User Generated Content: How Good is It?

Ricardo Baeza-Yates
VP, Yahoo! Research
Barcelona, Spain & Santiago, Chile

Disclaimer

• I do not work in Web credibility
  – Thanks to Marteen de Rijke, Carlos Castillo, Miriam Metzger, ...

• Facts vs. Beliefs
  – I believe I am credible
Agenda

• What is on the Web?
  – People: The Law of Large Numbers

• Credibility as a concept

• Credibility though a Search Engine:
  – Spam detection
  – Content quality
  – Blog credibility

• Concluding Remarks

Different Views on Data
The Web

• Largest public repository of *data*
  (more than 20 billion static pages?)

• Today, there are more than 230 million Web servers
  (Apr 09) and more than 625 million hosts (Jan 09)

• Well connected graph with out-link and in-link power
  law distributions

\[ \log y = -\beta \log x \]

Self-similar &
Self-organizing

The Structure of the Web
The Wisdom of Crowds

- James Surowiecki, a *New Yorker* columnist, published this book in 2004
  - “Under the right circumstances, groups are remarkably intelligent”
- Importance of diversity, independence and decentralization

Aggregating data

“large groups of people are smarter than an elite few, no matter how brilliant—they are better at solving problems, fostering innovation, coming to wise decisions, even predicting the future”.

The Wisdom of Crowds

- Crucial for Search Ranking
- Text: Web Writers & Editors
  - not only for the Web! (TF-IDF)
- Links: Web Publishers
- Annotations: Web 2.0 Users
  - Tags, bookmarks, comments, ratings, etc.
- Queries: All Web Users!
  - Queries and actions (or no action!)
Credibility

- Expertise & trustworthiness
  - Source attractiveness & dynamism

- Leveling effect of the Web (Burbules 98)

- Checklists:
  - accuracy, authority (reputation), objectivity, currency, coverage

Web Credibility

- Web sites:
  - Presentation, information, motives, reputation

- Free or Fee based?

- Techniques (Meola 2004, Metzger 2007):
  - Promotion, consistency, corroboration
  - Seal programs, rating systems, digital signatures
Web Credibility

- How much we can automate?
- Sources of evidence?
- Example: Link analysis
  - Credible sites
  - Credibility diffusion
  - Weighted PageRank or HITS

Fight Spam

- Adversarial Web Retrieval
- Text Spam (e.g. Cloaking)
- Link Spam (e.g. Link Farms)
- Metadata spam
- Ad spam (e.g. Clicks, Bids)
Many world-class athletes, from all sports, have the ability to get in the right state of mind and when looking for women looking for love the state of mind is most important. [...] You should have the same attitude in looking for women looking for love and we make it easy for you.

Many world-class athletes, from all sports, have the ability to get in the right state of mind and when looking for texas boxer dog breeders the state of mind is most important. [...] You should be thinking the same when you are looking for texas boxer dog breeders and we make it easy for you.
Spam detection

• Machine-learning approach --- training

Web Pages \rightarrow Features
\[0.3\]
\[0.9\]
\[1.7\]
\[4.5\]
\[3.2\]
\[0.0\]

\rightarrow Machine Learning System (ML)

Learning

Training Labels

Content-based spam detection

• Machine-learning approach --- prediction

New Web Pages \rightarrow Features
\[0.3\]
\[0.9\]
\[1.7\]
\[4.5\]
\[3.2\]
\[0.0\]

\rightarrow Machine Learning System (ML)

Predictions
Normal
Spam
Dataset: Web Spam Challenge

• Label “spam” nodes on the host level
  – agrees with existing granularity of Web spam
• Based on a crawl of .uk domain from May 2006
• 77.9 million pages
• 3 billion links
• 11,400 hosts

Collaborative effort

• 20+ volunteers tagged a subset of host
• Labels are “spam”, “normal”, “borderline”
• Hosts such as .gov.uk are considered “normal”
• In total 2,725 hosts were labelled by at least two judges
• hosts in which both judges agreed, and “borderline” removed
• Dataset available at
  http://www.yr-bcn.es/webspam/
Dependencies among spam nodes

- Spam nodes in out-links
- Spam nodes from in-links
**Exploiting dependencies**

**Stacked learning**

- First pass:

<table>
<thead>
<tr>
<th>Baseline</th>
<th>in</th>
<th>out</th>
<th>both</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positive rate:</td>
<td>78.7%</td>
<td>84.4%</td>
<td>78.3%</td>
</tr>
<tr>
<td>False positive rate:</td>
<td>5.7%</td>
<td>6.7%</td>
<td>4.8%</td>
</tr>
<tr>
<td>F-Measure:</td>
<td>0.723</td>
<td>0.733</td>
<td>0.742</td>
</tr>
</tbody>
</table>

- Second pass:

<table>
<thead>
<tr>
<th>Baseline</th>
<th>1st pass</th>
<th>2nd pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positive rate:</td>
<td>78.7%</td>
<td>85.2%</td>
</tr>
<tr>
<td>False positive rate:</td>
<td>5.7%</td>
<td>6.1%</td>
</tr>
<tr>
<td>F-Measure:</td>
<td>0.723</td>
<td>0.750</td>
</tr>
</tbody>
</table>

---

**Quality and Frequency**

Quality and Quantity

The Push for Quality

User-generated

Traditional publishing
Resolved Question

What's the best way to get telemarketers off my back?

I have caller ID and usually don't answer. How can I get them to stop calling (I hear the don'tcall registry doesn't work) and if I do pick up the phone aside from immediately hanging up what can I say to deter additional calls?

1 year ago

Report it

Best Answer - Chosen by Askerno

Register at the online do not call registry. Cell phones, business and home phones can be registered. You will still get some calls for about 30 days. Just tell anyone who calls in that time period that you are registered with the do not call registry and to please remove you from their calling list. If they give you any hassle advise them that you will file a report.

I had to do this too and every solicitor I spoke to was immediately ready to get off the phone and apologized quickly. Keep a log next to your phone for the first 30 days and file it with your phone bill after that. You will then have a
¼ questions want an opinion: informal polls

¾ questions seek for information or advice

17%-45% of answers were correct.

65%-90% of questions had at least one correct answer.

There are top contributors ... 

... but they don't have all the answers.
Question quality and answer quality are not independent and can be predicted reasonably well (Castillo et al., 2008).

Predict high-quality questions

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-grams (N)</td>
<td>65%</td>
<td>48%</td>
<td>0.52</td>
</tr>
<tr>
<td>N+ text analysis</td>
<td>76%</td>
<td>65%</td>
<td>0.65</td>
</tr>
<tr>
<td>N+ clicks</td>
<td>68%</td>
<td>57%</td>
<td>0.58</td>
</tr>
<tr>
<td>N+ relations</td>
<td>74%</td>
<td>65%</td>
<td>0.66</td>
</tr>
<tr>
<td>All</td>
<td>79%</td>
<td>77%</td>
<td>0.76</td>
</tr>
</tbody>
</table>
Credibility in Blogs

- Link analysis not successful yet in UGC
  - Links are social not authority-related
  - Not well understood in blogs

- But, we have local evidence
  - Textual content, structure, behavior, etc.
Text Analysis Features

– Word length, vocabulary irregularities, uncharacteristic co-occurrences
– Topicality, word and phrase frequencies
– Punctuation and grammatical errors
– Readability (e.g., Flesch-Kincaid)
– User profiles
– Value system (community based?)
– Objectivity, honesty, ...


• Blogger’s expertise and offline identity disclosure
• Name, location, credentials, affiliations, hyperlinks to others, stated competencies, mode of knowing
• Blogger’s trustworthiness and value system
• Biases, beliefs, opinions, honesty, preferences, habits, slogans
• Information quality
• Completeness, accuracy, appropriateness, timeliness, organization, match to prior expectations, match to information need
• Appeals and triggers of a personal nature
• Aesthetic appeal, literary appeal, curiosity trigger, memory trigger, personal connection
Indicators & Estimation

• Literary appeal - capitalization, emoticons, spelling, shout
• Completeness - post length
• Accuracy - semantic distance, spaminess
• Appropriateness - semantic distance, spaminess
• Timeliness - timeliness
• Match to prior exp. - spaminess
• Match to inf. need - spaminess
• Habits - regularity, consistency
• Credentials - number of comments

Results

• Based in language models
• Credibility helps improve retrieval effectiveness
• Assessors like credible posts
• Some individual indicators helpful on their own
  – Post length, comments
  – Combination most helpful
Relating Queries (Baeza-Yates, 2007)

Qualitative Analysis

<table>
<thead>
<tr>
<th>Graph</th>
<th>Strength</th>
<th>Sparsity</th>
<th>Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>Medium</td>
<td>High</td>
<td>Polysemy</td>
</tr>
<tr>
<td>Session</td>
<td>Medium</td>
<td>High</td>
<td>Physical sessions</td>
</tr>
<tr>
<td>Click</td>
<td>High</td>
<td>Medium</td>
<td>Multitopic pages</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Click spam</td>
</tr>
<tr>
<td>Link</td>
<td>Weak</td>
<td>Medium</td>
<td>Link spam</td>
</tr>
<tr>
<td>Term</td>
<td>Medium</td>
<td>Low</td>
<td>Term spam</td>
</tr>
</tbody>
</table>

q1 ← q2

common words

q3
q4

common session

common clicks

common terms

queries

clicks

pages

links
Implicit Folksonomy?
Identical sets: **equivalence**

Subsets: **specificity**

- directed edges

Non empty intersections (with threshold)

- degree of relation

Dual graph: URLs related by queries

- High degree: multi-topical URLs

**Implicit Knowledge? Webslang!**

Baeza-Yates & Tiberi
ACM KDD 2007
• A simple measure of similarity among queries using ODP categories

- Define the similarity between two categories as the length of the longest shared path over the length of the longest path

- Let $c_1, \ldots, c_k$ and $c'_1, \ldots, c'_k$ be the top $k$ categories for two queries. Define the similarity (@$k$) between the two queries as $\max\{\text{sim}(c_i, c'_j) \mid i, j = 1, \ldots, K\}$

• Suppose you submit the queries “Spain” and “Barcelona” to ODP.

• The first category matches you get are:

  - Regional/ Europe/ Spain

  - Regional/ Europe/ Spain/ Autonomous Communities/ Catalonia/ Barcelona

• Similarity @1 is 1/2 because the longest shared path is “Regional/ Europe/ Spain” and the length of the longest is 6
Experimental Evaluation

- We evaluated a 1000 thousand edges sample for each kind of relation.
- We also evaluated a sample of random pairs of not adjacent queries (baseline).
- We studied the similarity as a function of $k$ (the number of categories used).
Open Issues

• Explicit vs. implicit social networks
  – Any fundamental similarities?

• How to evaluate with partial knowledge?
  – Data volume amplifies the problem

• User aggregation vs. personalization
  – Optimize common tasks
  – Move away from privacy issues

Concluding Remarks

• Social Science 2.0 (Duncan Watts)
  – We can use large samples (e.g. social networks, Mturk)
  – What do you believe might be wrong

• Social reputation & leadership
  – Correlated with credibility
Influence Leadership (Bopal et al, 2008)

- Influence of social graph in particular actions
  - Social graph: Yahoo! Instant Messenger
  - Actions log: Yahoo! Movies
    - Action = user $u$ rated movie $m$ at time $t$
  - joined through common users identifiers
- Started from Yahoo! Instant Messenger subgraph of “most active” users (110M nodes) and 21M ratings from Yahoo! Movies.
  - Ended with 217.5K nodes, 221.4K edges and 1.8M ratings.

Leaders vs. Tribe leaders

![Graph showing the comparison between Leaders and Tribe leaders in Yahoo! Movies, Genre, $\pi = 9$ weeks, $\sigma = 5$.]
Epilogue

- The Web is scientifically young
- The Web is intellectually diverse
- The technology mirrors the economic, legal and sociological reality
- Web Mining: large potential for many applications
  - A fast prototyping platform is needed
- Plenty of open problems

Questions?

Advertising:

SPIRE 2009
in Lappland
Deadline: May 1st

Web Credibility

Ricardo Baeza-Yates
Yahoo! Research
Barcelona, Spain & Santiago, Chile
ricardo@baeza.cl