CredibleWeb: A Platform for Web Credibility Evaluation

Zhicong Huang
EPFL EDOC-IC
INR 012
CH-1015 Lausanne Switzerland
zhicong.huang@epfl.ch

Alexandra Olteanu
EPFL IC LSIR
BC 147
CH-1015 Lausanne Switzerland
alexandra.olteanu@epfl.ch

Karl Aberer
EPFL IC LSIR
BC 108
CH-1015 Lausanne Switzerland
karl.aberer@epfl.ch

CHI 2013 Extended Abstracts, April 27–May 2, 2013, Paris, France.
ACM 978-1-4503-1952-2/13/04.

Abstract

The web content is the main source of information for many users. However, due to the open nature of today's web anyone can produce and publish content, which, as a result, is not always reliable. As such, mechanisms to evaluate the web content credibility are needed. In this paper, we describe *CredibleWeb*, a prototype crowdsourcing platform for web content evaluation with a two-fold goal: (1) to build a social enhanced and large scale dataset of credibility labeled web pages that enables the evaluation of different strategies for web credibility prediction, and (2) to investigate how various design elements are useful in engaging users to actively evaluate web pages credibility. We outline the challenges related with the design of a crowdsourcing platform for web credibility evaluation and describe our initial efforts.

Author Keywords

Web Credibility; Crowdsourcing; Gamification; Recommendations

ACM Classification Keywords

H.5.m [**Information interfaces and presentation (e.g., HCI)**]: Miscellaneous.

Introduction

The Web can be seen as a double-edged sword: while it provides people rich information, it is also exposes spurious and malicious content due to its' open nature. This problem becomes even more crucial when people rely on online information in their day by day decisions (e.g., health related). While, as shown by the the Pew Research Center's report, 81% of american adult population have used the web[1], number that grows to 93% for teens [2], there is not systematic way to moderate the web content. For classic publications "the cost of printing acted as a barrier" [6], yet the web enabled everyone with access to a computer and an internet connection to publish and disseminate content with low or no cost.

In this context, several studies with the purpose of helping users to evaluate the credibility of online information [11, 14], or even to automate the credibility assessments have emerged [5, 12, 9]. The main drawback in testing these systems is the lack of a large scale and comprehensive dataset to validate the generality of the proposed approaches. For instance, when employing a machine learning based approach one usually needs training and test sets, which are composed of a set of web pages labeled with "ground truth" credibility ratings. Alas, most of the studies use only small datasets (no more than 1000 web pages [9]) with respect to the size and diversity of the web content. This leads to models with poor predictive performance that overfit the training data.

A popular way to ensure the quality of the published content is peer-reviewing. Such systems embed the "wisdom of crowds" concept and are widely used to ensure the quality of scholarly publications. However, employing such a system at the scale of today's web is not straightforward and, probably, yet unpractical: (1) *reviewer selection* – while in these systems the reviewing process is ensured by volunteers that are validated by topical-communities as experts and are interested in maintaining high standards within their communities, on the Web it is harder to identify and incentivize experts to review and endorse the content; (2) *economic barrier* – is an important aspect since an expert can evaluate only a limited and rather small amount of information; to scale such a system the aggregation of evaluation from a wider range of users needs to be considered; (3) *various metrics* – the web involves skewed service interest and long-tail content. As such, the information credibility assessment varies depending on the topic, user involvement, task, knowledge, context, etc.; to collect reliable evaluations it is important to offer users proper incentives to evaluate to the best of their knowledge.

Although few independent efforts that allow users to evaluate the content at the phrase/sentence-level have been made in this direction [4, 2], they are still at an incipient stage. We take a similar direction, but, in contrast, we look at the overall credibility of a web page.

In this paper, we introduce *CredibleWeb* a social enhanced crowdsourcing platform for web content evaluation on which users can rate web pages, share comments, vote other users comments to gain points and reputation, and consult others' opinions about content of interest. *CredibleWeb* has as main goal the building of a large scale dataset that enables the evaluation of various types of recommendation and prediction strategies when used for automatic assessment of web pages credibility. To this end, we highlight the data requirements and describe how

[1] http://pewinternet.org/Trend-Data-(Adults)/Internet-Adoption.aspx, 05.01.2013

[2] http://pewinternet.org/Static-Pages/Trend-Data-(Teens)/Internet-Access.aspx, 05.01.2013

we design our prototype in order to meet these requirements. The major contribution of this work stays in the design of a platform that (1) aims to collect both the users credibility evaluations and the social relations between them and (2) employs game mechanisms to engage users to rate webpages by the means of our platform.

System Design

In this section we detail the requirements imposed by our data acquisition goals and discuss how we address them through *CredibleWeb*'s initial design. We group the design points by the key challenges that a crowdsourcing platform design needs to answer [8]:

Zhicong Huang
Profile
35 points

Figure 1: *Points awarding.* On *CredibleWeb* users can earn points for their actions (e.g., rating web pages, making friends, submitting URLs, etc.

What Data to Gather?

The main goal of this work is to build a comprehensive and large scale dataset to enable the evaluation of various state of the art prediction and recommendation tasks when used to automatically assess web pages credibility. Most of the tasks that we account for have the following basic data requirements:

A set of webpages. We start with an initial collection composed of two sets of webpages: one collected by the authors of [11], and a mirroring set build in the same way but at a latter date. These add up to around 1700 URLs falling into 5 different categories (e.g., health, finance, politics, celebrities, environment). To extend this collection we enable users to submit URLs of web pages of their choice to be evaluated on our platform.

A set of evaluations per web page. To gather multiple evaluations per each web page and meet different data needs, we employ several recommendation strategies to assign users web pages for evaluation. (1) For instance, to use a machine learning approach to automatically predict the web pages credibility one needs reliable (i.e., “ground truth”) evaluations to avoid building predictive models that encompass possible evaluations bias. In this end, we keep recommending a web page to users for evaluation until we observe an agreement among raters. When such an agreement cannot be reached due to polarized opinions after gathering a significant number of ratings, the system can label the web page with two different credibility scores. In addition, we also recommend users to rate web pages that match their declared expertise, under the assumption that expert users make more reliable evaluations. (2) Another way to predict the credibility of an unseen web page by a certain user is to make use of collaborative filtering techniques. Yet, to successfully predict a web page credibility score, these techniques need a reasonable number of evaluations per user and per item. In this regard, we recommend users to evaluate scarcely rated web pages.

The social connections among users. Trust or social-based recommendation strategies exploit the relations between users in order to make predictions. To this end, besides creating a new account on our platform, we also allow users to sign in on our platform with various social network accounts (e.g., Facebook, Twitter) and to merge them. This allows the building of a multi-layered social graph in which each layer encompasses information from a certain social network, and enables one to evaluate how different networks influence the recommendation performance, e.g., interest-based (Twitter) vs. acquaintances-based (Facebook) networks.

How to recruit contributors?

Contributors recruitment is an important aspect of a crowdsourcing system [8]. We plan to start with a initial and relatively small set of *seed* users that volunteer to use

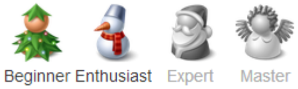


Figure 2: Achievement levels are a reputation measure and reflect the number of points a user gained over time.

Figure 3: Leaderboard. The big red point represents the current user, while the small grey ones represents the top users.

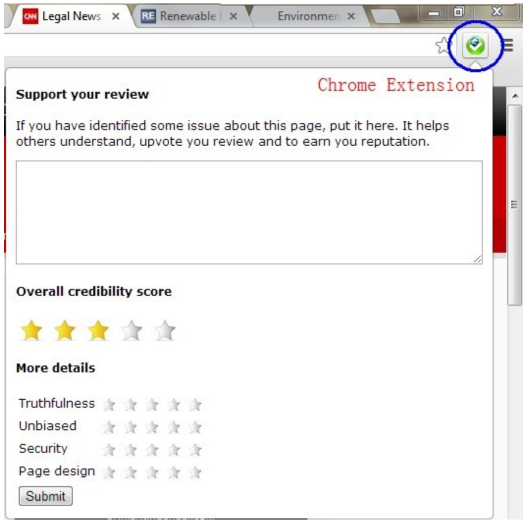
the platform and advertise it within their social entourage. From behavioural economics we know that when trying to make people change or adopt a habit, other people behaviour plays an important role [7]. As such, we target three main types of users as seed users [7]: (1) *experts* whose advice would be taken by most of the other users, (2) *popular users* which have a large number of friends, and, thus, have a good dissemination potential, and (3) *salesmen* which are users that know how to persuade others to adopt the platform.

How to Retain Users?
Even though to recruit seed users we mainly rely on members of our department, the main downside of volunteering lays in the inability to predict the number of users that will join and adopt the system. To this end, we employ several state-of-the-art strategies to retain and encourage users to contribute evaluations.

First, we employ *gamification* techniques to engage users. Gamification has been successfully used in human-based computation task, with examples such as *Stack Overflow* [3] and *ESP Game* [13]. Hereof, we use three gamification techniques: *points awarding* (Figure 1), *achievement levels* (Figure 2), and *leaderboards* (Figure 3) to measure and highlight users reputation.

Second, it is known that people are motivated to "do the right thing" and they need to feel effective in order to change their behaviour [7]. As such, we try to provide *instant gratification* to users by showing them how their contribution is used to compute the credibility of the web pages they rate [8].

Third, to make it easy for users to contribute we both (1) employ several recommendation strategies to indicate users what web pages to evaluate, and (2) provide a plugin extension to allow them to rate arbitrary web pages as they browse them, Figure 4. Concretely, we recommend a user to rate web pages which: (1) match her declared expertise (to determine a web page category we use AlchemyAPI [1]), (2) are scarcely rated (e.g., have less then 5 evaluations), or (3) are controversial (i.e., there is no agreement between the raters).



Support your review

If you have identified some issue about this page, put it here. It helps others understand, upvote you review and to earn you reputation.

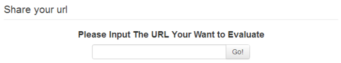
Overall credibility score

More details

Figure 4: Using the browser extension (i.e, Chrome extension) to evaluate a webpage.

What a contributor can do?
Our initial prototype allows users to either log in with their social network accounts (currently CredibleWeb supports Facebook and Twitter), or create a separate account on our platform. Once logged in, users can perform several operations:

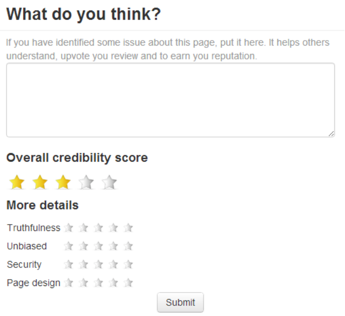
Evaluate web pages that are recommended by our system in the evaluation pane, or they can submit URLs for



Share your url

Please Input The URL Your Want to Evaluate

Figure 5: Users can submit URLs for evaluation.



What do you think?

If you have identified some issue about this page, put it here. It helps others understand, upvote you review and to earn you reputation.

Overall credibility score

More details

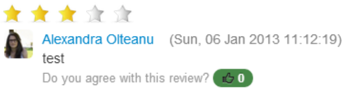
Truthfulness

Unbiased

Security

Page design

Figure 6: The dialog window that appears when the user clicks the "Evaluate" button in the IFrame evaluation interface.



Alexandra Olteanu (Sun, 06 Jan 2013 11:12:19)

test

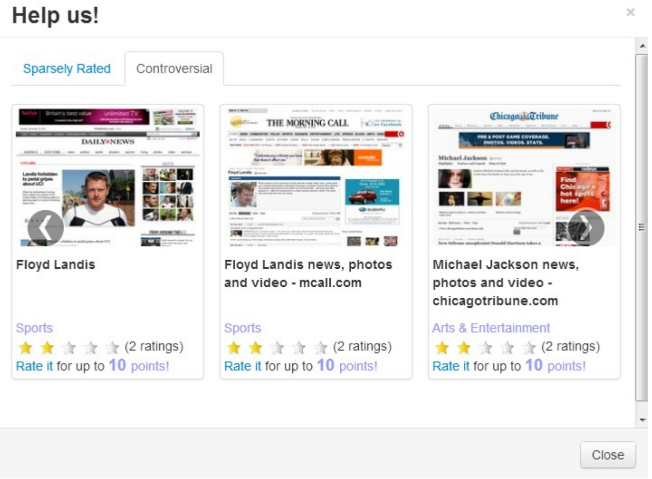
Do you agree with this review? 0

Figure 7: Users can vote a review if they find it useful.

evaluation and check other users opinion about them, Figure 5. In the evaluation pane, we use the three recommendation strategies previously discussed.

Once a recommended web page is clicked or an URL is submitted for evaluation, the web page is opened up within an Iframe, which we refer to as the *evaluation interface*. At the bottom of the window there is a bar that contains brief information about the user and the web page, along with three buttons: one for evaluation that opens a dialog window to allow the user to edit and submit his review or check past reviews (Figure 6), and two buttons that recommend users to evaluate web pages under two recommendation conditions. The recommendation conditions try to capture two different engagement strategies: (1) *I want to help!* which exploits a user desire to "do the right thing" and recommend users to rate scarcely rated or controversial web pages (Figure 8), and (2) *I'm bored!* which tries to entertain users by making "surprise" recommendations [10].

CredibleWeb also tries to reconstruct the underlying **social network** between users. For this, it fetches the friend lists from users social networking profiles when they join the platform using a social network account, and also allows users to connect with friends directly on the platform. When a user friend joins the platform with her social network account, we identify the social connection and connect them on *CredibleWeb* as well. In addition, a user can merge her social accounts allowing us to find common links across different social networks and build a multi-layered social graph (e.g., each layer represents the social links within different social networks like Facebook and Twitter). We note that users can gain points for all the actions that they perform on the platform (e.g., evaluate web pages, add friends, merge accounts).



Help us!

Floyd Landis

Sports

(2 ratings)

Rate it for up to 10 points!

Floyd Landis news, photos and video - mcall.com

Sports

(2 ratings)

Rate it for up to 10 points!

Michael Jackson news, photos and video - chicagotribune.com

Arts & Entertainment

(2 ratings)

Rate it for up to 10 points!

Figure 8: The dialog window that recommends users controversial web pages for evaluation.

Finally, *CredibleWeb* allows users to **moderate** other users contributions by voting up the good ones (Figure 7). This is meant to help the system to manage the potential abuse, by promoting good evaluations. When an evaluation is voted both the author and the rater win points.

How to manage abuse?
We want to be able to manage random and malicious contributions. To this end, we consider two schemes to ensure the reliability of our data: **Manage abuse when collecting data.** For this we employ (1) a self moderation mechanism that tries to promote good reviews by allowing users to vote them up, and (2) limit the amount of points (and, thus, impacting the speed with which a potential malicious user can gain reputation) that one can gain by rating multiple times the same web page or by rating web pages in automatic fashion (i.e., using a bot). Yet, we

note that, in general, it is hard to judge on the fly if a user evaluations are honest given the multiple strategies that a potential attacker can use to manipulate the system.

Manage abuse after collecting data. The data can also be cleaned after collection (e.g., by applying statistical tools to filter out spam ratings). Yet, this also have drawbacks as, for instance, it is hard to identify outliers for scarcely rated web pages.

Conclusion and Future Work

This paper introduces *CredibleWeb*, a crowdsourcing platform that aims to build a large scale dataset of credibility labeled webpages. We discussed some of the main challenges in designing our platform by grouping them by the key challenges that a crowdsourcing platform needs to address. However, there are still few aspects that need to be addressed, such as: (1) how to aggregated evaluations (e.g., when there is evidence of polarized opinions, for scarcely rated web pages vs. popular ones), (2) how to manage evaluation on dynamic web pages (e.g., dealing with old ratings on stale content), and (3) how to quantify the influence of old evaluation (if available to users) on the later ones.

Acknowledgements

We are grateful to Jonas Warner for sharing with us the initial prototype for *Heaven Points* application. We are grateful to Stefan Bucur, Yu Chen, Xin Liu, Tri Kurniawan Wijaya, LSIR group members and the anonymous reviewers for their valuable feedback. This work is partially supported by the grant *Reconcile: Robust Online Credibility Evaluation of Web Content* from Switzerland through the Swiss Contribution to the enlarged European Union.

References

[1] Alchemy API. http://alchemyapi.com.

[2] Hypothes.is. http://hypothes.is/.

[3] Stack Overflow. http://stackoverflow.com.

[4] Truth Goggles Demo. http://truthgoggl.es/demo.html.

[5] Aggarwal, S., and Van Oostendorp, H. An attempt to automate the process of source evaluation. In *Proc. of ACE* (2011).

[6] Arms, W. What are the alternatives to peer review? quality control in scholarly publishing on the web. *Journal of Electronic Publishing* (2002).

[7] Dawnay, E., and Shah, H. Behavioural economics: seven principles for policy makers. Tech. rep., New Economics Foundation, 2007.

[8] Doan, A., Ramakrishnan, R., and Halevy, A. Crowdsourcing systems on the world-wide web. *Communications of the ACM* (2011).

[9] Olteanu, A., Peshterliev, S., Liu, X., and Abereri, K. Web credibility: Features exploration and credibility prediction. In *Proc. of ECIR* (2013).

[10] Onuma, K., Tong, H., and Faloutsos, C. Tangent: a novel,'surprise me', recommendation algorithm. In *Proc. of KDD* (2009).

[11] Schwarz, J., and Morris, M. Augmenting web pages and search results to support credibility assessment. In *Proc. of CHI* (2011).

[12] Sondhi, P., Vydiswaran, V. G. V., and Zhai, C. Reliability prediction of webpages in the medical domain. In *Proc. of ECIR* (2012).

[13] Von Ahn, L., and Dabbish, L. Labeling images with a computer game. In *Proc. of CHI* (2004).

[14] Yamamoto, Y., and Tanaka, K. Enhancing credibility judgment of web search results. In *Proc. of CHI* (2011).