id</td>
<td>Score, name, screen_name, pic, Topics.</td>
</tr>
<tr>
<td>Actor</td>
<td>Graph</td>
<td>peerindex_id, twitter screen_name, twitter id</td>
<td>Score, name, screen_name, pic, Influence Graph</td>
</tr>
</tbody>
</table>
Come to London Web Submit's HackDay

We will have our latest API there to be played with
We will be Open Sourcing Some Stuff

Our SeDer for HIVE

  We have tried all of them, they all have some issue or another.

Our Storm Web Crawler

  With rate limiting, and politeness code

HIVE_TEST and Improvement to MrUnit

Java VW Model Classifier
Questions ?