ISSDM Collaborative Research Program Proposal

Project Title: Bringing Trust to Collaborative Content

Principal Investigator: Luca de Alfaro
Email: luca@soe.ucsc.edu
Phone: (831) 459-4982
Web: `http://www.soe.ucsc.edu/~luca`
Computer Engineering Dept.
University of California, Santa Cruz

On-line collaboration is fast becoming one of the most effective ways in which information is being created and shared. The poster child of this phenomenon is the Wikipedia, a collaborative encyclopedia which is becoming one of the most commonly used sources of information in the world.

Effective collaboration benefits from two mechanisms: incentive systems that encourage constructive behavior, and trust systems that help information users judge the reliability of the information resulting from the collaborative process. Specifically, we will develop a reputation system for authors, and a trust system for content. The reputation system will be driven by content evolution, and it will give higher reputation to authors who provide the most long-lived, and thus useful, contributions. The trust system will compute a value of trust for the content, at word granularity: the trust of a word will depend on the reputation of its author, as well as the reputation of the authors who revised the text where the word appears. Content trust will be displayed via an intuitive, trust-based coloring of the text background (see Figure 1).

We have already implemented embryos of these systems. The embryos work on static dumps of Wikipedia information, and are not suited to dynamic interaction with visitors: still they provide a hint of their potential value. This project will develop the key elements that are required to turn these off-line embryos into working on-line systems, that can provide real-time trust information to visitors. Making the systems suitable to on-line deployment requires the investigation, and solution, of two main research challenges, which form the technical core of the project.

The first challenge consists in making the reputation and trust systems resistant to attacks. In an on-line implementation, some visitors may want to game the systems to gain reputation or to cause information of their choice to be marked as trusted. We propose to investigate methods that exploit the content-driven nature of the reputation and trust algorithms to achieve algorithms that are resistant to attack. Resistance to attacks is especially important when the information carries larger technical, political, or safety significance. The second challenge consists in developing techniques that can use the trust information to automatically select high-quality, stable versions of the content. Many information users would rather access selected versions of content that are trusted and reasonably recent, rather than the most recent versions, which may be inaccurate. The methods developed in the course of the project will be applicable to a wider context than wikis: the crucial requirement is that *users can modify previous contributions given by other users.*

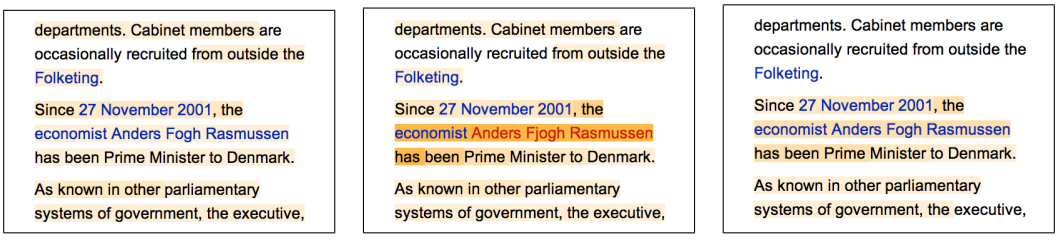

Figure 1: Three consecutive Wikipedia revisions showing an an attempt to modify the spelling of the Danish Prime Minister's last name, from Fogh, to Fjogh (in Danish, a fjog is a fool); Text with orange background is low-trust; the initial revision is 77625823. Such subtle changes can be difficult to spot without a trust coloring.

1 Project Description: Bringing Trust to Collaborative Content

We are rapidly moving into an era where information is created by communities that can span the whole world, and where the main interaction is taking place via on-line real-time cooperation. A growing proportion of the content on the Web is created in collaborative fashion: from the Wikipedia, which currently ranks as the 8^{th} most-visited web site [Ale07], to YouTube, Flickr, TripAdvisor, and to the innumerable discussion groups and wiki-powered information and documentation sites. This collaborative process of content creation makes it difficult for users to form an idea of the reliability of the information presented to them. An example of this is provided by the Wikipedia, perhaps the most successful example of collaborative content creation to date. Wikipedia articles are constantly changing, and the contributors range from domain experts to vandals, from dedicated editors to superficial contributors not fully aware of the quality standards the Wikipedia aspires to attain. Wikipedia visitors are presented with the latest version of each article: this latest version does not offer them any simple insight into how the article content has evolved into its most current form, nor does it offer a measure of how much the content can be relied upon. Unsurprisingly, the reliability of the Wikipedia has frequently been called into question in the media (see, e.g., [See05, Leh06, HR06, Dav06, Str06, Sch06, Sta07]).

To help users make sense of collaboratively-generated information, we are developing algorithmic notions of *information trust* [AdA07, ABC$^+$07]. Our prototype system for the Wikipedia displays the computed values of text trust by coloring the background of text: white corresponds to high-trust, and orange to low-trust. Against the white background of stable articles, any recent modification that has not yet been sufficiently revised stands out clearly (see the examples in Figure 2, taken from `http://trust.cse.ucsc.edu`). Our previous work has been done on the basis of archived Wikipedia data and batch analysis. In this project, we propose to investigate two issues that are central to deploying a live system, operating on data updated in real-time. The first issue consists in making the trust system resistant to attacks and manipulations by users. The second issue consists in developing algorithms for the selection and construction of high-quality content on the basis of the trust information.

The methods developed in the course of the project will be applicable to a wider context than wikis: the crucial requirement is that *users can modify previous contributions given by other users,* so that each document, or article, results from the integration of the contributions of all users who worked on it. The project focuses on wikis because, as the Wikipedia illustrates, they are arguably the most successful large-scale on-line collaboration venue to date.

1.1 Background

Our trust system for the Wikipedia computes text trust in two steps: first, the system computes a value of reputation for each author; second, the trust of text is computed on the basis of the reputation of the authors who have created and revised the text.

Content-driven reputation. We compute author reputation via content analysis, rather than on the basis of user-to-user comments. In our system, authors who perform long-lived contributions gain reputation; authors whose contributions are reverted, or removed in short order, lose reputation.

To measure the longevity of contributions, our system analyzes at word level the revision history of all wiki articles, in chronological order. Suppose that an author A performs an edit that changes an article from a version v_{i-1} to a version v_i. When a subsequent author B perform an edit, producing in turn a version v_j, $j > i$, we update A's reputation by adding to it *text* and *edit* increments [AdA07]:

- The *text increment* is proportional to the number of words introduced in v_i that are still present in v_j.

Figure 2: Text coloring of the Italian cuisine article resulting after an anonymous author (unrelated to the PI) added a comparison of the virtues of the Italian and French cuisines (revision id: 104258868).

- If v_i is more similar to v_j than v_{i-1}, A's reputation is given a positive *edit increment,* since A's edit was "in the right direction"; otherwise, As' reputation receives a negative edit increment.

To evaluate the effectiveness of the reputation system, we considered all edits to the English Wikipedia; our results indicate that:

- *Recall.* 84.5% of reverted edits are performed by authors of low reputation.
- *Precision.* The probability that an edit performed by a low-reputation author is reverted is 74.2%.

Text trust. Our trust system assigns newly-inserted words an initial value of trust proportional to the reputation of their author. When subsequent authors edit the page, and leave the text surrounding a word unchanged, we raise the trust of the word: the fact that the text around the word was unchanged means that the subsequent authors implicitly agree with it. On the other hand, if subsequent authors modify or rearrange the text in proximity to the word, the trust in the word will in general decrease [ABC$^+$07]. A virtue of the system is that it makes it hard to maliciously and surreptitiously change the content of Wikipedia articles: every change, including text deletions, leaves a low-trust mark that fades in future revisions only as the text is further revised. Another property of our trust algorithm is that nobody can single-handedly create fully trusted text: full trust (signaled via a white background) can only be achieved via the revisions of multiple authors. We assessed the quality of the trust labeling in a data-driven way, using the idea that *trust should be a predictor for text stability* [ZAFM06]:

- *Recall.* We show that text in the lowest 50% of trust values constitutes only 3.4% of the text of articles, yet corresponds to 66% of the text that is deleted from one revision to the next.
- *Precision.* We show that text that is in the bottom half of trust values has a probability of 33% of being deleted in the very next revision, in contrast with the 1.9% probability for general text. The deletion probability raises to 62% for text in the bottom 20% of trust values.

The above results were obtained by analyzing 1,000 articles selected randomly from the English Wikipedia articles with at least 200 revisions.

1.2 Proposed Research

The goal of this proposed project is to develop the two missing elements for an effective on-line trust system: resitance to attacks, and the ability to automatically select trusted information.

Attack-resistant reputation and trust. We anticipate that as soon as our reputation and trust systems will be made available on the Wikipedia, visitors will try to find weaknesses that can be exploited to gain reputation without effort, or to cause invalid information to be inserted with a high value of trust. For the reputation system, we are primarily concerned with attacks in which users try to gain reputation without giving useful contributions. This is possible in the current system via a *Sybil attack,* in which a single user assumes various identities, using "secondary" identities to give positive feedback to a "primary" identity, that thus gains reputation [Dou02, CF05, SGJ05]. Sybil attacks are notoriously hard to defeat [Dou02, LSM06]. However, we believe that our content-driven reputation system can be made resistant to Sybil attacks by exploiting the characteristics of the content-driven algorithms we use to attribute reputation. An author can increase in reputation only as a consequence of editing pages, and pages are public: there is no way for an author to covertly give positive feedback on another author. Furthermore, there are no pages that are reserved to groups of authors: on every page, it is virtually guaranteed to have a regular influx of edits coming from the population of wiki contributors at large. We will investigate modified reputation-attribution algorithms that take advantage from these facts to thwart Sybil attacks.

The main attack mode to the trust system consists in users that try to raise the trust of a portion of text by repeatedly performing minor revisions of the article. We plan to develop a randomized signature algorithm, which tags each word of the text of the most recent revision of each article with a signature. The goal is to remember the authors who have recently raised the trust of a word, and prevent them from doing so again until a sufficient number of unrelated authors has also affected the trust of the word.

Automated selection of trustworthy information. A common desire of Wikipedia visitors is to be able to choose to see high-quality revisions of articles, rather than the latest revisions. High-quality revisions are better suited to environment such as schools, where absence of spam and inappropriate language is prized. Search engines would also benefit, in their indexing of Wikipedia content, from the ability to index high-quality revisions, rather than the latest revisions, which may be incomplete or which may contain unrelated content. The need for automated trusted information selection is in fact a very general one, extending beyond the particular setting of the Wikipedia.

We propose to develop algorithms for the automated selection and creation of high-quality article revisions, on the basis of the trust information for the article words. When selecting high-quality revisions, the naïve idea of looking at the minimum trust value of words may not work well in practice, as it does not differentiate between revisions that are mostly high-trust, but have a small recently-edited portion of intermediate trust, and revisions in flux, where most of the text is of intermediate trust values. Similarly, basing the choice on average trust may not work well, as it fails to differentiate among high-trust revisions before, and after, a small amount of low-trust spam has been added. Rather, we propose to select high-quality revisions by analyzing the revision history of articles. As articles are edited, there are usually periods in their history with consecutive revisions that consist of text that is high trust, except for small portions, which have been subject to minor edits. These small low-trust portions may become high-trust in later revisions. Thus, by comparing these consecutive revisions, it may be possible to infer that a revision is overall high trust, using the future revisions to validate the small low-trust portions.

References

[ABC+07] B.T. Adler, J. Benterou, K. Chatterjee, L. de Alfaro, I. Pye, and V. Raman. Assigning trust to wikipedia content. Technical Report UCSC-CRL-07-09, School of Engineering, University of California, Santa Cruz, CA, USA, 2007.

[AdA07] B.T. Adler and L. de Alfaro. A content-driven reputation system for the Wikipedia. In *Proc. of the 16th Intl. World Wide Web Conf. (WWW 2007)*. ACM Press, 2007.

[Ale07] Dec. 2, 2007. `http://www.alexa.com`.

[CF05] A. Cheng and E. Friedman. Sybilproof reputation mechanisms. In *Proc. of the ACM SIGCOMM workshop on Economics of peer-to-peer systems*. ACM Press, 2005.

[Dav06] M. Davis. Congress "made Wikipedia changes". *BBC News*, Feb. 9, 2006.

[Dou02] J.R. Douceur. The sybil attack. In *Peer-to-Peer Systems: First Intl. Workshop*, volume 2429 of *Lect. Notes in Comp. Sci.*, pages 251–260, 2002.

[HR06] M. Hickman and G. Roberts. Wikipedia — separating fact from fiction. *The New Zealand Herald*, Feb. 13 2006.

[Leh06] E. Lehmann. Rewriting history under the dome. *The Sun*, Jan. 27, 2006.

[LSM06] B.N. Levine, C. Shields, and N.B. Margolin. A survey of solutions to the sybil attack. Technical Report Technical Report 2006-052, Univ. of Massachussets Amherst, 2006.

[Sch06] S. Schiff. Know it all: Can Wikipedia conquer expertise? *The New Yorker*, Jul. 31, 2006.

[See05] K.Q. Seelye. Snared in the web of a Wikipedia liar. *The New York Times*, Dec. 4, 2005.

[SGJ05] J.-M. Seigneur, A. Gray, and C.D. Jensen. Trust transfer: Encouraging self-recommendations without sybil attack. In *Trust Management*, volume 3477 of *Lect. Notes in Comp. Sci.* Springer-Verlag, 2005.

[Sta07] BBC Staff. Fake professor in Wikipedia storm. *BBC News*, Mar. 6, 2007.

[Str06] R. Stross. Anonymous source is not the same as open source. *The New York Times*, Mar. 12, 2006.

[ZAFM06] H. Zeng, M. Alhossaini, R. Fikes, and D.L. McGuinness. Mining revision history to assess trustworthiness of article fragments. In *Proc. of the 2nd Intl. Conf. on Collaborative Computing: Networking, Applications, and Worksharing (COLLABORATECOM)*, 2006.