WEBCENTIVES'09

1st International Workshop on Motivation and Incentives

co-located with the WWW 2009, Madrid, Spain

Summary

The Web 2.0 movement has brought a new generation of usability and socio-technical change to the Web. At the same time, several so-called Web 2.0 applications had enormous success: Wikipedia, del.icio.us, Flickr, YouTube, Facebook, Twitter, Geni – to name just a few. Having differing objectives, they all have something in common: huge amounts of enthusiastic users contributing and creating a plethora of content. The high acceptance of these applications with Web users from all over the world prove that they are usable and – more importantly – provide some kind of benefit. Each of the applications has incentive structure well in place, triggering user interest and involvement.
The aim of the workshop is to address the following questions around incentives and motivation of Web applications: what is the motivation for a user to (install and) use a tool? Which incentive structures can be applied to the Web, which cannot? Moreover, incentives are a crucial topic for future Web generations: Web paradigms, like the Semantic Web or the 3D Web, that are novel and unfamiliar to end users, aim to involve wide user bases. **WEBCENTIVES** will attract contributions analyzing, applying, and designing incentive structures for Web applications. We want to emphasize that the workshop also aims at failures, i.e. cases where incentives failed, in order to understand why they failed and to disseminate the lessons learned.

Program Committee

- Sofia Angeletou, Open University, UK
- Anupriya Ankolekar, HP, US
- Sinuhe Arroyo , University of Alcala, Spain
- Sören Auer, University of Leipzig, DE
- Phoebe Ayers, University of California, USA
- Jie Bao, RPI, USA
- Chris Bizer, FU Berlin, DE
- Kurt Bollacker, Metaweb, US
- Danah Boyd, University of California in Berkeley, US
- Dan Brickley, The FOAF Project, UK
- Tom Heath, Talis, UK
- Andreas Hotho, University of Kassel, DE
- David Huynh, Metaweb, US
- Nick Kings, British Telecom, UK

- Andrew Lih, University of Hong Kong, China
- Eyal Oren, VU Amsterdam, NL
- Valentina Presutti, CNR, IT
- Mathias Schindler, Wikimedia, DE
- Andreas Schmidt, FZI, DE
- Hideaki Takeda, University of Tokyo, NII, JP
- David Weinberger, Harvard, US

Organizers

- Elena Simperl - STI Innsbruck, University of Innsbruck, Austria
- Katharina Siorpaes - STI Innsbruck, University of Innsbruck, Austria
- Denny Vrandecic - AIFB, Universität Karlsruhe, Germany
- Tim Bartel - Wikia Inc., Country Manager Germany

Contact

http://webcentives09.sti-innsbruck.at/

Is Editing More Rewarding Than Discussion?

A Statistical Framework to Estimate Causes of Dropout from Wikipedia

Ulrik Brandes
University of Konstanz, Germany
Ulrik.Brandes@uni-konstanz.de

Patrick Kenis
TiasNimbas Business School & Tilburg University, Netherlands
p.kenis@tiasnimbas.edu

Jürgen Lerner
University of Konstanz, Germany
lerner@inf.uni-konstanz.de

Denise van Raaij
Tilburg University, Netherlands
D.P.A.M.Korssen-vanRaaij@uvt.nl

ABSTRACT

In this paper we address the question: what causes formerly active Wikipedians to stop contributing? Seen from a different angle, we estimate characteristics of users, pages, or the whole system that increase or decrease the probability of dropout. We propose a general statistical method with which hypothetical causes of dropout can be tested. With this method it can be analyzed whether the emerging structures in Wikipedia function as incentives preventing Wikipedians to stop contributing. Applying this method to a selection of active users reveals, among others, that participation in discussion pages, as well as editing controversial pages, increases the dropout hazard, whereas editing general content pages has an attenuating effect on dropout. Although our method is solely illustrated on Wikipedia, it can be easily applied to other Web 2.0 applications.

Keywords

Wikipedia, lifetime-analysis, missing Wikipedians, motivation, frustration

1. INTRODUCTION

As any Web 2.0 application, Wikipedia needs, in order to grow and improve, a large number of motivated contributers. Given this fact, it is crucial and insightful for Web 2.0 researchers to learn about the causes to contribute and, as the other side of the coin, learn about the causes to stop contributing. Here we are interested in emerging mechanisms in Wikipedia that either motivate and reward contributors or frustrate users making them to leave as Wikipedians. Although these mechanisms can have an implicit nature (i. e., have not been designed as systematic feedback systems that aim at rewarding contributors [12]), increased knowledge in their functioning could be a first step in helping system designers and administrators to sustain enthusiastic users. In this paper, we focus on active users (i. e., users who performed a certain minimum number of contributions) and attempt to find factors that influence the probability whether such a user *survives as a Wikipedian* (i. e., continues to contribute) or *dies as a Wikipedian* (i. e., not contributes anymore). The restriction to active users is mostly due to statistical reasons (for inactive users we do not have sufficient data) but, arguably, the active users are also the more interesting ones.

Causes for dropout can be manifold and we distinguish between factors that are *exogenous* and factors that are *endogenous* to Wikipedia. Exogenous factors include demographic variables such as age, gender, education level, marriage status, profession, or occupation as well as external events such as getting a new job or getting children. Endogenous factors include everything that can be determined from the history of Wikipedia, i. e., information about edits, discussion, elections for administrator status, featured article voting, user blocking, page blocking and so on. While many exogenous factors may strongly influence the decision to not contribute anymore (in some cases, simply for the reason that the user does no longer have time to spend days or nights editing Wikipedia), we do not use them in this paper. The major reason for this decision is that we are attempting to uncover which features that are endogenous to the system function as incentives for sustained contribution and, vice versa, which endogenous features trigger dropout of Wikipedians. Such information can (to some degree) be used to design and shape Web 2.0 applications in order to enhance motivation of contributors.

Since we do not use exogenous factors—although they might influence the dropout probability—it seems to be obvious that there will be cases of dropout that are not well described by our model. We emphasize that we do not attempt to maximize the precision of predicting dropouts; rather, the goal of our analysis is to test statistically whether specific endogenous factors do, yes or no, increase or decrease the probability of leaving Wikipedia—thereby getting a better understanding which emerging and often implicit mechanisms contribute to sustain users. Such results are very useful because designers or administrators of Web 2.0 applications might use them to mitigate causes for dropout or add features that decrease dropout probability—even if the empirical time-to-dropout data contains unexplained variance due to exogenous factors. An additional consequence of our approach is that we are able to better understand the social collaboration process in Wikipedia by detecting characteristics that distinguish high-quality collaboration from low-quality collaboration; while an obvious quality dimension would be the quality of the encyclopedic entries, we claim that keeping contributors motivated is another very

WWW2009, April 20-24, 2009, Madrid, Spain.
.

important aspect of quality of the system (also see Sect.2.1).

Even if we do not use exogenous predictors for dropout in this paper, we emphasize that the general statistical method presented in Sect. 3 is applicable to all kinds of predictors, independent on whether they stem from log-data, demographic data, or questionnaire-based surveys.

In Sect. 2 we put the topic of this paper into the context of a broader research project, provide background on statistical methods for lifetime analysis, and review related work on Wikipedia research. Section 3 presents our statistical framework to model dropouts from Wikipedia. In Sect. 4 we report on the results of an empirical analysis using this model and Sect. 5 indicates future work.

2. BACKGROUND

2.1 Dropout Hazard as a Proxy for Quality

The long-term goal of this project is to gain insight into the social collaboration process in community forms of organizations (in contrast to formal or hierarchical organizations) that rise at the Internet and that we refer to as *webbased information communities* (*WebICs*). "WebICs are defined as work systems facilitated by the Internet infrastructure and composed of voluntary actors that attempt to produce a product or service such as software or encyclopedic information [2]." WebICs are organized in an informal way and are governed and coordinated by flows and linkages between actors [11]. Based on existing knowledge in the field of organization studies we argue that one of the success factors of WebICs is this implicit and emergent governance and coordination structure. However, since WebICs are not successful by definition, our research attempts to find out the characteristics of high-quality and low-quality collaboration structures.

Quality of Wikipedia most often refers to the quality of its encyclopedic entries: For instance it has been suggested that various forms of vandalism are indicators of (low) quality of articles [18]. Others have applied self-assessment criteria developed in Wikipedia, such as distinctions between excellent featured pages and worth-reading featured pages [15], or featured versus controversial pages [2]. Another way to assess the quality of articles is to present a number of selected Wikipedia entries to scientific experts [6].

However, quality of Wikipedia does not only mean quality of its encyclopedic articles; instead we argue that the dropout hazard of Wikipedians can also function as a proxy for quality of the system or certain parts of it (needless to say that a high dropout hazard is interpreted as pointing to low quality). This approach is based on the observation that: "Wikipedia operates from the presumption that any individual's knowledge is by definition incomplete and that ongoing revisions enabled by mass collaboration tools and involving a large group of eyeballs will produce a reliable yet continually evolving knowledge repository [5, p.361]." As a consequence, the ability of Wikipedia to prevent turnover and motivate Wikipedians to continue to contribute can be understood as a quality indicator of its governance and coordination structure. Turnover in formal, hierarchical organizations is associated with the loss of human capital and thus the loss of hiring and training investments [14]. Turnover in the context of Wikipedia can be associated with the loss of work force, their skills and knowledge and consequently, the decrease of production of encyclopedic knowledge. While formal, hierarchical organizations can manage employee commitment through, among others, economic incentives, formal training, contracts, and formal supervision procedures, Wikipedia can only rely on non-economic incentives to sustain contributors commitment [12]. Hence, if Wikipedia is able to preserve large numbers of highly contributing users, it is likely to produce higher outcome quality than if it lacks the ability to motivate contributors.

2.2 Statistical Methods for Lifetime Analysis

Lifetime analysis (also referred to as *time-to-event analysis*, *failure analysis*, or *survival analysis*) is an area of statistics that is concerned with modeling the elapsed time until a specific event happens; a general reference is given by Lawless [9]. Using a customary vocabulary, lifetime analysis models the time until a certain object *dies*, where *death* is sometimes meant in a metaphorical way. Lifetime analysis is frequently used in medicine, engineering, social science, and political science, among others. For instance, in medicine researchers are interested in how long a patient suffering a certain illness survives; engineers might be concerned with how long it takes until a manufactured item (e. g., a computer) breaks down. In this paper we are interested in the dropout of Wikipedians, i. e., in the events in which formerly active Wikipedia users stop contributing.

Besides estimating the actual survival times, another goal of lifetime analysis is to discover factors that increase or decrease the probability to die. Returning to the above examples, a specific pharmaceutical treatment may or may not empirically increase the survival time of patients; the lifetime of a computer may be dependent on the specific machine that manufactured it (potentially pointing to faults of machines). As already mentioned, we are attempting to uncover the reasons for dropout in Wikipedia, i. e., which factors increase or decrease the probability of dropout.

Lifetime analysis is often confronted with specific properties of the data that require special care. In many cases (and also in our case) lifetime analysis is faced with so-called *right-censoring*, meaning that some of the selected instances have not died at the time of data collection. Ignoring these survivors would introduce a serious bias into the analysis (intuitively, it would be hard to learn about the causes of survival, if surviving instances were discarded). Instead our model has to deal with the fact that for one part of the instances (namely those that died, later in this paper referred to as *dropouts*) we know the time when the individual died and for the other part of instances (later in this paper referred to as *survivors*) we only know that they survived beyond a certain point in time. See Sects. 3.3 and 3.5 how these instances are treated differently. Another issue to take care of is the definition of when a specific individual enters the *risk set* (i. e., the set of individuals that have a non-zero probability to die). We note first that in our case individuals (i. e., contributors of Wikipedia) enter the risk set at different time points, namely at the time of their first edit. However, since we restricted our analysis to *active* users (see Sect. 3.2 for a definition of an active user) we introduced a further bias: by discarding inactive users the probability of reaching the active state is artificially set to one (if a user died before, it would not be in our set of instances). Section 3.5 shows how to correct for this bias. Nevertheless, we stress that even with this correction it would not be valid to generalize findings to inactive users: those that dropout

quickly might do so for totally different reasons than those that reach the active state.

2.3 Further Related Work

Wikipedia—besides being a popular Web page—has become a popular case in academic research. Several papers visualize certain aspects of the history of Wikipedia pages, i. e., the development of their content over time. The *history flow* visualization [18, 19] shows how sentences persist over time or get deleted at later revisions. Other researchers constructed and visualized networks encoding how users interact with the edits of others, e. g., [8, 16, 3, 2]. The revision history of Wikipedia articles has been further used to distinguish the edit behavior of different user groups [7], to define reputation or Wikipedians [1], to estimate the impact of vandalism [13], and to identify controversial articles [20]. We are not aware of any work that quantitatively analyzes causes for dropout of Wikipedians, which is the topic of the current paper. However, Lento et al. [10] examined causes for continued participation in the Wallop Weblogging system; a difference to their approach is that our method takes the effects of *time-varying* explanatory variables into account.

3. METHOD

3.1 Data

The selection of instances and the extraction of the explanatory variables is mostly based on the so called stub files from the latest available database dump of the English Wikipedia (see `http://download.wikimedia.org`). These stub files contain metadata (most notable page title, username, and timestamp) of every revision on every page (including talk pages etc.) since the launch of Wikipedia. The dump that we used for this paper has been started on October 8th, 2008. Although the file contains edits with later timestamps, we ignore these and take October 8th, 2008 as the day of data collection. The uncompressed XML-file has a size of 66 gigabytes. Although this is quite large, it is nevertheless manageable since the needed information can be extracted in a sequential manner.

Besides the history stub file, the content of two additional Wikipedia pages have been used: The list of users on the page `Wikipedia:Missing Wikipedians` is helpful for selecting dropout instances (see Sect. 3.2 for details) and the page `Wikipedia:List of controversial issues` is used for the computation of one of the explanatory variables (see Sect. 3.4 for details).

3.2 Selection of Instances

As already noted in the introduction, we restrict our analysis to active users which are defined as users that performed a given minimum number of edits. These active users are later partitioned into *dropouts* (those who are known to have stopped contributing at a certain moment) and *survivors* (those who are known to continue editing beyond the time of data collection). We note that some active users fall between these two categories, i. e., for those users we do not have sufficient information to decide whether they are dropouts or survivors; those users are discarded.

More precisely, the dropouts are (a subset of) users listed on the page `Wikipedia:Missing Wikipedians`. This page has been mentioned in *The Economist* in an article about Wikipedia stating that "It serves as a reminder that frustration at having work removed prompts many people to abandon the project [4]." The first lines of the missing Wikipedian page already give an intuitive definition of what is a missing Wikipedian:

> This is a list of Wikipedians who are no longer an integral part of the community. [. . .] Wikipedians who no longer edit due to confirmed death should instead be added to `Wikipedia:Deceased Wikipedians`.
>
> [. . .]
>
> Please do not add people to this list who were never an integral part of the community. Don't add users with fewer than about 1,000 edits. Do not add people unless you are certain they have left, do not add anonymous users identified by their IP address (they could have created an account and still be contributing, or they might have a roaming IP address) and do not add yourself.

To make things precise we define (motivated by the above quotation) an *active Wikipedian* to be a logged-in user (in contrast to anonymous users identified by IP addresses) who is not a robot (i. e., not a software program that performs routine tasks) and who has performed at least 1,000 edits. From the database dump we derive that slightly more than 19,000 users qualify as active Wikipedians.

Dropout instances. To define the set of dropouts we start with all users listed on `Wikipedia:Missing Wikipedians`, yielding 501 users. From this set we deleted all those that made fewer than 1,000 edits, leaving us with 465 users. In order to not just trust the editors of the missing Wikipedian page we further delete all those that edited on or later than September 1st, 2008 (a bit more than one month before data collection). This gives us our final set of dropouts containing 413 users.

Survivor instances. For the survivors we start with the active Wikipedians, delete all those that are listed on the page of missing Wikipedians, and further delete all those that performed less than 30 edits in the time from July 1st, 2008 until the day of data collection. With the last step we want to exclude Wikipedians that do not qualify as dropouts but that are nevertheless not very active anymore; these users are simply harder to interpret. However, we suggest that formerly active Wikipedias that have not been listed on the page of missing Wikipedians (the *un-missed* dropouts) are an interesting population for future research. Altogether, the set of survivors contains 10, 454 users.

3.2.1 Notes on the Selection of Instances

We have chosen to select dropouts via the list of missing Wikipedians since this gives us some confidence that those users have indeed decided to stop participating, rather than just taking a break. However, it should be noted that this selection strategy implies that, strictly spoken, we estimate the causes for ending up on the page of missing Wikipedians, rather than the causes for dropout. Since only Wikipedians that are (well) known to at least one other user are put on this page, this selection procedure could introduce a bias in the analysis. We will analyze in future work the pros and

cons of alternative ways to divide active users into dropouts and survivors.

3.3 Statistical Model for Time-to-Dropout

While the procedure to select dropouts and survivors from Sect. 3.2 reflects a particular choice—giving emphasis to users that are recognized as missing by others—the model that is presented now is independent on the particular selection strategy and is (with a slight adaption in notation) also not restricted to Wikipedia.

3.3.1 Notation

Let $U = \{u_1, \ldots, u_n\}$ denote the selected users, where for an n_0 between one and n the set $D = \{u_1, \ldots, u_{n_0}\} \subseteq U$ contains exactly the dropouts. Let $u \in U$ be any selected Wikipedian. The random variable encoding u's dropout time is denoted by $T_u^{(drop)}$. The actual value of $T_u^{(drop)}$ is only observed if $u \in D$; in this case the observed dropout time of u is denoted by $t_u^{(drop)}$. Each user potentially starts (i. e., makes her first edit) at a different time point, denoted by $t_u^{(start)}$. By definition, selected users have performed at least a thousand edits; the time when u performed her thousandth edit is denoted by $t_u^{(1000)}$. Finally, the time point of data collection (i. e., October 8th, 2008) is denoted by $t^{(end)}$; it is equal for all users.

Turning to the explanatory variables, for a time point t let W_t denote the *history of Wikipedia* up to time t, i. e., information about every edit, discussion, voting, blocking (and so on) that took place on or before t. Later we let the risk of dropout at time t depend on W_t—more precisely, on particular *statistics* computed from W_t, see Sect. 3.4—and on nothing else. With $W = W_{t^{(end)}}$ we denote the history at the time of data collection, i. e., the entire data that we use to compute explanatory variables.

3.3.2 Survival, Hazard, and Probability Density

The methodology outlined in this section is not restricted to model dropouts; it is rather standard methodology for lifetime analysis in general, see [9].

As before, let $u \in U$ be any selected Wikipedian. The function

$$f_u(t) = \lim_{\Delta t \to 0} \frac{Pr\left(t \leq T_u^{(drop)} < t + \Delta t\right)}{\Delta t} \qquad (1)$$

is the probability density for u's dropout time being equal to t; f_u is defined on the real interval $[t_u^{(start)}, \infty[$.

At a first glance it seems that we could use f_u to test hypothetical causes of dropout by specifying f_u as parametrically dependent on covariates (encoding the potential causes) and testing whether those covariates show the predicted effect: covariates that empirically increase f_u (i. e., the risk to drop out) would then be interpreted as causes of dropout. However, this approach would not take into account an intrinsic dependency in lifetime data: an instance that dies at time t must necessarily survive up to this time point. To illustrate this on a simple example, assume that we were modeling the lifetime of humans. It is plausible that only a small percentage of people dies at the age of 100 years. However it would be wrong to conclude that people in their hundredth year are at a low risk of dying; the low percentage is rather due to the fact that very few people ever survive up to their hundredth year.

Returning to the case of Wikipedians but keeping the above example in mind, we see that we should rather model the *conditional* probability of users dropping out at time t, under the precondition that they survived up to t. This conditional probability density

$$h_u(t) = \lim_{\Delta t \to 0} \frac{Pr\left(t \leq T_u^{(drop)} < t + \Delta t \middle| t \leq T_u^{(drop)}\right)}{\Delta t}$$

is called the *hazard function* [9]; h_u is defined on the real interval $[t_u^{(start)}, \infty[$.

The hazard to drop out at time t is modeled as a function of various *statistics* $s_i(u; W_t)$, $i = 1, \ldots, k$ (characterizing certain aspects of the Wikipedia history at time t around user u) and parameters $\theta = (\theta_1, \ldots, \theta_k)$ that encode whether the respective statistics have a decreasing or increasing effect (or none) on the hazard to drop out. More precisely, we model the dropout rate in the following functional form:

$$h_u(t) = h_u(W_t; \theta) = \exp\left(\sum_{i=1}^{k} \theta_i \cdot s_i(u; W_t)\right) \qquad (2)$$

The estimated parameter values give information about the causes of dropout: if, for instance, a statistic $s_i(u; W_t)$ encodes how much u participates in discussion and if the associated parameter θ_i is significantly positive (negative), then participation in discussion is correlated with a higher (lower) probability to drop out. (Actually, it turns out that discussion is correlated with a *higher* probability to drop out, see Sect. 4.)

The general model outlined so far can be applied to test hypotheses about the interplay between characteristics of the Wikipedia system and the dropout hazard of Wikipedians. The model is specialized to test concrete hypotheses by plugging appropriate statistics into Eq. (2). The statistics that we take in this paper are defined in Sect. 3.4.

Equation (2) formalizes the assumption that the time dependence of the dropout hazard is completely captured in W_t. In other words, we assume that only endogenous factors are responsible for triggering dropout and, given the history of Wikipedia W_t, the hazard is conditionally independent of time.

While the hazard rate is convenient for parametric modeling, we nevertheless need for parameter inference (Sect. 3.5) the probability density f_u, see Eq. (1), and the *survivor function*

$$S_u(t) = Pr\left(t \leq T_u^{(drop)}\right);\ t \in [t_u^{(start)}, \infty[$$

(denoting the probability to survive as a Wikipedian beyond time t). However, specifying the hazard function h_u is sufficient since it determines both, the survivor function S_u and the probability density f_u by (cf. [9])

$$\begin{aligned} S_u(t) &= \exp\left(-\int_{t_u^{(start)}}^{t} h_u(x)\, dx\right) \text{ and} \\ f_u(t) &= h_u(t) \cdot \exp\left(-\int_{t_u^{(start)}}^{t} h_u(x)\, dx\right) . \end{aligned}$$

3.4 Explanatory Variables

In this section we define the concrete statistics that we take in this paper as the determinants of the dropout hazard, see Eq. (2). Each statistic corresponds to a hypothetical factor that might increase or decrease the hazard to drop out.

The estimation of the associated parameter (see Sects. 3.5 and 4) reveals whether such a hypothetical dependency can be empirically validated.

The statistics that we take in this paper are quite simple from a computational point of view. Other more involved statistics will be treated in future research (also see Sect. 5).

3.4.1 Editing, Discussing, and Organizing

The first family of statistics is constructed to answer the question: do users become more robust against dropout when they accumulate a growing number of contributions? A positive answer to this question would imply that users are more likely to drop out at the beginning of their career than at later stages. A negative answer would imply that users wear out and their dropout hazard increases with a growing number of contributions. However, since users can contribute to Wikipedia in different ways, we distinguish between three different kinds of contributions: (1) editing encyclopedic entries, (2) discussing, and (3) performing organizational work in Wikipedia.

To provide some background on this distinction, we recall that the set of Wikipedia pages is partitioned into various *namespaces* representing different types of pages (see the page `Wikipedia:Namespaces`). The *main namespace* comprises the set of encyclopedic articles. In the following, we denote contributions to the main namespace as *editing*. Besides the articles pages—whose creation is the main purpose of Wikipedia—there are pages which are concerned with various kinds of organizational work. These include pages in the namespaces `Wikipedia (Project)`, `Portal`, `User`, `File`, `MediaWiki`, `Template`, `Category`, `Help`, `Media`, and `Special`. In the following, we denote contributions to these namespaces as *organizing*. Finally, pages of all namespaces except `Media` and `Special`, but including the main namespace, have associated *talk pages* providing space for discussion. In the following, we denote all contributions to the talk pages as *discussing*.

Several researchers, including [19, 8], pointed out that discussion and organization work increased more rapidly over the last years than editing main articles. In this paper we analyze whether contributions to these three types of pages have different implications for the dropout hazard.

To define the statistics encoding how much a particular user u contributed to these three types of pages up to a timepoint t, let $E_{u,t}$ denote the set of revisions that u performed on pages of the main namespace on or before time t; let $T_{u,t}$ denote u's revisions to discussion pages on or before t; and let $O_{u,t}$ denote u's revisions to pages in all other namespaces (listed above) on or before time t. The respective statistics, to be used in Eq. (2), are defined by

$$\begin{aligned} \texttt{edit}(u; W_t) &= \log(1 + |E_{u,t}|) \\ \texttt{discuss}(u; W_t) &= \log(1 + |T_{u,t}|) \\ \texttt{organize}(u; W_t) &= \log(1 + |O_{u,t}|) \ . \end{aligned}$$

The logarithmic scaling of the number of revisions has been chosen due to the extremely skewed distribution (there are users who performed more than 100,000 revisions, while most of the selected users have a count of only slightly more then 1,000).

The interpretation of the associated parameters is as follows. A significantly positive (negative) parameter associated with `edit` implies that users with a higher number of revisions to the main namespace have a higher (lower) hazard to drop out. The interpretation for the parameters associated with `discuss` and `organize` is analogous.

3.4.2 Feedback

Another likely determinant of the dropout probability is the feedback that a user receives from others. Positive feedback is likely to have a motivating effect and, thus, might reduce the dropout hazard. On the other hand, negative feedback is likely to be frustrating and might increase the dropout hazard. Feedback can be provided to a user via her *user talk page* (see the page `Wikipedia:User talk page`). Since we want to rely in this paper only on automatic (and simple) methods, we do not evaluate whether feedback is positive or negative but only count the number of revisions made to the talk page of a particular user. Additionally we count how many contributions to the talk page of user u are made by u herself; thereby we can distinguish between users who reply to feedback given to them and users who do not (or less) reply.

More precisely, let $T_t^{(u)}$ denote the set of revisions to the user talk page of user u that are performed by any user on or before time t. Similarly, let $T_{u,t}^{(u)}$ denote the revisions made by u to her own user talk page on or before t. The respective statistics, to be used in Eq. (2), are defined by

$$\begin{aligned} \texttt{getFeedback}(u; W_t) &= \log(1 + |T_t^{(u)}|) \\ \texttt{replyFeedback}(u; W_t) &= \log(1 + |T_{u,t}^{(u)}|) \ . \end{aligned}$$

A significantly positive (negative) parameter associated with `getFeedback` implies that users with a higher number of revisions made to their user talk page have a higher (lower) hazard to drop out.

3.4.3 Controversy

Another reason for dropping out might be that Wikipedians are frustrated from ongoing controversies or edit wars with other users. To analyze this we look at how much a certain user edits *controversial* pages, i. e., pages mentioned on `Wikipedia:List of controversial issues`. Similar as above, let $C_{u,t}$ denote the set of revisions that a user u made to any controversial page on or before time t and define the respective statistic by

$$\texttt{editControversial}(u; W_t) = \log(1 + |C_{u,t}|) \ .$$

A significantly positive (negative) parameter associated with `editControversial` implies that users with a higher number of revisions made to controversial articles have a higher (lower) hazard to drop out.

3.5 Parameter Inference from Observations

This section provides details about how the parameters θ_i in Eq. (2) are computed from a set of observed dropout users and survivors. Readers not interested in this may directly continue with Sect. 4 (note that the parameters can be interpreted without knowledge of the estimation algorithm).

Let $U = \{u_1, \ldots, u_n\}$ denote the selected users, where for an n_0 the set $D = \{u_1, \ldots, u_{n_0}\} \subseteq U$ contains exactly the dropouts. Any observation of a $u \in U \setminus D$ (i. e., each survivor) gives us the information that u survived beyond time $t^{(end)}$. Since all selected users have at least thousand edits, the probability for surviving up to $t_u^{(1000)}$ is equal to

one. Thus, the probability for observing $u \in U \setminus D$ is

$$\begin{aligned}
&Pr\left(t^{(end)} \leq T_u^{(drop)} \middle| t_u^{(1000)} \leq T_u^{(drop)}; W; \theta\right) \\
&= \frac{S_u(t^{(end)}; W; \theta)}{S_u(t_u^{(1000)}; W; \theta)} \\
&= \frac{\exp\left(-\int_{t_u^{(start)}}^{t^{(end)}} h_u(W_x; \theta)\, dx\right)}{\exp\left(-\int_{t_u^{(start)}}^{t_u^{(1000)}} h_u(W_x; \theta)\, dx\right)} \\
&= \exp\left(-\int_{t_u^{(1000)}}^{t^{(end)}} h_u(W_x; \theta)\, dx\right) \\
&= \texttt{survive}_u(W, \theta)
\end{aligned}$$

For each $u \in D$ (i. e., for each dropout instance) we know that u dropped out at $t_u^{(drop)}$. As above, we have to correct for the fact that we selected only users with at least thousand edits. Thus, the probability density for observing $u \in D$ is

$$\begin{aligned}
&f_u(t_u^{(drop)} \big| t_u^{(1000)} \leq T_u^{(drop)}; W; \theta) \\
&= \frac{f_u(t_u^{(drop)}; W; \theta)}{S_u(t_u^{(1000)}; W; \theta)} \\
&= h_u(W_{t_u^{(drop)}}; \theta) \cdot \frac{S_u(t_u^{(drop)}; W; \theta)}{S_u(t_u^{(1000)}; W; \theta)} \\
&= h_u(W_{t_u^{(drop)}}; \theta) \cdot \exp\left(-\int_{t_u^{(1000)}}^{t_u^{(drop)}} h_u(x; W_x; \theta)\, dx\right) \\
&= \texttt{dropout}_u(W, \theta)
\end{aligned}$$

The joint probability density to observe the complete set of selected users U is

$$f(U, \theta) = \left(\prod_{i=1}^{n_0} \texttt{dropout}_{u_i}(W, \theta)\right) \cdot \left(\prod_{i=n_0+1}^{n} \texttt{survive}_{u_i}(W, \theta)\right)$$

(Here we assumed that dropouts are conditionally independent, given the history of Wikipedia W, i. e., we assume that W captures all the necessary information that determines dropout. For instance, an agreement between two users of the kind "I drop out, if you drop out" would violate this independence assumption; nevertheless, if two users drop out due to the same endogenous factor these dropout events are *conditionally* independent, although not independent.)

For a fixed observation U, we obtain a likelihood function L on the space of parameters $\Theta = \mathbb{R}^k$ by

$$L\colon \Theta \to \mathbb{R};\ \theta \mapsto f(U, \theta)$$

and we estimate those parameters $\hat{\theta} = \operatorname{argmax} L$ that maximize L (maximum likelihood principle, cf. [21]).

Computational simplification. We note that the state of Wikipedia W_t changes only when an edit is performed, i. e., only at finitely many time points (albeit a lot). Hence, if the statistics $s_i(u; W_t)$ have no explicit time-dependency, they are piecewise constant functions and the integrals in the equations above are equal to weighted sums (where the weights correspond to the lengths of the time intervals during which the state of Wikipedia remains unchanged). For practical and computational reasons we will simplify this further and approximate the state of Wikipedia in the sense that we let W_t change only once a day. Thus, the statistics $s_i(u; W_t)$ are constant for each day and the integrals reduce to a manageable number of summands.

Thus, from now on we assume that time is given by integer numbers denoting a counter for days. In particular, $\sum_{x=t_1}^{t_2} h_u(x; \theta)$ denotes the sum over $h_u(x; \theta)$, where the day counter x goes from t_1 to t_2.

Estimation algorithm. The maximum likelihood estimates of the parameters are computed by the established Newton-Raphson algorithm. First, we note that parameters $\hat{\theta}$ maximize L if and only if $\hat{\theta}$ maximize $\log L$; however, $\log L$ has a simpler functional form. It is

$$\begin{aligned}
&\log L(\theta) = \\
&\left(\sum_{i=1}^{n_0} \log \texttt{dropout}_{u_i}(W, \theta)\right) \\
&+ \left(\sum_{i=n_0+1}^{n} \log \texttt{survive}_{u_i}(W, \theta)\right) = \\
&\left(\sum_{i=1}^{n_0} \log h_{u_i}(W_{t_{u_i}^{(drop)}}; \theta) - \int_{t_{u_i}^{(1000)}}^{t_{u_i}^{(drop)}} h_{u_i}(W_x; \theta)\, dx\right) \\
&+ \left(\sum_{i=n_0+1}^{n} -\int_{t_{u_i}^{(1000)}}^{t^{(end)}} h_{u_i}(W_x; \theta)\, dx\right) \quad ,
\end{aligned}$$

where $h_{u_i}(W_x; \theta) = \exp\left(\sum_{j=1}^{k} \theta_j \cdot s_j(u_i; W_x)\right)$. With the convention that we make changes to W_t only once a day (see above) we obtain

$$\begin{aligned}
&\log L(\theta) = \\
&\sum_{i=1}^{n_0} \sum_{j=1}^{k} \theta_j \cdot s_j(u_i; W_{t_{u_i}^{(drop)}}) \\
&- \sum_{i=1}^{n_0} \sum_{x=t_{u_i}^{(1000)}}^{t_{u_i}^{(drop)}} \exp\left(\sum_{j=1}^{k} \theta_j \cdot s_j(u_i; W_x)\right) \\
&- \sum_{i=n_0+1}^{n} \sum_{x=t_{u_i}^{(1000)}}^{t^{(end)}} \exp\left(\sum_{j=1}^{k} \theta_j \cdot s_j(u_i; W_x)\right)
\end{aligned}$$

The first order partial derivative with respect to $\ell = 1, \ldots, k$ is

$$\begin{aligned}
&\frac{\partial}{\partial \theta_\ell} \log L(\theta) = \\
&\sum_{i=1}^{n_0} s_\ell(u_i; W_{t_{u_i}^{(drop)}}) \\
&- \sum_{i=1}^{n_0} \sum_{x=t_{u_i}^{(1000)}}^{t_{u_i}^{(drop)}} s_\ell(u_i; W_x) \cdot h_{u_i}(W_x; \theta) \\
&- \sum_{i=n_0+1}^{n} \sum_{x=t_{u_i}^{(1000)}}^{t^{(end)}} s_\ell(u_i; W_x) \cdot h_{u_i}(W_x; \theta)
\end{aligned}$$

The second order partial derivative with respect to $\ell, \ell' = 1, \ldots, k$ is

$$\frac{\partial^2}{\partial\theta_{\ell'}\partial\theta_\ell}\log L(\theta) =$$
$$-\sum_{i=1}^{n_0}\sum_{x=t_{u_i}^{(1000)}}^{t_{u_i}^{(drop)}} s_\ell(u_i;W_x)\cdot s_{\ell'}(u_i;W_x)\cdot h_{u_i}(W_x;\theta)$$
$$-\sum_{i=n_0+1}^{n}\sum_{x=t_{u_i}^{(1000)}}^{t^{(end)}} s_\ell(u_i;W_x)\cdot s_{\ell'}(u_i;W_x)\cdot h_{u_i}(W_x;\theta)$$

Let

$$\nabla\log L(\theta) = \left(\frac{\partial}{\partial\theta_\ell}\log L(\theta)\right)_{\ell=1,\ldots,k}$$

denote the vector of first order derivatives and let

$$H(\theta) = \left[\frac{\partial^2}{\partial\theta_{\ell'}\partial\theta_\ell}\log L(\theta)\right]_{\ell,\ell'=1,\ldots,k}$$

denote the matrix of second order derivatives. Start with initial parameter values $\theta^{(0)}$ and update for $i = 0, \ldots,$ *max-iter* by setting

$$\theta^{(i+1)} = \theta^{(i)} - \left(H(\theta^{(i)})\right)^{-1}\cdot\nabla\log L(\theta^{(i)}) \ ,$$

until $\nabla\log L(\theta^{(i)})$ is sufficiently close to zero. This $\theta^{(i)}$ is then a good approximation for the maximum likelihood estimate $\hat{\theta}$.

4. RESULTS AND DISCUSSION

We estimated the model outlined in Sect. 3.3 with the six explanatory statistics (editing, discussing, organizing, getting feedback, replying to feedback, and editing controversial articles, defined in Sect. 3.4) plus an additional constant parameter. The main information resulting from this analysis is whether the associated parameters are significantly positive (revealing a tendency for increased dropout hazard) or significantly negative (revealing a tendency for decreased dropout hazard). The constant just normalizes the model to the empirical time scale in which one unit corresponds to the expected time-to-dropout of a (hypothetical) user for which the effects of all other statistics add up to zero. The value of this constant does not provide much information; if we had started with another time unit (the time unit of our model is one day) we would have obtained another value as constant.

The estimated parameter values and estimated standard errors are reported in Table 1. The parameters are significantly different from zero at the 5%-level, if the resulting t-ratio (the absolute value of the parameter divided by the standard error) is at least 1.96, cf. [21]. All six parameters turned out to be significant at this level. The interpretation of the results is below.

The parameter associated with `edit` is negative, indicating that the dropout hazard of a user decreases with a growing number of edits to the main namespace (i. e., the set of encyclopedic articles). Thus, users are more likely to drop out early in their career and gain robustness against leaving Wikipedia while they perform more and more edits to article pages.

Table 1: Estimated parameters, standard errors (in brackets), and t-ratios. Parameters are significantly different from zero at the 5%-level if the t-ratio is at least 1.96. Significantly positive (negative) parameters indicate a higher (lower) hazard to drop out.

statistic	parameter (s.e.)	t-ratio
`edit`	-0.410 (0.061)	6.78
`discuss`	0.137 (0.068)	2.01
`organize`	0.220 (0.060)	3.69
`getFeedback`	0.365 (0.078)	4.66
`replyFeedback`	-0.140 (0.057)	2.44
`editControversial`	0.177 (0.036)	4.98
constant	-10.604 (0.405)	26.18

This is different for participation in discussion: the parameter associated with `discuss` is positive, indicating that users become more likely to drop out when they participated more in discussion pages. This dependency—which lead us, together with the result for the `edit` parameter, to the choice of our title—is not necessarily a causal relationship. It might be the case that users accumulate frustration due to some other unknown reason which, at the same time, has an increasing effect on the frequency of contributions to discussion. To get into the vicinity of causality it will be analyzed in future research whether different forms of discussion (e. g., un-replied threads vs. replied threads, or discussion patterns that resemble a flame war, see [17]) have different effects on the dropout hazard. Thereby we would gain insight into *how* Wikipedians should discuss such that reasons to drop out are attenuated. The participation on pages concerned with the organization of Wikipedia also has an increasing effect on the dropout hazard (positive value of the `organize` parameter).

Turning to the effects of feedback on user talk pages, we observe that if user u gets revisions on her own user talk page, then the dropout hazard of u increases (positive value of the `getFeedback` parameter); this effect is attenuated, if u herself participates to the discussion on her user talk page (negative value of the `replyFeedback` parameter). A possible explanation for the `getFeedback` parameter is that users might become involved into disputes which could result into the two effects that (1) they get complaints from other users on their user talk page and (2) they become more likely to drop out due to frustration. The negative value of the `replyFeedback` parameter indicates that users who respond to comments on their user talk page have a lower dropout hazard than users who do not respond—potentially being explained that the latter ones do not care anymore since they are already pondering about stop participating. Similar to the `discuss` statistics, it seems to be an important topic for future research to distinguish between positive feedback and negative feedback or, more generally, to find out how conversation on user talk pages should look like such that users are retained in Wikipedia.

The positive value of the `editControversial` parameter indicates that users editing controversial pages have a higher dropout hazard. This relationship seem to be very plausible since editing controversial pages involves confrontation with vandalism or edit wars, which might be a frustrating experience.

5. CONCLUSION AND FUTURE WORK

We presented a statistical framework to assess hypothetical causes for dropout from Wikipedia. The model defined in this paper is generally applicable to all kinds of data that may explain dropout in Wikipedia or other Web 2.0 applications—independent on whether the explanatory data stems from log files, questionary-based surveys, or other sources. The general model can be specialized to test specific hypotheses by plugging appropriate statistics into Eq. (2). The explanatory variables defined in Sect. 3.4 and used in Sect. 4 reflect a particular choice of hypothetical factors for dropout that will be extended in future research.

The most intriguing empirical result obtained in this paper is that participation in discussion seems to cause dropout rather than preventing it. Although several researchers have reported an increase in discussion in Wikipedia (e. g., [19, 8]), we are not aware of any previous quantitative work analyzing the effects of discussion. However, it is obvious that the results obtained in this paper are still very coarse, since only the number of contributions to talk pages has been counted and we did not distinguish between different forms of conversation. It is a promising topic for future research to relate various discussion patterns (see, e. g., [17]) to the dropout hazard, thereby revealing how frustrating discussion and how motivating discussion looks like. Furthermore, as we outlined in Sect. 3.2.1, our analysis is based on a specific selection of dropouts via the page of missing Wikipedians; it will be analyzed in future work whether alternative selection strategies lead to different and potentially more reliable results.

Another promising avenue for future research is to focus more on the effects of collaboration structure on the dropout hazard. We defined in [2] the *edit network* of Wikipedia pages encoding how users contribute to the page and how they interact with each other. It is very likely that certain patterns of users in these edit networks (e. g., getting deleted, getting restored, being a provider of novel content) or patterns of the global collaboration structure (e. g., bipolarity) influence the dropout hazard. If this can be validated we would identify collaboration patterns that are rewarding and motivating for Wikipedians and patterns that frequently lead to dropout and therefore to the loss of human capital.

Acknowledgments. We gratefully acknowledge financial support by the *Netherlands Organization for Scientific Research* (program *Networks of Networks*).

6. REFERENCES

[1] B. T. Adler and L. de Alfaro. A content-driven reputation system for the Wikipedia. In *Proc. 16th Intl. Conf. WWW*, pages 261–270, 2007.

[2] U. Brandes, P. Kenis, J. Lerner, and D. van Raaij. Network analysis of collaboration structure in Wikipedia. In *Proc. 18th Intl. World Wide Web Conf. (WWW2009)*, 2009, to appear.

[3] U. Brandes and J. Lerner. Visual analysis of controversy in user-generated encyclopedias. *Information Visualization*, 7:34–48, 2008.

[4] The battle for Wikipedia's soul. *The Economist*, March 6th, 2008.

[5] R. Garud, S. Jain, and P. Tuertscher. Incomplete by design and designing for incompleteness. *Organization Studies*, 29(3):351–371, 2008.

[6] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438:900–901, 2005.

[7] A. Kittur, E. Chi, B. A. Pendleton, B. Suh, and T. Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *Proc. SIGCHI Conf. Human Factors in Computing Systems*, 2007.

[8] A. Kittur, B. Suh, B. A. Pendleton, and E. H. Chi. He says, she says: conflict and coordination in Wikipedia. In *Proc. SIGCHI Conf. Human Factors in Computing Systems*, pages 453–462, 2007.

[9] J. F. Lawless. *Statistical Models and Methods for Lifetime Data*. Wiley, 2nd edition, 2003.

[10] T. Lento, H. T. Welser, L. Gu, and M. Smith. The ties that blog: Examining the relationship between social ties and continued participation in the Wallop weblogging system. In *Proc. 3rd Annual Workshop on the Weblogging Ecosystem*, 2006.

[11] P. R. Monge and N. S. Contractor. *Theories of Communication Networks*. Oxford University Press, 2003.

[12] J. Y. Moon and L. S. Sproull. The role of feedback in managing the Internet-based volunteer work force. *Information Systems Research*, 19(4):494–515, 2008.

[13] R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, destroying, and restoring value in Wikipedia. In *Proc. Intl. ACM Conf. Supporting Group Work*, pages 259–268, 2007.

[14] S. A. Snell and J. W. Dean Jr. Integrated manufacturing and human resource management: A human capital perspective. *The Academy of Management Journal*, 35(3):467–504, 1992.

[15] K. Stein and C. Hess. Does it matter who contributes? A study on featured articles in the German Wikipedia. In *Proc. 18th ACM Conf. Hypertext and Hypermedia (Hypertext 2007)*, pages 171–174, 2007.

[16] B. Suh, E. H. Chi, B. A. Pendleton, and A. Kittur. Us vs. them: Understanding social dynamics in Wikipedia with revert graph visualizations. In *Proc. IEEE VAST*, pages 163–170, 2007.

[17] T. C. Turner, M. A. Smith, D. Fisher, and H. T. Welser. Picturing usenet: Mapping computer-mediated collective action. *Journal of Computer-Mediated Communication*, 10(4), 2005.

[18] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proc. SIGCHI Conf. Human Factors in Computing Systems*, pages 575–582, 2004.

[19] F. B. Viégas, M. Wattenberg, J. Kriss, and F. van Ham. Talk before you type: Coordination in Wikipedia. In *Proceedings HICSS*, 2007.

[20] B.-Q. Vuong, E.-P. Lim, A. Sun, M.-T. Le, and H. W. Lauw. On ranking controversies in Wikipedia: Models and evaluation. In *Proc. Intl. Conf. Web Search and Web Data Mining*, pages 171–182, 2008.

[21] G. A. Young and R. L. Smith. *Essentials of Statistical Inference*. Cambridge University Press, 2005.

A new life for a dead parrot: Incentive structures in the Phrase Detectives game

Jon Chamberlain
University of Essex
School of Computer Science and Electronic Engineering
jchamb@essex.ac.uk

Massimo Poesio
University of Essex
School of Computer Science and Electronic Engineering
poesio@essex.ac.uk

Udo Kruschwitz
University of Essex
School of Computer Science and Electronic Engineering
udo@essex.ac.uk

He's passed on! This parrot is no more! He has ceased to be! He's expired and gone to meet his maker! He's kicked the bucket, he's shuffled off his mortal coil, run down the curtain and joined the bleedin' choir invisibile! THIS IS AN EX-PARROT! [1]

ABSTRACT

In order for there to be significant improvements in certain areas of natural language processing (such as anaphora resolution) large linguistically annotated resources need to be created which can be used to train, for example, machine learning systems. Annotated corpora of the size needed for modern computational linguistics research cannot however be created by small groups of hand-annotators. Simple Web-based games have demonstrated how it might be possible to do this through Web collaboration. This paper reports on the ongoing work of *Phrase Detectives*, a game developed in the ANAWIKI project designed for collaborative linguistic annotation on the Web. In this paper we focus on how we recruit and motivate players, incentivise high quality annotations and assess the quality of the data.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human factors; Human information processing; H.2.7 [**Artificial Intelligence**]: Natural Language Processing

Keywords

Web-based games, incentive structures, user motivation, distributed knowledge acquisition, anaphoric annotation

1. INTRODUCTION

The statistical revolution in natural language processing (NLP) has resulted in the first NLP systems and components which are usable on a large scale, from part-of-speech (POS) taggers to parsers [7]. However it has also raised the problem of creating the large amounts of annotated linguistic data needed for training and evaluating such systems. Potential solutions to this problem include semi-automatic annotation and machine learning methods that make better use of the available data. Unsupervised or semi-supervised techniques hold great promise, but for the foreseeable future at least, the greatest performance improvements are still likely to come from increasing the amount of data to be used by supervised training methods. These crucially rely on hand-annotated data. Traditionally, this requires trained annotators, which is prohibitively expensive both financially and in terms of person-hours (given the number of trained annotators available) on the scale required.

Recently, however, Web collaboration has emerged as a viable alternative. Wikipedia and similar initiatives have shown that a surprising number of individuals are willing to help with resource creation and scientific experiments. The Open Mind Common Sense project [16] demonstrated that such individuals are also willing to participate in the creation of databases for Artificial Intelligence (AI), and von Ahn showed that simple Web games are an effective way of motivating participants to annotate data for machine learning purposes [23].

The goal of the ANAWIKI project[1] is to experiment with Web collaboration as a solution to the problem of creating large-scale linguistically annotated corpora, both by developing Web-based annotation tools through which members of the scientific community can participate in corpus creation and through the use of game-like interfaces. We will present ongoing work on *Phrase Detectives*[2], a game designed to collect judgments about anaphoric annotations. We will also report results which include a substantial corpus of annotations already collected.

2. RELATED WORK

Related work comes from a range of relatively distinct research communities including, among others, Computational Linguistics / NLP, the games community and researchers working in the areas of the Semantic Web and knowledge representation.

Large-scale annotation of low-level linguistic information (part-of-speech tags) began with the Brown Corpus, in which very low-tech and time consuming methods were used. For the creation of the British National Corpus (BNC), the first 100M-word linguistically annotated corpus, a faster methodology was developed consisting of preliminary annotation with automatic methods followed by partial hand-correction [1]. This was made possible by the availability of relatively high quality automatic part-of-speech taggers (CLAWS).

With the development of the first high-quality chunkers, this methodology became applicable to the case of syntactic annotation. It was used for the creation of the Penn

[1] `http://www.textfiles.com/media/petshop`

WWW2009, April 20-24, 2009, Madrid, Spain.

[1] `http://www.anawiki.org`
[2] `http://www.phrasedetectives.org`

Treebank [10] although more substantial hand-checking was required.

Medium and large-scale semantic annotation projects (for wordsense or coreference) are a recent innovation in Computational Linguistics. The semi-automatic annotation methodology cannot yet be used for this type of annotation, as the quality of, for instance, coreference resolvers is not yet high enough on general text. Nevertheless the semantic annotation methodology has made great progress with the development, on the one end, of effective quality control methods [4] and on the other, of sophisticated annotation tools such as Serengeti [20].

These developments have made it possible to move from the small-scale semantic annotation projects, the aim of which was to create resources of around 100K words in size [14], to the efforts made as part of US initiatives such as Automatic Context Extraction (ACE), Translingual Information Detection, Extraction and Summarization (TIDES), and GALE to create 1 million word corpora. Such techniques could not be expected to annotate data on the scale of the BNC.

Collaborative resource creation on the Web offers a different solution to this problem. The motivation for this is the observation that a group of individuals can contribute to a collective solution, which has a better performance and is more robust than an individual's solution as demonstrated in simulations of collective behaviours in self-organizing systems [6].

Wikipedia is perhaps the best example of collaborative resource creation, but it is not an isolated case. The gaming approach to data collection, termed *games with a purpose*, has received increased attention since the success of the ESP game [22]. Subsequent games have attempted to collect data for multimedia tagging (*OntoTube*[3], *Tag a Tune*[4]) and language tagging (*Verbosity*[5], *OntoGame*[6], *Categorilla*[7], *Free Association*[8]). As Wikipedia has demonstrated however, there is not necessarily the need to turn every data collection task into a game. Other current efforts in attempting to acquire large-scale world knowledge from Web users include Freebase[9] and True Knowledge[10].

The *games with a purpose* concept has now also been adopted by the Semantic Web community in an attempt to collect large-scale ontological knowledge because currently "the Semantic Web lacks sufficient user involvement almost everywhere" [17].

It is a huge challenge to recruit enough users to make data collection worthwhile and, as we will explore later, it is also important to attract the right kind of player. Previous games have attracted exceptional levels of participation such as the ESP game (13,500 players in 4 months) [22], Peekaboom (14,000 players in 1 month) [24] and OpenMind (15,000 users) [16] which encourages one to believe mass participation might be possible for similar projects.

[3] http://www.ontogame.org/ontotube
[4] http://www.gwap.com/gwap/gamesPreview/tagatune
[5] http://www.gwap.com/gwap/gamesPreview/verbosity
[6] http://www.ontogame.org
[7] http://wordgame.stanford.edu/categorilla.html
[8] http://wordgame.stanford.edu/freeAssociation.html
[9] http://www.freebase.com
[10] http://www.trueknowledge.com

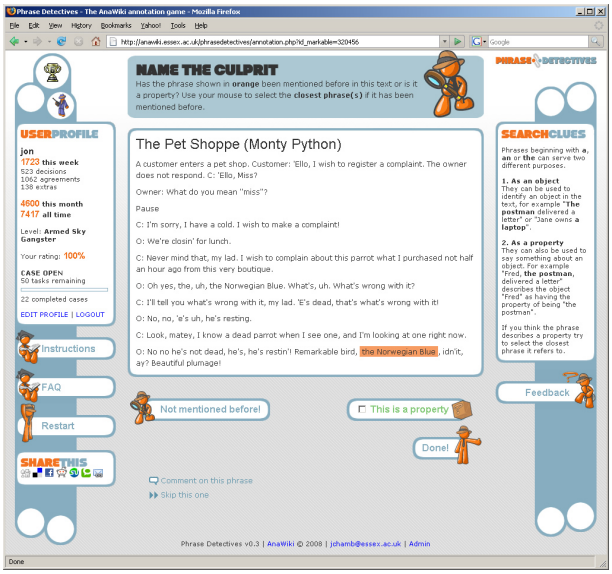


Figure 1: A screenshot of the Annotation Mode.

3. THE PHRASE DETECTIVES GAME

Phrase Detectives is a game offering a simple interface for non-expert users to learn how to annotate text and to make annotation decisions [2]. The goal of the game is to identify relationships between words and phrases in a short text. An example of a task would be to highlight an anaphor-antecedent relation between the markables (sections of text) *'This parrot'* and *'He'* in *'This parrot is no more! He has ceased to be!'* Markables are identified in the text by automatic pre-processing. There are two ways to annotate within the game: by selecting a markable that corefers to another one (Annotation Mode, called *Name the Culprit* in the game); or by validating a decision previously submitted by another player (Validation Mode, called *Detectives Conference* in the game).

Annotation Mode (see Figure 1) is the simplest way of collecting judgments. The player has to locate the closest antecedent markable of an anaphor markable, i.e. an earlier mention of the object. By moving the cursor over the text, markables are revealed in a bordered box. To select it the player clicks on the bordered box and the markable becomes highlighted. They can repeat this process if there is more than one antecedent markable (e.g. for plural anaphors such as *'they'*). They submit the annotation by clicking the *Done!* button. The player can also indicate that the highlighted markable has not been mentioned before (i.e. it is not anaphoric), that it is non-referring (for example, *'it'* in *'Yeah, well it's not easy to pad these Python files out to 150 lines, you know.'*) or that it is the property of another markable (for example, *'a lumberjack'* being a property of *'I'* in *'I wanted to be a lumberjack!'*). Players can also make a comment about the markable (for example, if there is an error in the automatic text processing) or skip the markable and move on to the next one.

In Validation Mode (see Figure 2) the player is presented with an annotation from a previous player. The anaphor markable is shown with the antecedent markable(s) that the previous player chose. The player has to decide if he agrees with this annotation. If not he is shown the Annotation

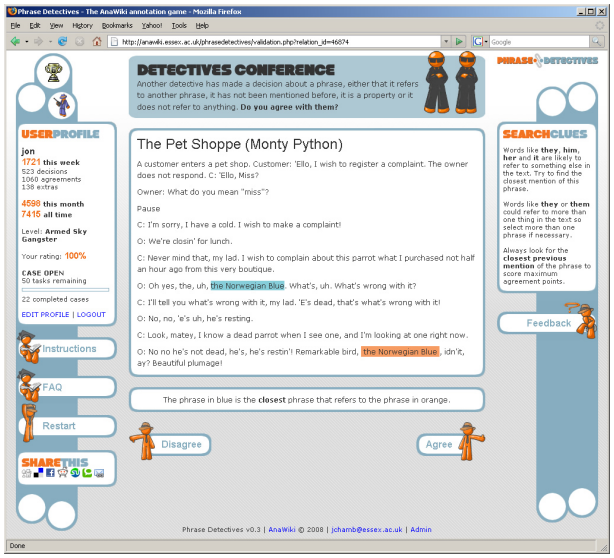


Figure 2: A screenshot of the Validation Mode.

Mode to enter a new annotation. The Validation Mode not only sorts ambiguous, incorrect and/or malicious decisions but also provides a social training mechanism [9].

When the users register they begin with the training phase of the game. Their answers are compared with Gold Standard texts to give them feedback on their decisions and to get a user rating, which is used to determine whether they need more training. Contextual instructions are also available during the game.

The corpus used in the game is created from short texts including: Wikipedia articles selected from the 'Featured Articles' and the page of 'Unusual Articles'; stories from Project Gutenberg including Aesop's Fables, Sherlock Holmes and Grimm's Fairy Tales; and dialogue texts from Textfile.com including Monty Python's Dead Parrot sketch. Selections from the GNOME and ARRAU corpora are also included to analyse the quality of the annotations.

4. THE SCORING SYSTEM

One of the most significant problems when designing a game that collects data is how to reward a player's decision when the correct answer is not known (and in some cases there may not be just one correct answer). Our solution is to motivate players using comparative scoring (awarding points for agreeing with the Gold Standard) and collaborative scoring (increasing the reward the more the players agree with each other).

In the game groups of players work on the same task over a period of time as this is likely to lead to a collectively intelligent decision [21]. An initial group of players are asked to annotate a markable. For each decision the player receives a 'decision' point. If all the players agree with each other then they are all awarded an additonal 'agreement' point and the markable is considered complete.

However it is likely that the first group of players will not agree with each other (62% of markables are given more than one relationship). In this case each unique relationship for the markable is validated by another group of players. The validating players receive an 'agreement' point for every player from the first group they agree with (either by agreeing or disagreeing). The players they agree with also receive an 'agreement' point.

This scoring system motivates the initial annotating group of players to choose the best relationship for the markable because it will lead to more points being added to their score later. The validating players are motivated to agree with these relationships as they will score more agreement points.

Contrary to expectations [3] it took players almost twice as long to validate a relationship than to annotate a markable (14 seconds compared to 8 seconds).

5. INCENTIVE STRUCTURES

The game is designed to use 3 types of incentive structure: personal, social and financial. All incentives were applied with caution as rewards have been known to decrease annotation quality [12]. The primary goal is to motivate the players to provide high quality answers, rather than large quantities of answers.

- Document topic
- Task speed
- User contributed documents
- Leaderboards
- Collaborative scoring
- Weekly and monthly prizes

5.1 Personal incentives

Personal incentives are evident when simply participating is enough of a reward for the user. For example, a Web user submitting information to Wikipedia does not usually receive any reward for what they have done but are content to be involved in the project. Similarly the progress of a player through a computer game will usually only be of interest to themselves, with the reward being the enjoyment of the game.

Generally, the most important personal incentive is that the user feels they are contributing to a worthwhile project. News and links to the research were posted on the homepage to reinforce the credibility of the project.

Also important for the players of Phrase Detectives is that they read texts that they find interesting. The choice of documents is important in getting users to participate in the game, to understand the tasks and to keep playing. Players can specify a preference for particular topics, however only 4% do so. This could be an indication that the corpus as a whole was interesting but it is more likely that they simply didn't change their default options [11].

It is also important for the players to read the documents at a relatively normal speed whilst still being able to complete the tasks. By default the tasks are generated randomly (although displayed in order) and limited (50 markable tasks selected from each document) which allows a normal reading flow. Players are given bonus points if they change their profile settings to select every markable in each document (which makes reading slower). Only 5% of players chose to sacrifice readability for the extra points.

In early versions of the game the player could see how long they had taken to do an annotation. Although this had no

Figure 3: A screenshot of the player's homepage.

influence on the scoring, players complained that they felt under pressure and that they didn't have enough time to check their answers. This is in contrast to previous suggestions that timed tasks motivate players [23]. The timing of the annotations is now hidden from the players but still recorded with annotations. The relationship between the time of the annotation, the user rating and the agreement will be crucial in understanding how a timed element in a reading game influences the data that is collected.

The throughput of Phrase Detectives is 450 annotations per human hour (compared to the ESP game at 233 labels per human hour [23]). There is, however, a difference in data input between the 2 games, the former only requiring clicks on pre-selected phrases and the latter requiring the user to type in a phrase. The design of a game task must consider the speed at which the player can process the input source (e.g. text, images) and deliver their response (e.g. a click, typing) in order to maximise throughput and hence the amount of data that is collected.

We allowed users to submit their own text to the corpus. This would be processed and entered into the game. We anticipated that, much like Wikipedia, this would motivate users to generate content and become much more involved in the game. Unfortunately this was not the case, with only one user submitting text. We have now stopped advertising this incentive however the concept may still hold promise for games where the user-submitted content is more naturally created (e.g. collaborative story writing).

5.2 Social incentives

Social incentives reward users by improving their standing amongst their peers (in this case their fellow players).

Phrase Detectives features the usual incentives of a computer game, including weekly, monthly and all-time leaderboards, cups for monthly top scores and named levels for reaching a certain amount of points (see Figure 3). Interesting phenomenon have been reported with these reward mechanisms, namely that players gravitate towards the cut-off points (i.e. they keep playing to reach a level or high score before stopping) [24]. The collaborative agreement scoring in Phrase Detectives prevents us from effectively analysing this (as players continue to score even when they have stopped playing) however our high-scoring players can be regularly seen outscoring each other on the leaderboards.

In addition to the leaderboards that are visible to all players, each player can also see a leaderboard of other players who agreed with them. Although there is no direct incentive (as you cannot influence your own agreement leaderboard) it reinforces the social aspect of how the scoring system works. The success of games integrated into social networking sites like Sentiment Quiz[11] on Facebook indicates that visible social interaction within a game environment motivates the players to contribute more.

5.3 Financial incentives

Financial incentives reward effort with money. We introduced a weekly prize where a player is chosen by randomly selecting an annotation made during that week. This prize motivates low-scoring players because any annotation made during the week has a chance of winning (much like a lottery) and the more annotations you make, the higher your chance of winning.

We also introduced monthly prizes for the 3 highest scorers of the month. The monthly prize motivates the high-scoring players to compete with each other by doing more work, but also motivates some of the low-scoring players in the early parts of the month when the high score is low.

The weekly prize was £15 and the monthly prizes were £75, £50 and £25 for first, second and third places. The prizes were sent as Amazon vouchers by email.

6. QUALITY OF DATA

The psychological impact of incentive structures, especially financial ones, can create a conflict of motivation in players (i.e. how much time they should spend on their decisions). They may decide to focus on ways to maximise rewards rather than provide high quality answers. The game's scoring system and incentive structures are designed to reduce this to a minimum. We have identified four aspects that need to be addressed to control annotation quality: ensuring users understand the task; attention slips; malicious behaviour; and genuine ambiguity of data [9].

Further analysis will reveal if changing the number of players in the annotating and validating groups will effect the quality of the annotations. The game currently uses 8 players in the annotating group and 4 in the validating group with an average of 18 players looking at each markable. Some types of task can achieve high quality annotations with as few as 4 annotators [18] but other types of tasks (e.g anaphor resolution) may require more [15].

7. ATTRACTING & MOTIVATING USERS

The target audience for the game are English-speakers who spend significant amounts of time online, either playing computer games or casually browsing the Internet.

In order to attract the number of participants required to make a success of this methodology it is not enough to develop attractive games, but also successful advertising. Phrase Detectives was written about in local and national press, on science websites, blogs, bookmarking websites and

[11] `http://www.modul.ac.at/nmt/sentiment-quiz`

gaming forums. The developer of the game was also interviewed by the BBC. At the same time a pay-per-click advertising campaign was started on the social networking website Facebook, as well as a group connected to the project.

We investigated the sources of traffic since live release using Google Analytics. Incoming site traffic didn't show anything unusual: direct (46%); from a website link (29%); from the Facebook advert (13%); from a search (12%). However the bounce rate (the percentage of single-page visits, where the user leaves on the page they entered on) revealed how useful the traffic was. This showed a relatively consistent figure for direct (33%), link (29%) and search (44%) traffic. However for the Facebook advert it was significantly higher (90%), meaning that 9 out of 10 users that came from this source did not play the game. This casts doubt over the usefulness of pay-per-click advertising as a way of attracting participants to a game.

The players of Phrase Detectives were encouraged to recruit more players by giving them extra points every time they referred a player and whenever that player gained a level. The staggered reward for referring new players was to discourage players from creating new accounts themselves in order to get the reward. The scores of the referred players are displayed to the referring player on the recruits leaderboard. 4% of players have been referred by other players.

Attracting large numbers of players to a game is only part of the problem. It is also necessary to attract players who will make significant contributions. Since its release the game has attracted 750 players but we found that the top 10 players (5% of total) had 60% of the total points on the system and had made 73% of the annotations. This indicates that only a handful of users are doing the majority of the work, which is consistent with previous findings [18], however the contribution of one-time users should not be ignored [8]. Most of the players who have made significant contributions have a language-based background.

Players are invited to report on their experiences either through the feedback page or by commenting on a markable. Both methods send a message to the administrators who can address the issues raised and reply to the player if required. General feedback included suggestions for improvements to the interface and clarification of instructions and scoring. Frequent comments included reporting markables with errors from the pre-processing and discussing ambiguous or difficult markable relations.

It was intended to be a simple system of communication from player to administrator that avoids players colluding to gain points. However it is apparent that a more sophisticated community message system would enhance the player experience and encourage the development of a community.

8. IMPLEMENTATION

Phrase Detectives is running on a dedicated Linux server. The pre-processed data is stored in an MySQL database and most of the scripting is done via PHP.

The Gold Standard is created in Serengeti (a Web-based annotation tool developed at the University of Bielefeld [20]) by computational linguists. This tool runs on the same server and accesses the same database.

The database stores the textual data in Sekimo Generic Format (SGF) [19], a multi-layer representation of the original documents that can easily be transformed into other common formats such as MAS-XML and PAULA. We apply a pipeline of scripts to get from raw text to SGF format. For English texts this pipeline consists of these main steps:

- A pre-processing step normalises the input, applies a sentence splitter and runs a tokenizer over each sentence. We use the *openNLP*[12] toolkit to perform this process.
- Each sentence is analysed by the *Berkeley Parser*[13].
- The parser output is interpreted to identify markables in the sentence. As a result we create an XML representation which preserves the syntactic structure of the markables (including nested markables, e.g. noun phrases within a larger noun phrase).
- A heuristic processor identifies a number of additional features associated with markables such as person, case, number etc. The output format is MAS-XML.

The last two steps are based on previous work within the research group at Essex University [15]. Finally, MAS-XML is converted into SGF. Both MAS-XML and SGF are also the formats used to export the annotated data.

9. RESULTS

Before going live we evaluated a prototype of the game interface informally using a group of randomly selected volunteers from the University of Essex [2]. The beta version of *Phrase Detectives* went on-line in May 2008, with the first live release in December 2008. Over 1 million words of text have been added to the live game.

In the first 3 months of live release the game collected over 200,000 annotations and validations of anaphoric relations. To put this in perspective, the GNOME corpus, produced by traditional methods, included around 3,000 annotations of anaphoric relations [13] whereas OntoNotes[14] 3.0, with 1 million words, contains around 140,000 annotations.

The analysis of the results is an ongoing issue. However, by manually analyzing 10 random documents we could not find a single case in which a misconceived annotation was validated by other players. This confirms the assumptions we made about quality control. It will need to be further investigated by more thorough analysis methods which will be part of the future work.

10. CONCLUSIONS

The incentives structures used in Phrase Detectives were successful in motivating the users to provide high quality data. In particular the collaborative and social elements (agreement scoring and leaderboards) seem to offer the most promise if they can be linked with existing social networks.

The methodology behind collaborative game playing has become increasingly more widespread. Whilst the good-will of Web volunteers exists at the moment, there may be a point of saturation, where it becomes significantly more difficult to attract users and more novel incentive structures will need to be developed.

[12] `http://opennlp.sourceforge.net`
[13] `http://nlp.cs.berkeley.edu`
[14] `http://www.ldc.upenn.edu`

11. FUTURE WORK

We are progressively converting text for use in the game with the aim of having 100 million words. So far, mainly narrative texts from Project Gutenberg and encyclopedic texts from Wikipedia have been converted. We also plan to include further data from travel guides, news articles, and the American National Corpus [5].

It has become evident that working with a corpus of that size will require additional types of users. New tasks need to be developed, some as game tasks and others as admin player tasks that allow the management of players and documents to be handled by the users themselves. Motivating admin players will require very different incentive structures than have been used so far in the game.

The data collected by the game will be made available to the community through the Anaphoric Bank[15].

Ultimately, the usefulness of the annotated data will need to be shown by, for example, successfully training anaphora resolution algorithms that perform better than existing systems.

Acknowledgments

ANAWIKI is funded by a grant from the Engineering and Physical Sciences Research Council (EPSRC), grant number EP/F00575X/1. Thanks to Daniela Goecke, Nils Diewald, Maik Stührenberg and Daniel Jettka (University of Bielefeld), Mark Schellhase (University of Essex) and all the players who have contributed to the project, in particular *livio.robaldo*, *trelex*, *VB*, *TLS* and *Lupian*.

12. REFERENCES

[1] L. Burnard. The British National Corpus Reference guide. Technical report, Oxford University Computing Services, Oxford, 2000.

[2] J. Chamberlain, M. Poesio, and U. Kruschwitz. Phrase Detectives: A Web-based Collaborative Annotation Game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics'08)*, Graz, 2008.

[3] T. Chklovski and Y. Gil. Improving the design of intelligent acquisition interfaces for collecting world knowledge from web contributors. In *K-CAP '05: Proceedings of the 3rd international conference on Knowledge capture*, pages 35–42, New York, NY, USA, 2005. ACM.

[4] E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. OntoNotes: The 90% Solution. In *Proceedings of HLT-NAACL06*, 2006.

[5] N. Ide and C. Macleod. The American National Corpus: A Standardized Resource of American English. In *Proceedings of Corpus Linguistics*, Lancaster, 2001.

[6] N. L. Johnson, S. Rasmussen, C. Joslyn, L. Rocha, S. Smith, and M. Kantor. Symbiotic Intelligence: Self-Organizing Knowledge on Distributed Networks Driven by Human Interaction. In *Proceedings of the Sixth International Conference on Artificial Life*. MIT Press, 1998.

[7] D. Jurafsky and J. H. Martin. *Speech and Language Processing- 2^{nd} edition*. Prentice-Hall, 2008.

[8] B. Kanefsky, N. Barlow, and V. Gulick. Can distributed volunteers accomplish massive data analysis tasks? *Lunar and Planetary Science*, XXXII, 2001.

[9] U. Kruschwitz, J. Chamberlain, and M. Poesio. (Linguistic) Science Through Web Collaboration in the ANAWIKI Project. In *Proceedings of WebSci'09*, Athens, 2009.

[10] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of english: the Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[11] K. Markey. Twenty-five years of end-user searching, Part 1: Research findings. *Journal of the American Society for Information Science and Technology (JASIST)*, 58(8):1071–1081, June 2007.

[12] J. Mrozinski, E. Whittaker, and S. Furui. Collecting a why-question corpus for development and evaluation of an automatic QA-system. In *Proceedings of ACL-08: HLT*, pages 443–451, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[13] M. Poesio. Discourse annotation and semantic annotation in the gnome corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*, 2004.

[14] M. Poesio. The MATE/GNOME scheme for anaphoric annotation, revisited. In *Proceedings of SIGDIAL*, 2004.

[15] M. Poesio and R. Artstein. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation*, pages 76–83, 2005.

[16] P. Singh. The public acquisition of commonsense knowledge. In *Proceedings of the AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*, Palo Alto, CA, 2002.

[17] K. Siorpaes and M. Hepp. Games with a purpose for the semantic web. *IEEE Intelligent Systems*, 23(3):50–60, 2008.

[18] R. Snow, O'Connor, D. Jurafsky, and A. Ng. Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. *Proceedings of EMNLP-08*, Jan 2008.

[19] M. Stührenberg and D. Goecke. SGF - An integrated model for multiple annotations and its application in a linguistic domain. In *Proceedings of Balisage: The Markup Conference*, Montreal, 2008.

[20] M. Stührenberg, D. Goecke, N. Diewald, A. Mehler, and I. Cramer. Web-based annotation of anaphoric relations and lexical chains. In *Proceedings of the ACL Linguistic Annotation Workshop*, pages 140–147, 2007.

[21] J. Surowiecki. *The Wisdom of Crowds*. Anchor, 2005.

[22] L. von Ahn. Games with a purpose. *Computer*, 39(6):92–94, 2006.

[23] L. von Ahn and L. Dabbish. Designing games with a purpose. *Communications of the ACM*, 51(8):58–67, 2008.

[24] L. von Ahn, R. Liu, and M. Blum. Peekaboom: a game for locating objects in images. In *Proceedings of CHI '06*, pages 55–64, 2006.

[15] `http://www.anaphoricbank.org`

YouTube's Collaborative Annotations

Sigalit Bar, Aviad Barzilai, Isaac Elias, Michael Fink*, Julian Frumar, Herb Ho,
Ryan Junee, Nir Kerem, Simon Ratner, Jasson Schrock and Ran Tavory

YouTube Interactive Video Annotations Group

901 Cherry Ave. San Bruno, CA 94066, USA

ABSTRACT

More and more YouTube videos no longer provide a passive viewing experience, but rather entice the viewer to interact with the video by clicking on objects with embedded links. These links are part of *YouTube's Annotations* system, which enables content owners to add active overlays on top of their videos. YouTube Annotation overlays also enable adding dynamic speech bubbles and pop-ups which can function as an ever-changing layer of supplementary information and entertainment, augmenting the video experience. This paper addresses the question of whether the ability to add annotation overlays on a given video should be opened to the YouTube public. The basic dilemma in opening a video to *collaborative annotations* is derived from the tension between the benefits of collaboration and the risks of visual clutter and spam. We term the degree to which a video is open to external contributions as the *collaboration spectrum*, and describe several models that let content owners to explore this spectrum in order to find the optimal way to harness the power of the masses.

Keywords

YouTube, annotations, collaborative, spectrum, video, tagging.

1. INTRODUCTION

YouTube, the world's largest multi-media platform, reached its phenomenal success by catering to the needs of three sets of users. The first are **passive users** who merely want to enjoy YouTube videos and wish to receive smooth and high quality video streams. The second are **active users** who are mostly motivated by the chance of having their voice heard and are engaged on the YouTube site by commenting or rating. Finally the third group includes the **content contributing users** who generate the site's videos and are mostly motivated by the potential to increase their video and channel exposure. Naturally, the number of content contributing users is dominated by the number of users who might be active on the site but, do not actually upload videos. While YouTube was very successful in reaching an impressive market share, it remains diligently focused on improving other success metrics such as, increasing search quality, developing monetization opportunities and most importantly augmenting the satisfaction of the three user groups. We believe that many of these strategic goals can be addressed by harnessing user generated video metadata. Following this reasoning, YouTube launched in June 2008, the *Interactive Video Annotations* tool which enables content owners to add visual overlays on top of their YouTube videos. These overlays can contain texts (shaped as pop-ups. speech bubbles and spotlights), hyperlinks (to videos, channels, etc.) and time controls (used to seek or pause the video). In less than eight months, millions of videos have included overlay annotations. Profiling the existing usage of the tool, we see three levels of sophistication and user engagement:

1. **Commentary** with examples such director's commentary, animal dubbing and How-To videos for anything from Japanese origami to guitar and dance lessons.

2. **Outbound links** include anything from one link to a full menu leading to other independent videos, or channel pages.

3. **Interconnecting links** are used for creating branching storyboards with many videos and links woven into an elaborate scene of interaction. In this category we find interactive games, card tricks, truth or dare videos, and virtual pets [1].

Interactive videos are currently viewed tens of millions of times each day, with certain embedded links exhibiting high click-through-rates. Thus for example, the very first card trick video with 9M views, required viewers to choose one of six cards. It can be observed that the follow-up videos (representing the six individual card selections) sum up to 9M, a similar total of views.

2. COLLABORATIVE ANNOTATIONS

Owner generated annotations have enjoyed a high adoption rate however, we have previously pointed out that video owners are still a fairly small percentage of the YouTube community. This paper focuses on describing several models in which content owners can share the power to add annotation overlays with the general YouTube community. The basic dilemma in opening a video to *collaborative annotation* is derived from the tension between the benefits of collaboration and the risks of visual clutter and spam. We term the degree in which a video owner decides to open his video to external contributions as the *collaboration spectrum*.

Figure 1. The new in-video interface for editing YouTube Annotations contains a next / previous annotation navigator (A). By clicking on the video, users can add a default annotation (B), and using the contextual menu (C) they can embed links and format appearance. Finally, the temporal extent is defined using the time controls (D).

As a content owner gains trust that he will not be required to filter large quantities of clutter and spam from his videos, he will be more inclined to share the power to add annotations. On the other hand the collaborators are more inclined to add their contributions as the visibility and exposure of their contribution increases. In order to maintain the mutual trust between content owners and collaborators our first exploration of the collaboration spectrum enabled content owners to invite trusted family members, friends and peers to add annotations on their videos. The invitation process was based on sharing a URL which functions as an access key to the video's annotations' editor in Figure 1. It should be noted that the access key can be sent on by the collaborators to other users that they trust. However, what is important is that the model maintains a chain of trust. In order to extend this mutual trust and accountability we chose a user interface which clearly attributes each collaborative annotation to the contributing user (see for example the annotation by "devball2376" in Figure 2 B).

It could be expected that the owners' incentives to augment video experience will occasionally collide with the collaborators incentive to add their perspective to the video. We therefore expect a learning curve in which content owners gradually discover their trustworthy collaborators. In addition, we always maintain that a content owner have a simple interface for erasing any cluttering contributions and can easily revoke the existing access key and resend a new access key to a revised list of collaborators. Finally, it is important to note that we identify the collaborators' main incentive to contribute is proportional to the video's visibility. It was this motivation that caused us to reject models which provide an individual "layer" or copy of the video for each collaborator (a model favored by some competitors [2]).

3. PUBLIC ANNOTATIONS

The collaborative annotations model presented in the previous section is ideal for scenarios where the set of potential collaborators is known in advance. However, in certain scenarios content owners might wish to further explore the collaboration spectrum, and open their video to external contributions of the general YouTube public. We therefore launched an additional model of public annotation editing access which is granted by clicking on a *public collaborative link* embedded in the video by the owner (e.g. the red region in Figure 2 A). One nice example of this model can be found in "Kevin N' George must die" which is the very first collaborative sitcom. In this series Kevin (coat and mustache) and George (blonde wig and a yellow shirt) die at the end of each episode, unless the viewer intervenes and clicks on the rescuing object. This object has a hidden embedded link (e.g. the open cell phone in Figure 2 C) which saves the heroes by transferring them to the next episode. The YouTube community is challenged to generate more and more episodes in order to extend our heroes' lives. Once a user created a follow-up episode he adds the relevant clues and comments on the collaborative billboard and embeds a link on the object that transfers Kevin and George to the next episode. Although filtering spam and clutter is part of the daily task, these videos have become a bustling scene of activity with four episodes aired (and hopefully many more to come).

It should be noted that in this model of public annotations, content owners can filter annotations by spatial-temporal constraints, so that annotations not appearing within the designated areas (e.g. the collaborative billboard) are erased.

Figure 2. "Kevin N' George must die" the first collaborative sitcom challenges the viewers to generate the next episode.

Another model we have seen is based on the natural filtering of views. Many videos (such as family videos) are typically watched by a small clique of viewers and thus the incentive to search for them and spam them is negligible. A nice example of this scenario was a filmed interview of the French philosopher Claude Lévi Strauss which was uploaded with the *public collaborative link* stating "Help me translate this video". The video was fully translated in a collaborative manner within two weeks.

4. SUMMARY AND FUTURE WORK

We have described several models along the *collaboration spectrum*, in which content owners can share the power for collaborative annotations within a closed group of friends and family or with a wider viewer community. We also discussed how spatial-temporal constraints can reduce the clutter by collaborative annotations. Recently, we have added a new spotlight style that enables users to add hidden links that only appear on mouse-over so that content owners will be able to more freely explore the collaboration spectrum (e.g. as in Facebook image tagging [3]). We envision that using these hidden links users will be able to do anything from tagging objects in videos to creating elaborate treasure hunt games.

While we built the YouTube Annotations as a fun and engaging platform, they evolved to be one of the world's largest repositories of video metadata. This metadata will hopefully contribute to YouTube's content indexing and targeted advertising efforts. Rather than requiring the video owner to provide general tag words to the video during the upload process, we provide the incentive to augment the video experience by adding dynamic textual layers. These textual overlays can provide high-grain video metadata, localized in time and space. With the right incentives for collaborative annotations this process is expected to rapidly change the world of online video.

*** All correspondence should be directed to Michael Fink: Email: fink@google.com Tel: +972-542451115**

5. REFERENCES

[1] C. Lawton 2009. Video Sites Entice Users To Stay and Play. Wall Street Journal http://online.wsj.com/article/SB123370933597245913.html

[2] Plymedia www.plymedia.com.

[3] Facebook www.facebook.com

Mixing Financial, Social and Fun Incentives for Social Voting

Position Paper[1]

Frank Smadja
Toluna
MATAM POB 15075
Haifa 31905
frank.smadja@toluna.com

ABSTRACT

In this paper we examine the types of incentives available on the Web and compare them in terms of effectiveness, i.e., how well they motivate users. Our focus is on market-research, where the goal is to get people's opinions on a variety of topics and encourage them to answer surveys. We illustrate our point with examples taken from Toluna (http://www.toluna.com), a social voting site launched in 7 European countries and soon to be launched in the US. We show how a combination of financial and non-financial rewards is necessary to fuel continued interest and motivation on a user community. We classify incentives based on the intent behind them into : (1) recruiting incentives whose purpose is to allow attract new users, (2) content and retaining incentives that encourage existing users to remain active on the site, and finally (3) participation incentives that can entice users to participate in surveys, which might or might not be of interest to them. We discuss and compare the three types of incentives and give experimental results.

Categories and Subject Descriptors

H. Information Systems > H.1 Models And Principles > H.1.2 User/Machine Systems .

General Terms

Measurement, Documentation, Economics, Experimentation, Human Factors.

Keywords

Market research, polling, surveys, incentives and reward, financial incentives, social reward.

1. Our Experimental Ground: Toluna

Toluna is one of the most active social site focused on voting community, i.e., a Web2.0 site completely geared towards polls, surveys and opinions of users. Toluna members can voice their opinion on any topic but they can also poll the community and get other users' opinions. Some users actively participate in survey activities, while others are on the site out of pure curiosity and remain more passive. Toluna currently counts more than 2 million active users. In January 2009 alone, users voted 8 million times on the site, created 35,000 polls and topics and expressed some 200,000 full text opinions on a huge range of topics.

Toluna is a market research site in the sense that its focus is on getting people's opinions on many topics, whether commercial or mundane. The ultimate purpose of the site is for people to enjoy the site experience, while giving valuable opinions. Our conjecture is that the face of market research can be changed by adding social interaction, the same way it is done in other Internet activities such as Question Answering (See Yahoo! Answers), shopping (see Ebay.com), Eating out (Yelp.com), Videos (Youtube), Movies (Yahoo! Movies), etc.

Toluna is pioneering market research by being as far as we know the first "social voting" site, in the sense that users have their own user page and presence, produce valuable content, share activities in addition to the traditional survey-answering activity so as to make the entire visit a pleasant experience. We have conducted a number of experiments and surveys regarding incentives on Toluna and acquired insights in the process that we present and discuss in the rest of this paper.

2. Various Types of Incentives for Market Research

This section lists the various types of rewards and incentives mechanisms used on the Web and discuss them as tools for motivating users. Our terminology mostly draws on Raban [2].

2.1 Explicit Incentives

Explicit incentives refer to material rewards such as payment or other tangible benefit [2]. In market research, this encompasses two types of rewards: financial reward and prize draws.

A financial reward is anything that can be translated into cash or savings such as plain cash naturally, but also vouchers, coupons, frequent flyers mileage, etc. These incentives are more or less attractive to the user, and more or less costly to the company. But the principle is always the same, and it always translates into an item that can be directly or indirectly monetized by the user.

Lotteries and prize draws are explicit rewards, which are characterized by the fact that they are spread across fewer users and thus permit awarding significantly larger amounts of money per winning user. Prize draws can be periodic (monthly, yearly, etc.) or for one-time events specific to some survey participation. Prize draws range from pure cash to other material gifts, such as free tickets, free products to test, etc.

Lotteries and prize draws do attract some users but mostly as a complementary incentive program; they would not be enough to stimulate interest without the other types of incentives.

[1] Presented at Webcentives09, 1st International Workshop on Motivation and Incentives on the Web. Part of the 18th International World Wide Web conference, WWW2009, Madrid, April 2009. http://www2009.org

2.2 Implicit Incentives

Implicit incentives are typically driven by some kind of intrinsic motivation not based on anything tangible. For our purpose, we distinguish here between two types of implicit incentives: social and user experience incentives.

Social Incentives

Social incentives cover a wide range of mechanisms and features that allow the user to "feel good" as an active member of the community of users. These incentives are in general a way to highlight the impact of a specific user on the other users but can take various forms such as:

- Relationships among users: friends, people who know you, people who follow you, who visited your user page, people who agree with you, people who rated you, people who answered your polls, etc.
- Special placement or role on the site: the user can be part of the "most influent" users, the user picture can be placed on the homepage, user content could be elected as significant, etc.
- Awards and Social level: users with more seniority can be awarded titles such as "Expert user" that might grant the user some extra privileges.

A good example of a site making good use of social rewards is Yahoo! Answers (http://answers.yahoo.com), which offers a sophisticated way of rating users by their activity on the site (Yahoo! Answers points system is described here: http://answers.yahoo.com/info/scoring_system). In [7] Raghavan is quoted saying that 4 % to 6% or all Yahoo! users are drawn to contribute their energy for free on any of Yahoo! properties based on user generated content (Yahoo! movies, answers, etc.).

Money might be a great motivator, but users' motivation involves far more aspects than just financial. Social rewards seem to trigger deep psychological reactions that can sometimes bring very high motivation, it has been shown that social and financial rewards are coded differently in the brain, and even gender can have an influence here [1].

We have experienced this phenomenon on Toluna, when, on several occasions, users have directly expressed it as in the following example: "*I myself am a member of quite a few different survey sites, but I honestly prefer Toluna to all the other ones. Toluna though is not just a survey site it is a site that helps you to stay in touch with other members similar to yourself, even though you will never become a millionaire from using Toluna <…>*[2]"

Also Toluna has a social level grading system that allows users to move from the novice rank to the expert level and then to a VIP level. A user being promoted to a VIP level is usually very happy usually posts it on the site again as witnessed by their posts on the site[3].

[2] http://uk.toluna.com/opinions/233304/Which-other-survey-sites-member-which-your.htm

[3] http://www.toluna.com/opinions/236420/Muchas-gracias-experto.htm, http://fr.toluna.com/opinions/233211/Finalement-aussi-suis-passee-merci-tout-monde.htm, http://fr.toluna.com/opinions/232951/Me-voici-merci-tous.htm

Personal User Experience

The key additional implicit incentive is user experience on the site, which has a direct impact in our opinion on users' happiness. The user experience is a major motivator for users. More specifically we claim that the following two points are critical:

- Rich and entertaining content: The richer the content the more interesting the site and the more the user will be willing to spend time exploring the site.
- Good site design: This encompasses several aspects including look and feel, but also response time, navigation, search engine, integration with other social tools, etc.

On Toluna, the average user visit is longer than 7 minutes with an average of 8 pages viewed per visit. It is clear that much more than simple financial motivation is at play here and that users enjoy and are motivated to participate on the site.

In Figure 3, 4 and Figure 5 we show how users with different mindsets are attracted to different types of activities on the site. Figure 3 shows the overall responses to a random sample of 2,000 active users asking them to explain their motivation for being on the site. You can see that over 70% of them like the financial rewards but that also the fact that people can express their opinion, while looking at other people opinion is also popular. Figure 4 shows the favorite features of the users who selected any of the financial rewards in the first question; the answer is clear here, as expected, the large majority of users like the vouchers, the sponsored polls (bringing them points), etc. Finally, Figure 5 shows the favorite features of the users who selected the more social rewards, here we can see that the choices are more balanced and that more people like the social aspects in favor of the pure financial.

3. Explicit Financial Incentives, the Good, the Bad and the Ugly

The market research industry has a long tradition of relying on purely financial incentives in order to guarantee answers to all the surveys. Recruiting campaigns are mostly based on simple messages such as "Become an instant cash winner," "Get paid to give your opinion," "Get paid for completing surveys from your home," etc. The whole market research industry has been working solely on these types of messages for a long time. We discuss in this section how this approach is limited by looking at advantages and drawbacks of purely financial incentives.

3.1 The Good

Promising users some amount of money for spending time on a site or answering a few questions, always attracts users. Indeed, the appeal of cash is a simple message that is understood by all. In some cases, cash might even be the only way to attract people to answer personal questions. For example, while users easily discuss topics such as coffee, shopping habits, etc. it is harder to get them to talk about banking, insurance or even ear-care products. The difference in participation can vary by an order of magnitude if the topic is of interest or properly illustrated with images and/or video and attractive questions. Financial incentives are a sure way to attract new users and keep on fueling the ecosystem even for non-attractive surveys.

3.2 The Bad

Once users have been promised some kind of financial reward the relationship is changed, the expectations become different and not always for the best. For example, a user looking for quick money will tend to judge the site solely by how easy it is to get his reward.

In addition, a user clicking through a survey in order to make a few dollars might be inclined to give non necessarily accurate answers, s/he will tend to consider this task as one would consider a boring and poorly paid job. This is a sure way to generate frustration as well as poor quality answers.

Note that financial rewards do not systematically have a negative effect, and is in several cases not only worth the cost, but even necessary to fuel many aspects of the Web ecosystem. Raban [2] demonstrated how monetary incentives, when used in conjunction with social rewards are correlated with high quality content in some specific cases.

3.3 The Ugly

Using financial incentives as the main reward system is in fact turning the user community into the site's workforce and thus creating an implicit employer/employee relationship with all its problems like raises, timely payments, motivation and retention plans, on a very large scale. Dealing with a workforce of several millions is no small feat as can be imagined, but also the frustrated user will tend to be very vocal and intolerant to errors and other quirks that may affect their perceived performance of the system. This can have negative impact on the site's ability to maintain an active user base in the long term.

Moreover, financial incentives tend to be less effective with long-term users. Indeed, it might be very attractive for new users to get paid for answering surveys, but after a while stickiness is mostly affected by other factors such as quality of the content or the sense to belong to a community, the will to impact new and future products, etc. On Toluna, we have noticed that users that are involved exclusively for the money, tend to show fatigue and their response rate tend to decrease faster than users who combine several types of activities on the site. When reaching this fatigue stage, a pure financial incentive is simply ineffective and either the user finds other sources of motivation or the user stops being active. This explains that traditional market research companies usually have a very low "stickiness," users are attracted to the site, answer a few surveys, get some financial reward and get tired rapidly. The market research company has to keep on attracting new users to ensure a proper supply chain of panelists. In this respect we concur with Raban's insight [2] that a pure financial incentive is not enough.

4. Incentive and Intent

We distinguish between three types of intents or goals for incentives based on what the motivation of the site owner is; the three intents are: recruiting, improving content and ensuring participation in surveys. We describe them here.

- *Recruiting*: incentives are intended to attract new users to join the site; they are usually based on a simple message that should be enough to attract the curiosity of the "right" type of users. Conversion or bounce rates are good measures of the effectiveness of these incentives.
- *Improving Content:* incentives are intended to motivate existing users to generate content of good quality on the site. Stickiness, user fidelity, pageviews per visit, are good measures of effectiveness of these incentives.
- *Ensure participation in Commercial Survey*: incentives are intended to encourage a specific user to participate into a survey. Here market research common practice must be taken into account and very often the choice of incentives will be very limited. A good measure of effectiveness here is response rates to survey invitations.

The table below shows the results of several recruiting experiments we have conducted in the recent months. In the first table, one clearly sees the difference in effectiveness between a purely financial incentive and a non financial one. Financial incentives have 4 times the click through rate and 2.5 times the conversion rate than non-financial ones; on average they are thus ten times more effective.

Recruitment type	*Click-Through Rate*	*Conversion Rate*
Financial	0.2%	9%
Social	0.05%	4%

Similarly, we have conducted a series of experiments (some of which are still going on) on Toluna with all these kinds of rewards and intent and the compiled results are shown in the table below. We use a [1-5] range, where 5 indicates a best fit and 1 a worse fit.

Reward / Intent	*Recruiting*	*Improving content*	*Ensure participation*
Explicit: Financial	***5***	***2***	***5***
Implicit: Social	***1***	***5***	***2***
Explicit: Lotteries	***2***	***1***	***2***
Implicit: Experience	***2***	***2***	***1***

One key result of our experimentation is that financial rewards, which are the classical common practice for all purposes in market research sites are not appropriate for all intents and are not sufficient in most cases.

We verified that, when recruiting new users, financial reward is the leading solution. In addition, for increasing participation in most commercial surveys it is the only possible solution when for example:

- The topic of the survey cannot be revealed ahead of time (for various reasons) to the user, since interest is not a driver here, only financial rewards can be attractive
- The user can be screened out only after answering several preliminary questions (*e.g.*, because s/he does not belong to the specific demographic group targeted by the company ordering the survey), some type of explicit reward must be granted in order to avoid user's frustration and future participation in similar surveys.
- Surveys are obviously commercial; users might feel "cheated" if they do not see at this stage some type of revenue sharing.

In Figure 1, we clearly see that the users who are the most "socially" active also are the best survey takers. In our graph, we

show the response rate of a specific sample of users compared to their activity on the site. The activity of the site is measured by a simple number of "non financial" activities performed on the site. In Figure 2, we see a similar pattern, on countries in which we don't have a Toluna.com site translated, the users clearly show a steeper interest decline compared to similar countries on which Toluna.com is launched. In the figure, we compared the panels for Belgium (French and Dutch) compared to Spain and Italy.

5. Conclusion

.In this paper, we verified that the common practice of granting financial rewards in market research, is critical for several reasons, the most important one being to attract initial user's participation. A fair and reasonable reward system is essential and very important for good motivation and appeal to new users. However, we also discovered that financial reward alone cannot build a solid and rich user community. Like in other domains such as Q&A [3], it turned out that in market research as well, using a combination of explicit (financial and material) and implicit (social and user experience) incentives are key to a successful site.

6. ACKNOWLEDGMENTS

Our thanks go to Martin Lüttgersheiden and Laurent Sabouret for their help on the experimentation part. Without their insight we could not have done it.

7. REFERENCES

[1] Spreckelmeyer, Krach, Kohls, Rademacher, Irmak, Konrad, Kircher and Gründer, *Anticipation of monetary and social reward differently activates mesolimbic brain structures in men and women*, Social Cognitive and Affective Neuroscience Advance Access, January 27, 2009. *http://scan.oxfordjournals.org/cgi/content/abstract/nsn051*

[2] D. R. Raban, The Incentive Structure in an Online Information Market. Journal of the American Society for Information Science and Technology 59(14): 2284-2295, 2008.

[3] S. Rafaeli, D. R. Raban, et al., How Social Motivation Enhances Economic Activity And Incentives In The Google Answers Knowledge Sharing Market.International Journal of Knowledge and Learning 3(1): 1-11. 2007

[4] J. Tirole and R. Benabou, Intrinsic and extrinsic motivation. Review of Economic Studies, 70(3), 2003

[5] E. L. Deci, *Effects of externally mediated rewards on intrinsic motivation.* Journal of Personality and Social Psychology, 18(1), 1971.

[6] S. Baker, *Will Work for Praise*, Business week, Feb 2009. http://www.businessweek.com/magazine/content/09_07/b4119046650659.htm

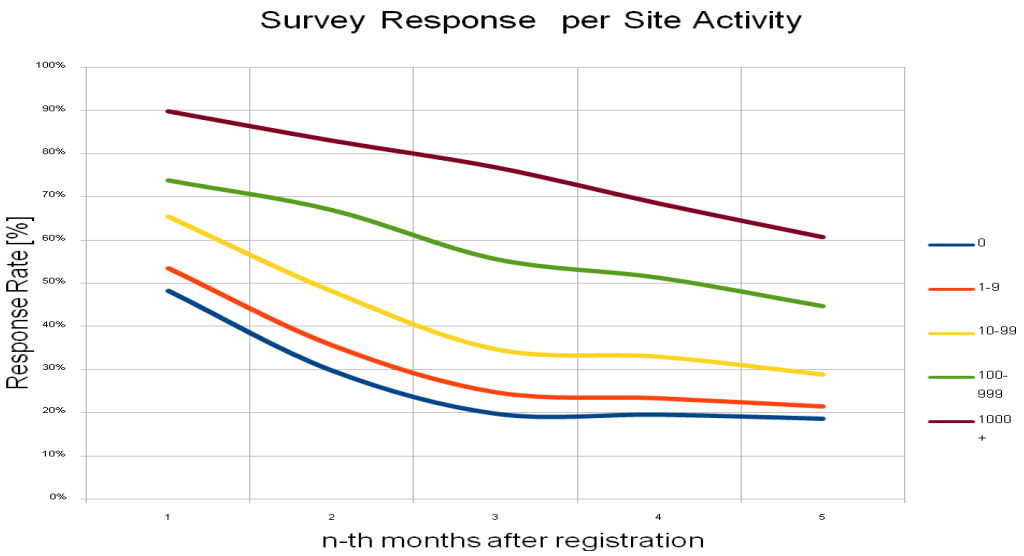


Figure 1: User Response compared to Site Activity

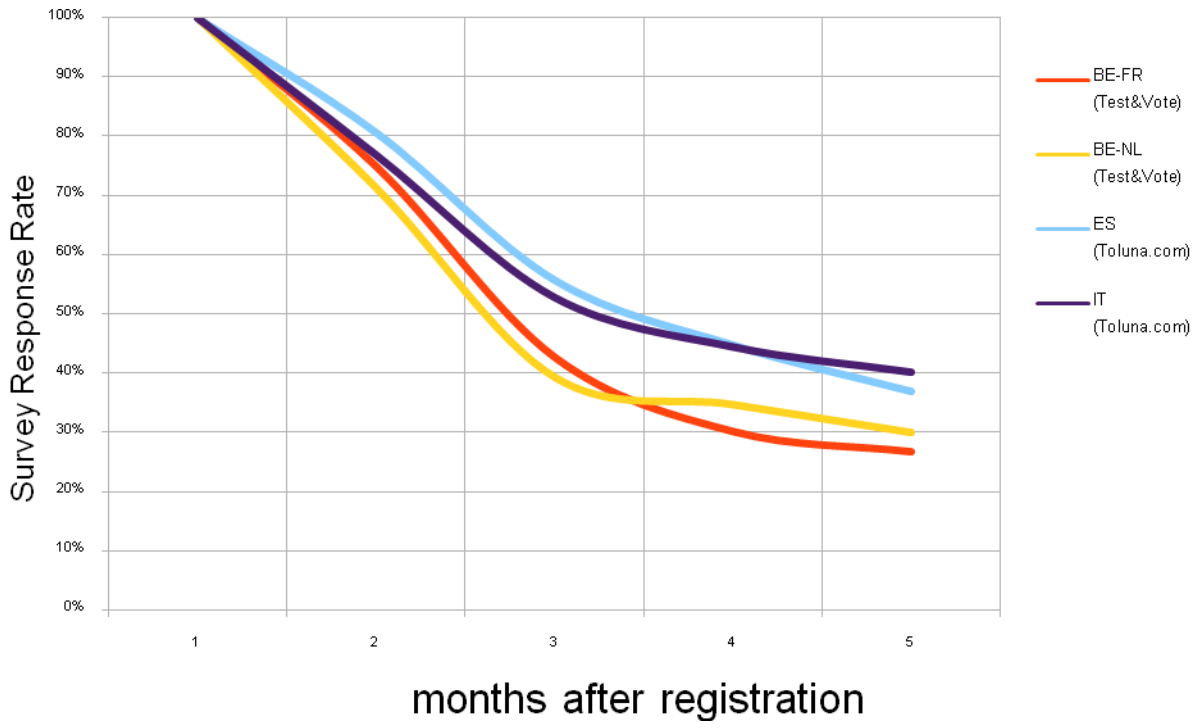


Figure 2: Comparing Response Rate for Several Panels

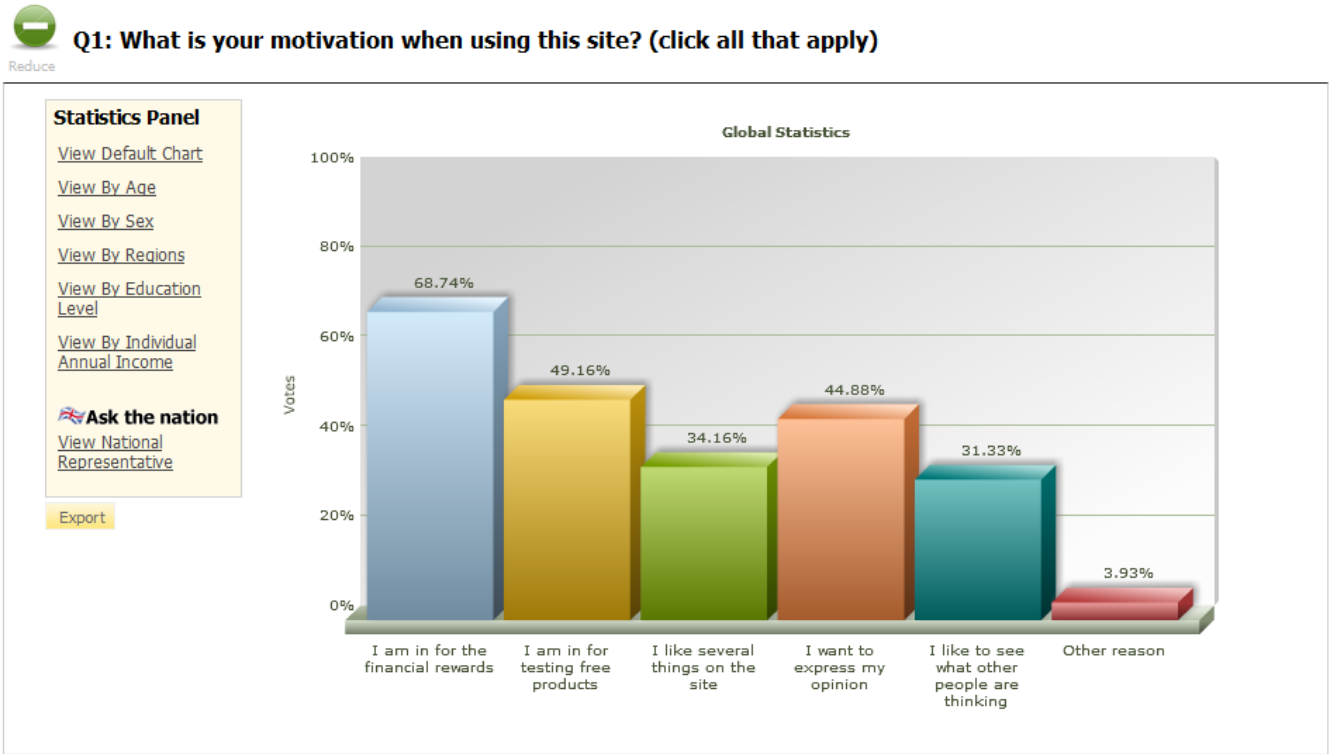


Figure 3: Motivation for new users

Reduce

Q2: What are the features you like most on Toluna? (click all that apply)

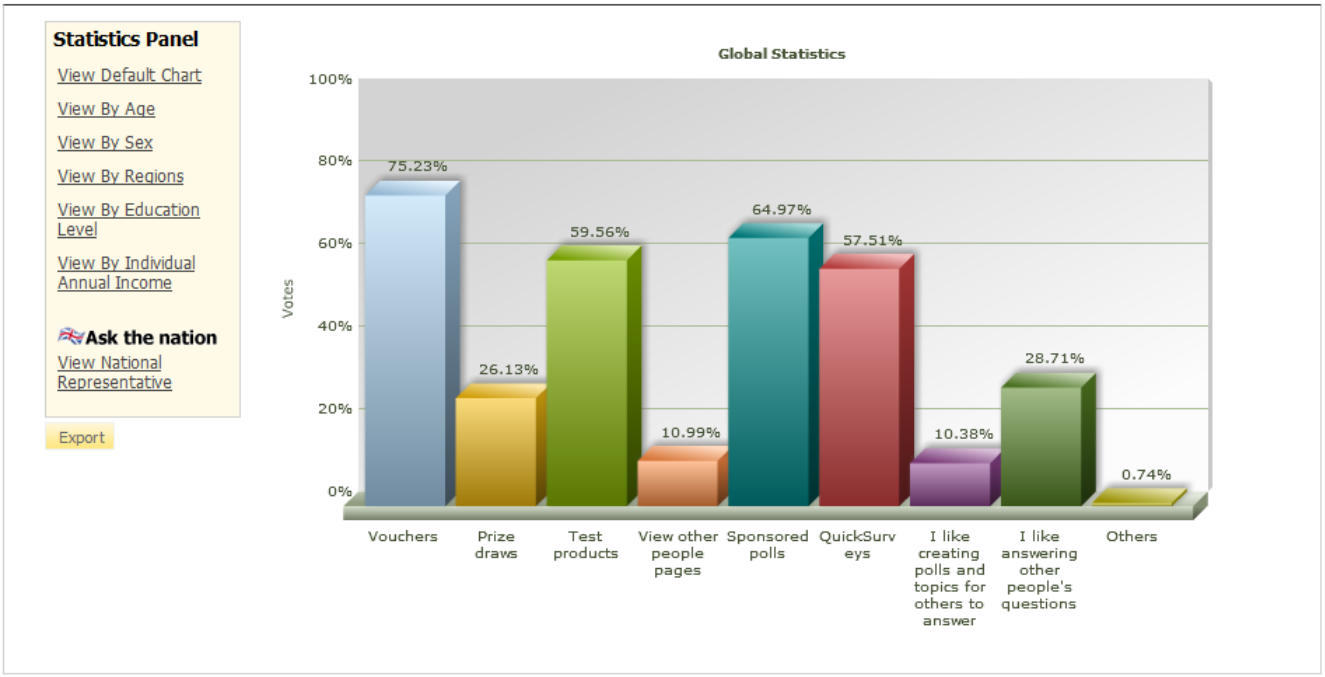


Figure 4: Interesting Features for Users Interested in Financial Rewards

Reduce

Q3: What are the features you like most on Toluna? (click all that apply)

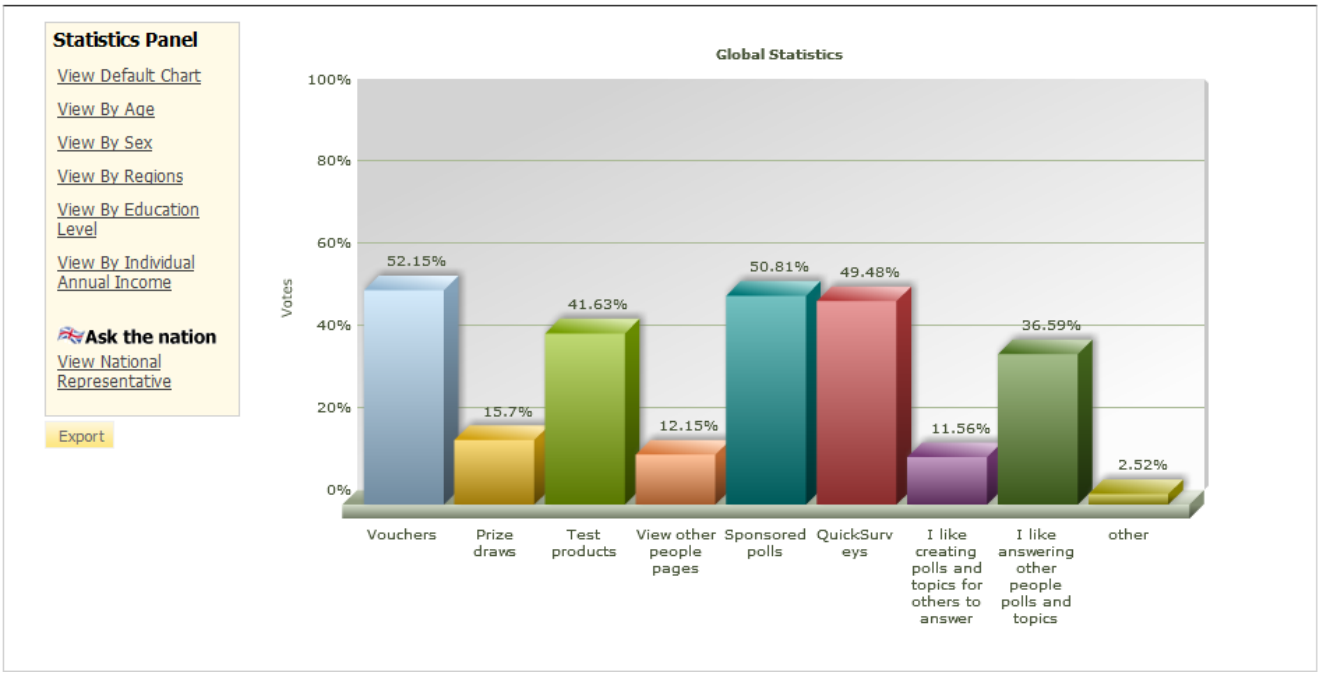


Figure 5: Interesting Features for Social Users

MultiRank: Reputation Ranking for Generic Semantic Social Networks

Xixi Luo
Computer Science Department
Beihang University
Beijing, China 100083
xixiluo.china@gmail.com

Joshua Shinavier
Tetherless World Constellation
Rensselaer Polytechnic Institute
Troy, New York 12180
shinaj@rpi.edu

ABSTRACT

This paper presents a technique for calculating "reputation" or influence of users and artifacts in semantic social networks: in particular, as an incentive mechanism to encourage reuse of complex resources such as ontologies. Adapting the PageRank algorithm to the relational schemas of typical social network applications, this technique allows the developer first to define via minimal rules the ways in which reputations of users and artifacts are likely to influence one another, then to obtain a mechanical, global ranking which reflects those rules in combination with the graph structure of the network. The mapping of multi-way relations such as usage and annotation to the binary-relational domain of PageRank is illustrated using the Actor-Concept-Instance model of ontologies. A lightweight software implementation,[1] currently under development, will provide a convenient way to add reputation-based functionality to Java-based community applications.

Keywords

Semantic Web, social network, reputation, incentive, PageRank, relational model

1. INTRODUCTION

Collaborative tagging systems have achieved tremendous popularity in the form of online media-sharing communities such as Delicious,[2] Flickr,[3] and CiteULike.[4] This is true in spite of the well-known shortcomings of tagging, including ambiguities of natural language such as variations in spelling, pluralization and part of speech [5]. Some of these shortcomings can certainly be addressed by Semantic Web technologies: for instance, by substituting controlled vocabularies for folksonomies. However, the obvious success of tagging systems indicates that their advantages outweigh their lack of clear semantics in many cases. On the other hand, the benefit of ontology-based annotation comes at the cost of significantly higher complexity. Emerging "Web 3.0" community applications such as Freebase[5] and various semantic wikis bridge this divide to some extent by providing a little more semantics than tagging systems, but a little more flexibility than typical ontology management tools. In such an environment, there is a need for effective incentive mechanisms to facilitate the complex task of building high-quality knowledge structures "from the bottom up". One such mechanism is the implicit "reputation" of ranking systems, which suggests to users the "best", most important, or most popular resources to use. The rest of this paper will focus on a specific ranking system, called MultiRank, which is based on an adaptation of the PageRank[3] algorithm to so-called semantic social networks.

1.1 The Actor-Concept-Instance model

As a minimal framework for semantic social networks, we will use a tripartite model of *actors* (human users or robots), *concepts* (tags, keywords, or possibly classes drawn from a controlled vocabulary) and *instances* (shared objects such as multimedia files, often contributed by actors themselves). All three elements are essential to the model: in a *semantic* social network, some notion of semantic annotation of instances with concepts is implied, whether this takes the form of simple folksonomy tagging or the sophisticated type system of a formal ontology. Furthermore, as the meaning of these artifacts is very much dependent on the context in which they are created and used [9], any measure of "reputation" should also take actors – authors, contributors – into account. This adds a social dimension to the otherwise bipartite model of traditional semantic networks (for example, of RDF graphs).

1.2 Multi-way relationships

We will use this tripartite model to illustrate the notion of *multi-way* relationships among actors, concepts and instances. Such a relationship may involve any number of elements, in contrast to the binary relationships of simple graphs. For example, the annotation of an instance – with a class – by an actor involves three distinct elements, and so cannot be completely represented by a simple binary edge. Multi-way relationships are common in database applications but complicate the otherwise simple recursive defini-

[1] http://multirank.googlecode.com/
[2] http://delicious.com/
[3] http://flickr.com/
[4] http://citeulike.org/
[5] http://www.freebase.com

tion of PageRank,[6] and we claim that there is no single *right* way to map them into the binary-relational domain. Instead, we provide a generic framework with which an application developer can define such mappings, in terms of patterns of ranking propagation between pairs of elements in a relation.

2. REPUTATION FROM RELATIONSHIPS

Intuitively, *reputation* is a collective measure of trustworthiness in the estimation of the community [7]. A user's reputation has both prescriptive and descriptive value: it is *prescriptive* in that it defines "good behavior" on the part of the user and thereby specifies the way in which users can gain reputation, and *descriptive* in that it provides a way to rank and classify users on the basis of their reputations [1]. The reputation of *resources* identifies the "best" or most important resources and thereby singles them out as candidates for imitation or reuse. For the purpose of this paper, reputation is an *implicit* statement of trustworthiness: much like the original formulation of PageRank, MultiRank is an attempt to measure human interest and attention based on the network of relationships within the community. Such an approach holds the possibility of making minimal demands on the user, while scaling well and delivering subjectively accurate results despite a high degree of heterogeneity in the quality and structure of the network.

2.1 Propagation of ranking

The notion of propagation of ranking (here: of reputation) through directed edges is the basis of the PageRank algorithm: if the sum of the ranking of the nodes with edges to a given node is high, then the ranking of the node itself should be high. The contribution of MultiRank is in the construction of a "virtual" binary-relational graph G_{prop}, on which to run PageRank in order to derive reputation values. The nodes of this graph are the the actors, concepts and instances of the semantic social network, while its edges are chosen so as to propagate ranking from node to node in a way that reflects the intended or expected flow of reputation within the network. In general, the reputation of an actor or artifact tends to increase the reputation of another item with which it associates (for instance, by "creating", "using", "knowing", or otherwise drawing attention to that item).

In the following, we list several informal and intuitive "rules" for the flow of reputation ranking in a hypothetical semantic social network (see also Figure 1), with which we motivate the idea of *propagation patterns* defined in the next section:

1. If actor a_1 knows actor a_2, then a_1's reputation should propagate to a_2. This rule reflects the fact that an actor benefits from being known by other, high-reputation actors. Note that we've chosen to let reputation propagate in only one direction. In this respect, our social network resembles an environment like Twitter,[7] in which it is possible to "follow" a popular individual without that individual necessarily following you.

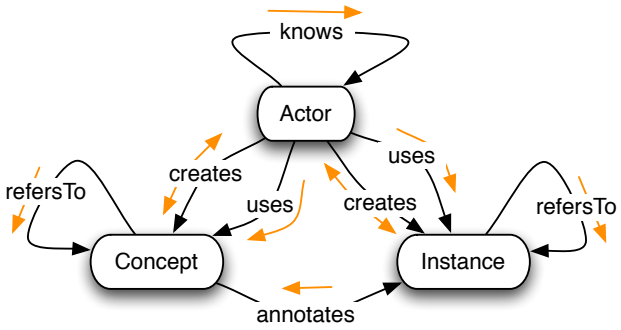
2. If actor *a creates* artifact t, then a's reputation should propagate to t, and vice versa. This reflects the fact that artifacts created by an actor with a high reputation are to some extent *authoritative*: their reputations benefit from association with their creator. Conversely, an actor's reputation benefits (perhaps to an even greater extent) from association with artifacts she has created which have achieved a high reputation.

3. If actor *a uses* artifact (concept or instance) t, then a's reputation should propagate to t. This is a somewhat weaker version of rule 2: an artifact should gain some reputation through a high-reputation actor who has associated himself with it. However, the reverse is not true: merely associating oneself with a great resource does not make one great.

4. If concept c *annotates* instance i, then reputation should propagate "backwards" from i to c, the direction of *annotates* being unimportant. Although each of these rules is debatable, we imagine the annotated instance as drawing attention – and thus reputation – to the annotating concept, but not the reverse.

5. If artifact t_1 *refers to* artifact t_2, then t_1's reputation should propagate to t_2. Again, a high-reputation item should increase the reputation of other items with which it associates. If the artifacts happen to be web pages and the links happen to be hyperlinks, then this rule is particularly close to ordinary PageRank.

Figure 1: Flow of reputation in an example network



2.2 Mapping to a binary-relational network

Now that we have an intuitive idea of propagation of ranking in the virtual network, let us formally describe the derivation of that network from a collection of multi-way relationships. For the purpose of clarity, we will introduce the notions of terms, variables, and bindings, in analogy to the SPARQL[11] query language and to relational databases. In the following, a *term* $t \in T$ is any item in the social network (be it an actor, concept or instance), a *variable* $v \in V$ is an abstraction which may be replaced with a term, and a *binding* $b \in V \times T$ is a pair which connects a variable to a term. Furthermore, a *relation* is an abstract relationship among variables which carries a particular meaning. For the purpose of calculating reputation, we reduce that meaning to a

[6] The PageRank of a node p in an unweighted, directed graph is generally given by $PR(p) = \frac{1-d}{N} + \sum_{q \in B(p)} \frac{PR(q)}{N_q}$, where N is the total number of nodes, $B(p)$ is the set of nodes q with edges to p, N_q is the out-degree of q and d is a *decay factor* which helps to dampen the effect of graph cycles. Weighted PageRank is defined similarly. Note that the expression $\frac{PR(q)}{N_q}$ represents the ranking of the edge (p, q).

[7] http://twitter.com/

Table 1: Propagation patterns of example relations

Relation	Variables	Propagation	Rules
actor knows actor	$\{a_1, a_2\}$	$(a_1, a_2, 0.6)$	1
actor creates concept	$\{a, c\}$	$(a, c, 0.4)$ $(c, a, 1.0)$	2
actor creates instance	$\{a, i\}$	$(a, i, 0.4)$ $(i, a, 1.0)$	2
actor uses concept to annotate instance	$\{a, c, i\}$	$(a, c, 0.2)$ $(a, i, 0.2)$ $(i, c, 0.8)$	3, 4
concept refers to concept	$\{c_1, c_2\}$	$(c_1, c_2, 0.6)$	5
instance refers to instance	$\{i_1, i_2\}$	$(i_1, i_2, 0.6)$	5

pattern of pairwise propagation of ranking among variables. For example, the *actor uses concept to annotate instance* relation in Table 1 involves three variables and propagates ranking among them according to rules 3 and 4 from the preceding section. Specifically, it propagates small amounts of ranking from the actor to the concept and to the instance (which the actor *uses*) and a larger amount of ranking from the instance to the concept (which *annotates* the instance). In general, we define a relation $R \subseteq V \times V \times \mathbb{R}$ as a set of abstract weighted edges between pairs of variables. Such an edge represents a type of path along which ranking is permitted to flow, while its weight permits fine-grained control over the *extent* of flow. This set of abstract edges is a *pattern* of ranking propagation which is to be pre-defined by the developer for each relation.

By combining a relation R with a set B of bindings, we obtain a multiset of weighted edges which propagate ranking among specific terms in T:

$$\begin{aligned} PropEdges(R, B) = \{(t_1, t_2, w) \quad : \quad & (\exists v_1, v_2 \in V, w \in \mathbb{R}) \\ & ((v_1, t_1) \in B) \\ & ((v_2, t_2) \in B) \\ & ((v_1, v_2, w) \in R)\} \end{aligned}$$

That is, the set of propagation edges for R and B are those which can be formed by replacing variables in the abstract edges of R according to the variable-to-term pairs in B. This combination of an abstract relationship with specific terms has a counterpart in SPARQL queries, in which variables in a query are resolved to specific terms in its solution, and in relational databases, in which column names resolve to specific cells in rows of data. It is easy to imagine a process which iterates through all of the results of a query or all of the rows in a table, applying $PropEdges$ to each result or row to generate a graph of all possible propagation edges:

$$G_{prop} = (T, \biguplus_{(R,B) \in S} PropEdges(R, B))$$

Here, S represents the set of all relation-binding pairs, or concrete semantic relationships which make up the social network. Depending on the application, these pairs may be drawn from a SPARQL query, a SQL database, or any other source of tabular, relational data. G_{prop}, then, is the weighted, directed multigraph of propagation edges generated by means of S over the set T of terms. This is the graph on which we will actually run PageRank to derive ranking results.

2.3 Applying PageRank

The propagation graph G_{prop} is an intermediate result in our calculation of ranking. To derive a final result, we have only to apply a weighted form of PageRank to this graph. The fact that G_{prop} is a *multigraph* presents no additional challenges: to transform it into a ordinary weighted graph G'_{prop}, we simply merge parallel edges, adding their weights together. Formally, we compute PageRank by iteratively solving for the vector $\pi \in \mathbb{R}^{|T|}$ in:

$$\pi = (1 - d)E + dG'_{prop}\pi$$

where d is the decay factor (typically chosen to be 0.85) and $E \in \mathbb{R}^{|T|}$ is a vector representing a source of ranking. If E is uniform over all $t \in T$, then the resulting π is a *global* measure of reputation in our semantic social network. However, by biasing E in favor of particular terms, any number of so-called *personalized*[10] PageRanks can be applied. Computing a personalized PageRank ranking (biased, for example, towards actors, concepts and instances which are *trusted* by a particular user) over G_{prop} brings MultiRank closer in spirit to trust-based mechanisms in recommendation systems [2] and shared content repositories [8].

3. INCENTIVE FROM REPUTATION

Our technique has been thoroughly described in the sections above. Having once computed the "reputation" vector π, the application is free to use it in application-specific ways. For instance, an ordering of actors by decreasing reputation can be used as a "Top X" list to which actors may aspire, raising the quality of the social network in the process. Similarly, actors may strive to get their own artifacts into "best of" lists of various kinds. These rankings, in turn, may help to ensure that the top actors in the network get the attention they deserve, and that the top artifacts, such as the elements of well-designed ontologies, are consistently re-discovered and re-used. Although in this paper we have focused on the ranking technique itself, we believe that the prescriptive value of subjectively accurate ranking results is an ample foundation for incentive mechanisms to motivate users to improve their own resources and connections.

4. RELATED WORK

There have been a number of Semantic Web tools which make use of PageRank. For example, the Swoogle search engine's OntoRank[4] is a variation of PageRank for ontologies. OntoRank takes a number of types of semantic links into account when calculating ranking, weighting links selectively according to these types. Similarly, the Semantic Web Search Engine's ReConRank[6] extends a graph of RDF resources with contextual edges, forming a compound graph which includes relevant provenance information. However, both of these technologies operate upon existing binary-relational semantic networks, whereas MultiRank is designed for relations which are not necessarily binary, introducing the notion of propagation patterns to first construct a "virtual" binary-relational network before applying PageRank to it. This technique was motivated by the notion of semantic-social hypergraphs in Peter Mika's tripartite ontology model, and builds upon previous work [12] in applying single-relational network analysis algorithms to multi-relational networks.

5. CONCLUSION AND FUTURE WORK

We have presented a PageRank-based reputation ranking system for semantic social networks, illustrating it with an actor-concept-instance model. The technique itself is very general: we make only the basic assumption that the structure of the network can be represented as a collection of multi-way relationships, such as the solution to one or more SPARQL queries or the contents of one or more tables in a relational database. MultiRank borrows from PageRank the implicit *reputation* of resources as expressed in network structure alone, while it adds application-specific *propagation patterns* which direct the flow of reputation according to a human's intuitive understanding of the social network. The algorithm proceeds in two stages: a loading stage in which relational data is processed row by row to derive a virtual binary-relational graph, and a computational stage in which the PageRank algorithm is applied to the virtual graph to generate ranking results. Due to the simplicity of the model and favorable performance characteristics of PageRank, we believe that a software implementation of MultiRank will provide a cheap and effective way to add reputation-based functionality to any of a variety of semantic social networks. Such an implementation is currently under development, building on the open-source Java Universal Network/Graph Framework (JUNG).[8] We intend to test this software in the near future using more than one social network data set, including a large dump of Freebase event logs. This will help us to estimate performance and memory usage, as well as to gauge the sensitivity of the computed ranking results with respect to the *weight* values of propagation patterns in different application scenarios. At that point, we will be ready to deploy and evaluate MultiRank-based incentive mechanisms in a live semantic social network environment.

6. ACKNOWLEDGEMENTS

This work has been supported by Rensselaer's Tetherless World Constellation. We would also like to thank Jim Hendler, Deborah McGuinness, Marko A. Rodriguez and Li Ding for their valuable feedback on various drafts of this paper.

7. REFERENCES

[1] Thomas B. Adler and Luca de Alfaro, *A content-driven reputation system for the Wikipedia*, WWW '07: Proceedings of the 16th international conference on World Wide Web, ACM Press, 2007, pp. 261–270.

[2] Reid Andersen, Christian Borgs, Jennifer Chayes, Uriel Feige, Abraham Flaxman, Adam Kalai, Vahab Mirrokni, and Moshe Tennenholtz, *Trust-based recommendation systems: an axiomatic approach*, WWW '08: Proceeding of the 17th international conference on World Wide Web, ACM, 2008, pp. 199–208.

[3] Sergey Brin and Lawrence Page, *The anatomy of a large-scale hypertextual web search engine*, Computer Networks and ISDN Systems **30** (1998), no. 1–7, 107–117.

[4] Li Ding, Tim Finin, Anupam Joshi, Rong Pan, Scott R. Cost, Yun Peng, Pavan Reddivari, Vishal Doshi, and Joel Sachs, *Swoogle: a search and metadata engine for the Semantic Web*, CIKM '04: Proceedings of the thirteenth ACM conference on Information and knowledge management, ACM Press, 2004, pp. 652–659.

[5] S. Golder and B. Huberman, *The structure of collaborative tagging systems*, Journal of Information Science **32** (2006), no. 2, 198–208.

[6] Aidan Hogan, Andreas Harth, and Stefan Decker, *ReConRank: A scalable ranking method for Semantic Web data with context*, In 2nd Workshop on Scalable Semantic Web Knowledge Base Systems, 2006.

[7] Audun Josang, Roslan Ismail, and Colin Boyd, *A survey of trust and reputation systems for online service provision*, Decision Support Systems **43** (2007), no. 2, 618–644.

[8] Deborah L. Mcguinness, Honglei Zeng, Paulo P. da Silva, Li Ding, Dhyanesh Narayanan, and Mayukh Bhaowal, *Investigations into trust for collaborative information repositories: A Wikipedia case study*, MTW (Tim Finin, Lalana Kagal, Daniel Olmedilla, Tim Finin, Lalana Kagal, and Daniel Olmedilla, eds.), CEUR Workshop Proceedings, vol. 190, CEUR-WS.org, 2006.

[9] Peter Mika, *Ontologies Are Us: A unified model of social networks and semantics*, International Semantic Web Conference, LNCS, Springer, 2005, pp. 522–536.

[10] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd, *The PageRank citation ranking: Bringing order to the Web*, Tech. report, Stanford Digital Library Technologies Project, 1998.

[11] Eric Prud'hommeaux and Andy Seaborne, *SPARQL query language for RDF*, Tech. report, W3C, January 2008.

[12] Marko A. Rodriguez and Joshua Shinavier, *Exposing multi-relational networks to single-relational network analysis algorithms*, Tech. Report LA-UR-08-03931, Los Alamos National Laboratory, 2008.

[8] http://jung.sourceforge.net/