Content-Driven Reputation and Trust on the Web

Luca de Alfaro
Jack Baskin School of Engineering
UC Santa Cruz

Research Review Day
October 2008
User Created Content on the Web

Wikipedia, YouTube, Picasa, Flickr, Blogs, Forums, Discussion groups, user-editable maps, ...

• If anyone can create it, how can we separate the wheat from the chaff?
  
  - By looking at the content (spam detection, etc.)
  
  - By looking at the authors, and at their interactions. If the authors are good, perhaps so is the content?

• In this talk we focus on the Wikipedia, and on wikis.
Italian cuisine is extremely varied: the country of Italy was only unified in 1861, and its cuisines reflect the cultural variety of its regions and its diverse history (with culinary influences from Greek, Roman, Norman and Arab civilizations). Italian cuisine is imitated all over the world. It also is way better than French food, the losers. To a certain extent, there is really no such thing as
Italian cuisine is extremely varied: the country of Italy was only unified in 1861, and its cuisines reflect the cultural variety of its regions and its diverse history (with culinary influences from Greek, Roman, Norman and Arab civilizations). Italian cuisine is imitated all over the world. It also is way better then French food, the losers.

To a certain extent, there is really no such thing as
The WikiTrust Project:

- Authors are assigned **reputation** based on their contributions
- Text is assigned **trust** based on the reputation of the author and of the revisors.
- Can be integrated in any wiki, any language.
Author Reputation + Text Trust

Author Reputation:
• Promote constructive behavior
• Provide information on author reliability

Text Trust:
• Give a guide to text reliability
• Provide alert for attempts to tamper with content
Content-driven reputation

- Authors of long-lived contributions gain reputation
- Authors of reverted contributions lose reputation
Content-driven reputation

- Authors of long-lived contributions gain reputation
- Authors of reverted contributions lose reputation
Content-driven reputation

- Authors of long-lived contributions gain reputation
- Authors of reverted contributions lose reputation

A wiki article

A

edits

B

builds on A’s edit
Content-driven reputation

- Authors of long-lived contributions gain reputation
- Authors of reverted contributions lose reputation
Content-driven reputation

- Authors of long-lived contributions gain reputation
- Authors of reverted contributions lose reputation
Content-driven reputation

• **No change to wiki user experience**
  - Transparent to casual users
  - No need to explicitly rate others and be rated, less stress

• **Everybody votes (via their edits)**
  - All users count.
  - Robust wrt. attacks.

• **Papers:**
  - **WWW 07**: First proposal, evaluation on Wikipedia
  - **AISec 08**: How to make it robust to attacks, including sock-puppet attacks.
Goals of reputation in Wikipedia

• **Prescriptive:** encourages people to behave in a good way (e.g., Ebay system).
  - We want to encourage lasting contributions.

• **Descriptive:** gives information to users (e.g., Pagerank, Ebay system).
  - Author reputation can be used as a rough guide to the trust in new text/edits.

• **Predictive:** Is reputation a good predictor for future behaviour? Few systems make this claim!
  - We use this as our evaluation criterion, and we show that our reputation can predict edit quality.
Does our reputation have predictive value?

Article 1
Article 2
Article 3
Article 4

Time

○ = edits by user A
Does our reputation have predictive value?

The reputation of author A at the time of an edit $E$ depends on the history before the edit.

The longevity of an edit $E$ depends on the history after the edit.

We will show a correlation between author reputation and edit longevity.
Building a Content-Driven Reputation System

At the heart, we have a very efficient text diff engine:

- **Block ops**: insert $I(n,k)$, delete $D(n,k)$, move $M(i,j,k)$
- **Greedy longest-common-block matching.**
- **Fast!** (Many hundreds of page pairs / sec).
- From the list of block ops, we can compute the **distance** between versions.

At the heart, we have a very efficient text diff engine:

- **Block ops**: insert $I(n,k)$, delete $D(n,k)$, move $M(i,j,k)$
- **Greedy longest-common-block matching.**
- **Fast!** (Many hundreds of page pairs / sec).
- From the list of block ops, we can compute the **distance** between versions.
Edit quality

\[ d(i, k) > d(j, k) \]

\[ d(i, k) < d(j, k) \]

\( j \) is good: \( j \) went towards the future

\( j \) is bad: \( j \) went against the future
Edit quality measures the fraction of change that agrees with the future page evolution.

- $\alpha_{\text{edit}} \sim +1$: edit $j$ is good
- $\alpha_{\text{edit}} \sim -1$: edit $j$ is reverted
From edit quality to reputation update

Repuation update:

\[ \text{rep}(A_j) = \text{rep}(A_j) + \alpha_{edit} \cdot w(\text{rep}(A_k)) \]

For a reputation \( r \),

\[ w(r) = \log(1 + r) \] is its “weight”.

Edit Quality:

\[ \alpha_{edit} = \frac{d(i, k) - d(j, k)}{d(i, j)} \]
Robust Reputation

By choosing appropriately the set of revisions to be compared, and by introducing rules to cap reputation increase, we ensure:

- **Theorem 1 (robustness):** A set of identities of reputation at most \( r \) cannot gain reputation above \( r \) without doing useful work (work that is considered of positive quality by people of reputation above \( r \)).

- **Theorem 2 (truthfulness):** If an author \( A \) wishes to do an edit \( E \), the best is to just do \( E \): author \( A \) cannot gain more reputation by splitting \( E \) in multiple pieces, or playing other tricks.
Results: English Wikipedia, in detail

% of edits below a given quality

<table>
<thead>
<tr>
<th>Bin</th>
<th>%_data</th>
<th>l&lt;0.8</th>
<th>l&lt;0.4</th>
<th>l&lt;0.0</th>
<th>l&lt;-0.4</th>
<th>l&lt;-0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>16.922</td>
<td>93.11</td>
<td>91.65</td>
<td>89.15</td>
<td>83.76</td>
<td>73.53</td>
</tr>
<tr>
<td>1</td>
<td>1.191</td>
<td>77.24</td>
<td>69.83</td>
<td>65.60</td>
<td>61.11</td>
<td>56.00</td>
</tr>
<tr>
<td>2</td>
<td>1.335</td>
<td>69.53</td>
<td>57.08</td>
<td>49.79</td>
<td>45.71</td>
<td>41.25</td>
</tr>
<tr>
<td>3</td>
<td>1.627</td>
<td>38.00</td>
<td>28.61</td>
<td>20.23</td>
<td>16.16</td>
<td>13.62</td>
</tr>
<tr>
<td>4</td>
<td>2.780</td>
<td>32.84</td>
<td>22.31</td>
<td>13.32</td>
<td>9.57</td>
<td>8.04</td>
</tr>
<tr>
<td>5</td>
<td>4.408</td>
<td>41.70</td>
<td>15.76</td>
<td>5.90</td>
<td>3.80</td>
<td>2.57</td>
</tr>
<tr>
<td>6</td>
<td>6.698</td>
<td>29.40</td>
<td>16.74</td>
<td>7.54</td>
<td>4.35</td>
<td>3.12</td>
</tr>
<tr>
<td>7</td>
<td>8.281</td>
<td>32.04</td>
<td>15.16</td>
<td>5.44</td>
<td>2.25</td>
<td>1.40</td>
</tr>
<tr>
<td>8</td>
<td>12.233</td>
<td>34.06</td>
<td>16.64</td>
<td>6.78</td>
<td>3.79</td>
<td>2.73</td>
</tr>
<tr>
<td>9</td>
<td>44.524</td>
<td>32.55</td>
<td>15.51</td>
<td>5.05</td>
<td>1.88</td>
<td>1.14</td>
</tr>
</tbody>
</table>
## Results: English Wikipedia, in detail

<table>
<thead>
<tr>
<th>Bin</th>
<th>% data</th>
<th>l&lt;0.8</th>
<th>l&lt;0.4</th>
<th>l&lt;0.0</th>
<th>l&lt;-0.4</th>
<th>l&lt;-0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>16.922</td>
<td>93.11</td>
<td>91.65</td>
<td>89.15</td>
<td>83.76</td>
<td>73.53</td>
</tr>
<tr>
<td>1</td>
<td>1.191</td>
<td>77.24</td>
<td>69.83</td>
<td>65.60</td>
<td>61.11</td>
<td>56.00</td>
</tr>
<tr>
<td>2</td>
<td>1.335</td>
<td>69.53</td>
<td>57.08</td>
<td>49.79</td>
<td>45.71</td>
<td>41.25</td>
</tr>
<tr>
<td>3</td>
<td>1.627</td>
<td>38.00</td>
<td>28.61</td>
<td>20.23</td>
<td>16.16</td>
<td>13.62</td>
</tr>
<tr>
<td>4</td>
<td>2.780</td>
<td>32.84</td>
<td>22.31</td>
<td>13.32</td>
<td>9.57</td>
<td>8.04</td>
</tr>
<tr>
<td>5</td>
<td>4.408</td>
<td>41.70</td>
<td>15.76</td>
<td>5.90</td>
<td>3.80</td>
<td>2.57</td>
</tr>
<tr>
<td>6</td>
<td>6.698</td>
<td>29.40</td>
<td>16.74</td>
<td>7.54</td>
<td>4.35</td>
<td>3.12</td>
</tr>
<tr>
<td>7</td>
<td>8.281</td>
<td>32.04</td>
<td>15.16</td>
<td>5.44</td>
<td>2.25</td>
<td>1.40</td>
</tr>
<tr>
<td>8</td>
<td>12.233</td>
<td>34.06</td>
<td>16.64</td>
<td>6.78</td>
<td>3.79</td>
<td>2.73</td>
</tr>
<tr>
<td>9</td>
<td>44.524</td>
<td>32.55</td>
<td>15.51</td>
<td>5.05</td>
<td>1.88</td>
<td>1.14</td>
</tr>
</tbody>
</table>

**Short-Lived**
From author reputation to text trust
<table>
<thead>
<tr>
<th>existing</th>
<th>text</th>
</tr>
</thead>
</table>
Trust: New Text

The color of new text is proportional to the author's reputation.
Even top-reputation authors cannot single-handedly create trusted text: trust always requires consensus.
Trust: Rearranging Blocks of Text

At every border between new neighbours, the text has the same trust as new text. The trust gradually returns to the original value as the distance from the disrupted border increases.
Trust: Text revision effect

![Graph showing trust changes over text revision effect]

- max trust
- trust 0
- text
- trust of existing text
Trust: Text revision effect

max trust

trust 0

text

trust of existing text

author reputation
If text has trust lower than the author's reputation, we update it as follows:

\[ trust_{\text{new}} = trust_{\text{old}} + \alpha \cdot (\text{reputation} - trust_{\text{old}}) \]
Cheerleading and word spotting

History

The real starter of cheerleader was Marte Klopstad! She was a happy little child who loved to dance and play with dolls. Cheerleading first started at Princeton University in the 1880s with the crowd chant, "Rah rah rah, tiger lion bear, sis sis sis, boom boom boom boom ahhhhhhh, Princeton Princeton Princeton!!" as a

Many tell us that we should use semantic analysis as well. But the present method is simpler, and language independent.
Finding non-obvious tampering

departments. Cabinet members are occasionally recruited from outside the Folketing.

Since 27 November 2001, the economist Anders Fogh Rasmussen has been Prime Minister to Denmark.

As known in other parliamentary systems of government, the executive,

The correct spelling is Fogh.
In Danish, “fjog” means “fool” or “goofy”
Low trust predicts text deletion

- **Recall wrt. deletions:** Text in the bottom half of trust values constitutes 3.4% of the text, yet corresponds to 66% of the text that is deleted in the next revision.

- **Precision wrt. deletions:** Text in the bottom half of trust values has a probability of 33% of being deleted in the very next revision, compared with 1.9% for general text. The probability raises to 62% for text in the bottom fifth of trust values.

Data obtained by analyzing 1,000 articles selected at random among those with at least 200 revisions.
Trust predicts text lifespan

- Expected life (n. of revisions)
  - Word trust

Graph showing the expected life of a text based on word trust.
WikiTrust: A Reputation/Trust System for Wikis

http://trust.cse.ucsc.edu/WikiTrust
(open source, BSD license)

Thanks to (in alphabetical order):
Bo Adler, Krishnendu Chatterjee, Ian Pye