An analysis of Social News Websites

Diego M. Virasoro
University of Southampton
University Road
Southampton, UK
dmv1c09@soton.ac.uk

Pauline Leonard
University of Southampton
University Road
Southampton, UK
pleonard@soton.ac.uk

Mark Weal
University of Southampton
University Road
Southampton, UK
mjw@ecs.soton.ac.uk

ABSTRACT

Social News Websites (SNW), such as `Digg` and `Reddit`, are a new type of on-line news aggregator where an article's visibility (e.g. its position) is determined "democratically" by the users' preferences (usually by vote). We offer an overview of SNWs in both generic and specific terms, as well as some common venues for variation. Finally we present two main results observed by continuously monitoring a `Reddit` and two more SNWs across a few days. First we observed a strongly cyclical variation in the site's activity and find that this affects the proportion of articles submitted at different time of the day/week that reaches the front page. Second we consider the problem of participation inequality and study the distribution of authorship, finding it consistent with Zipf's law.

Keywords

Collaborative filtering, Recommendation systems, Social media, Collaborative gatewatching, Collective intelligence, partecipation inequality, Reddit, Zipf's law

1. INTRODUCTION

The Internet (and in particular the Web) has brought a revolution that has all but removed distribution costs, disrupting many of the old equilibria and paradigms. In particular, and in contrast with previous technologies, it involves for authors and publishers very low initial costs coupled with a (potentially) global reach. This has resulted in an explosion of readily available information. However at this massive scale discovery and organisation are still a challenge: users are lost in a disorientating flood of information with few pointers to locate what they are after and judge its veracity and accuracy.

After human-generated online directories and completely automatised search engines, a new paradigm has recently emerged: *collaborative filtering*, a system that tries to condense a vastitude of subjective human judgements into one global (and hopefully "reliable") opinion via intelligent algorithms. While many applications of collaborative filtering have arisen (e.g. recommendation systems), there are additional challenges when applied to news, due to their time-sensitive nature: not only is the set of current news in permanent flux, but so is their perceived newsworthiness.

Social News websites (SNW), an application of collaborative filtering to news, are online aggregators (viz. a list of links to online articles) where the rankings are not chronological but instead are determined "democratically" by their members' preferences (usually by vote).

After `Digg`'s early success, which popularised the SNW paradigm, many such sites have sprung up mostly specialising in niche topics. There is still, however, a poor understanding of the sites' dynamics and especially how these are affected by the algorithms and the communities. Indeed there are no accepted algorithms (or best practices): different sites re-implement similar ideas in very different ways, without a clear view of the implications.

2. SOCIAL NEWS WEBSITES

Social News website (SNW) can be considered a new generation of online news aggregators that, following the trends of *Web 2.0*, harnesses the "collective intelligence" of its members to sort its content. In that sense it can be seen as substituting the *centralised gatekeeping* role of the newsroom editor [2], with a decentralised collaborative system that allows all news to be published contending itself with highlighting the most interesting articles: *collaborative gatewatching* [1].

The essence of the social news paradigm is simple and can be summarised in three properties:

News-aggregation A SNW is a type of news-aggregator, i.e. it links to newsworthy online articles but it rarely offers original content itself;

User submissions While generally the articles are submitted by the users themselves, there are cases where this is complemented or substituted by editors' submissions, automatic submissions, and more recently by paid advertisement-articles too. That said, we only considered sites that allow user-generated content.

User moderation This is the main innovation that sets SNW apart from other open news-aggregators. Rather

than displaying the news in pure reversed chronological order, the article's position, and therefore its visibility, is determined by a mixture of reversed chronological order and users' preferences (and/or activity). The result is a *positive feedback* as the "best" articles (as judged by the users) are rewarded by higher votes, which grants them higher visibility and hence a better chance to be voted by more users.

There are various ways in which social news websites may differ, with many alternative already proposed in other types of collaborative filtering (for a good overview see [6]). In particular:

Vote choices Because the user can always choose not to vote (the default) the simplest form of vote is a binary choice: `good`/default, or {+1,0}. This is the most popular set for SNWs, possibly due to its simplicity. Others introduce a third choice for a negative opinion (with respect to the default): e.g. `good`/default/`bad` or {+1,0,-1}. While not very popular, these sets could be further extended by adding more options such as the 5-star system used in Amazon, or changing the default value.

User weight Most websites are based on a strictly egalitarian principle where all votes have equal weight. Others, such as Linkibol use a meritocratic system: each user is assigned a karma value that grows with participation, and declines with anti-social behaviour and this value determines the weight of user's vote. Other possibilities could involve, for example, rewarding seniority (how long ago the user registered) or vitality (how active is the user).

Score computation This is the function that maps the tuple of all the user's votes (and their weights) to a real number, the score. By far the most common function is a straightforward sum of all votes, or — in some cases — an average. However depending on the site's design other alternatives could be preferable, such as a sum of cubes (to reward extreme values) or a truncated mean (to make it more robust). Moreover the algorithm could compensate for users voting in block by computing the correlations among the users choices, so that a moderately popular news across the whole community would gain a higher score than a very popular news across a small groups of like-minded individuals.

Visibility/Placement Finally the site needs to map the score of each article to a ranking. The simplest case is to rank each article to reward higher scoring and/or recent news. Alternative ranking could be based on the activity, such as the rate of score change (i.e. the score's first time derivative), or include factors such as the number of comments or the time of the day. Many sites offer alternative lists with different sorting rules: for example a slowly changing "best" list and a fast changing "newest" list; or may use hierarchical lists (e.g. see Digg below). Finally some sites may present personalised rankings for each user.

Manual moderation Some sites will have additional forms of moderation. This can be centralised (e.g. a team of moderators), decentralised (e.g. users can vote if an article or comment is inappropriate or inflammatory) or a mixtures of the two.

Social layer Most SNWs offer a space for members to discuss or comment on articles. A site may also offer different forms of meta-moderation (i.e. moderating the users actions) to penalise vandalism, spamming, and trolling. Finally many SNW offer the possibility of "befriend" (i.e. follow) other users. This may result, for example, in an increased weight for the friends' votes, a separate list only populated by friend's submissions, etcetera.

3. A PANORAMA OF SNW

The idea of a personalised newspaper that would aggregate the most relevant online news dates as far back as 1992 when an MIT student's project resulted in `FishWrap`. Similar ideas remerged in `CRaYON` at Buckell University, and `PointCast`. However it wasn't until Digg launched in December 2004 that the modern Social News Website arose, and its success ensured the wide acceptance of the paradigm.

Since then many such sites have sprung up, though with the exception of a few (such as `NewsVine` and `Reddit`) most failed to reach a sufficiently large community to produce an accurate algorithm and prosper. A few though have faired better by focusing on particular fields around which they gather a keen community. In this respect the most popular topic is undoubtedly technology (the original focus of `Digg` and `Reddit`) followed by design (e.g. `Design Float`, `DesignBump`), ecology (e.g. `Care2`), curiosities (e.g. `I am Bored`, `linkfilter.net`), celebrity gossip (e.g. Lipstick) and sports/cars (e.g. `Auto Spies`, `BallHype`). Others have prospered by focusing not on a field but on a culture (e.g. IndianPad and Muti) or non-english language (e.g. `Menéame` and `Linkibol`).

As mentioned above most sites differ profoundly in the details of their mechanics. Below we bring three examples:

Hacker News It was started in February 2007, and focuses on technological news, start-ups, and hacker's culture. Compare to other SNWs, the site is rather simple but still effective (thanks in part to its loyal and well respected community). Users can only submit and upvote (+1 vote) articles. All articles are presented in one main list (divided into 30 items per page) sorted via the ranking function[1]:

$$f_i = \frac{(v_i - 1)^{0.8}}{(t_i + 2)^k},$$

where v_i is the vote for the i-th article, t_i the number of hours from creation, and k is a decaying constant set to 1.8. Additionally there is a "newest" list which is strictly in reversed chronological order ($f_i = -t_i$), and a "best" list which ranks the last 1000 submissions only by the score ($f_i = v_i$). A comment thread is created for each article, which are themselves moderated by a similar voting systems.

[1]The code is available at:
http://github.com/nex3/arc/blob/master/news.arc

Reddit Created in April 2005, it has recently (2010) become the most popular SNW. It offers a vast number of features, such as allowing both up-votes (+1 vote) and down-vote (-1 vote). It is a federated system made of hundreds of sub-communities covering specific niches: in addition to the original ones, any member can create new sub-communities and configure its specific site's parameters as necessary (and moderate it if needed). Each user can select which sub-communities to follow: this will determine what articles will appear in the front page lists. There are four such lists, each sorted according to a different ranking function and displaying 25 items per page. The default list is the "hot" list sorted by[2]

$$f_i = \log_{10} v_i + \frac{v_i}{|v_i|}\frac{t_i}{k},$$

where the v_i is the vote for the i-th article, t_i the number of hours between creation and an arbitrary starting time, and k is a constant set to 12.5. The remaining three are: the "new" list (sorted in reversed chronological order), the "top" list (for the latest day/.../year, sorted by the vote), and the "controversial" list (sorted by the absolute value of the sum of the number of up and down votes divided by their difference). In addition users can choose to follow other users and all the articles submitted by these users will appear in a fifth list, the "friends" list.

Digg This was the first successful SNW, launched on December 2004. It remained the most popular SNW until 2010 when many users emigrated to its main competitor (`Reddit`), following the launch of a new version that took the power away from the community. Although `Digg` started with a focus on technology, it has since open up to all news. It features two main lists: the "top news" (the default), and the "upcoming" list. The latter will collect any new articles which can then be up-voted by the users. Articles that grew their score quickly are promoted (i.e. moved) to the "top news" list (the front page) where they are displayed in order of promotion time (from the newest to the oldest). While in the "top news" list, articles can still be upvoted which will may lead it to appear among the best articles of the last day/week/month. Finally there is a "hot news" list that current fast growing articles, and a "friends" list that displays articles submitted or upvoted by other users that one decides to follow. Unfortunately, and in contrast with `Reddit` and `Hacker News`, `Digg`'s code is strictly close-sourced, so the details of the algorithm are not known.

4. RESULTS

We monitored `Reddit` (non-member default front page) for 60 hours[3], by taking regular snapshots — at five minutes intervals — of the first 5 pages of the main list and the first 10 pages of the "new" list. A few results are shown below (similar results were obtained for `Hacker News` and `Tip'd` too.)

[2]The code is available at: https://github.com/reddit/reddit

[3]Between 28/Aug/2010 14:10:30 BST and 31/Aug/2010 04:09:12 BST

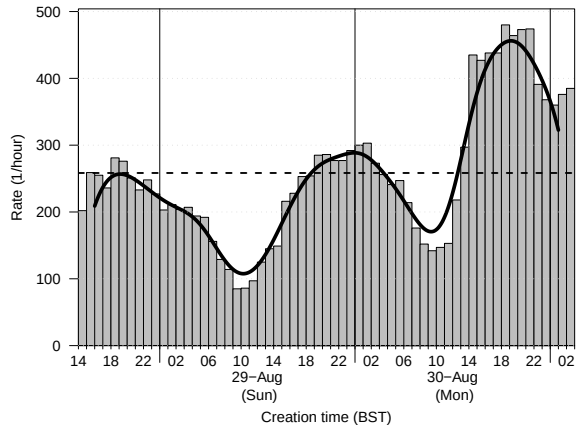

Figure 1: A histogram of the creation rate of articles on `Reddit` shows a strong cyclical component. Similar results were found when focusing only on the front page list, and is other SNWs such as `Hacker News` and `Tip'd`.

A characteristic that we noticed in all the sites studied is a strong cyclical variation in the site's activity with a 24 hour period, as shown in figure 3 for `Reddit`. This is not surprising since human activity follows the same pattern but it underlines the fact that the users are not distributed uniformly across the globe. A more careful analysis across several websites found the cycle consistent with peaks during working hours for the largest geographical community.

These fluctuations may causes more serious repercussions as they may undermines the accuracy of the site's algorithm. For example many sites have a very simple promotion algorithm, viz. promote any article whose score reaches a predetermined and constant threshold. Digg itself appeared to use just this algorithms in its earliest years [4]. Then, we would expect the promotion itself to be affected by the variability in activity, meaning that articles submission time would influence its likelihood of promotion!

Indeed a similar problem occurs on `Reddit` too, despite its different underlying algorithm. This can be seen more clearly in figure 4 where we can see that the cyclical variation in the site's activity results in an uneven probability of reaching the main list.

Another interesting problem to consider is the *participation inequality*: the extent of this inequality would have significant implications for the plurality of views presents in these communities, which would be crucial for effective gatewatching of sensitive news. This can be studied by analysing the authorship distributions of the number of submissions, viz. the percentage of articles submitted by the same person (a.k.a. the 'poster').

Previous studies had shown that activity in online communities follows Zipf's law [5]: in other words if the users are ranked in order of number of submissions, the proportion of articles submitted by the the r-th ranked user would be $f(r) \sim r^{-\omega}$. This was confirmed by our own results for the

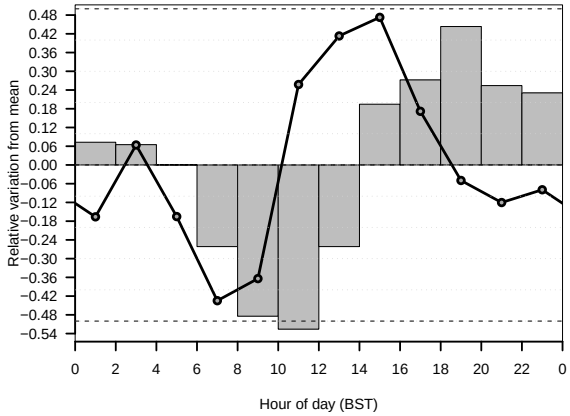

Figure 2: The relative variations from the respective means of the rate of article submission (bar-plot) and the proportion of these articles that reached at least once the top-125 of the front page (line-plot).

SNWs monitored, as shown in figure 4 for `Reddit` where we computed the best estimate of ω as explained in [3].

Further we tested if other distributions could explain these results better by computing the log-likelihood ratios. For `Reddit` we could rule out to 95% confidence the following distributions: exponential, Yule, Poisson and Weibull distributions as well as the log-normal distribution for the "new" list (but we could only reach a statistically significant conclusion for the default list.)

5. ACKNOWLEDGMENTS

This work was supported by an EPSRC Doctoral Training Centre grant (EP/G03690X/1).

6. REFERENCES

[1] A. Bruns. Gatewatching, not gatekeeping: Collaborative online news. *Media International Australia*, 107:31–44, 2003.

[2] A. Bruns. *Gatewatching: Collaborative online news production*. P. Lang, 2005.

[3] A. Clauset, C. Shalizi, and M. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[4] K. Lerman. Social information processing in news aggregation. *IEEE Internet Computing*, pages 16–28, 2007.

[5] J. Nielsen. Participation inequality: Encouraging more users to contribute. *Jakob NielsenâĂŹs Alertbox*, 9:2006, 2006.

[6] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen. Collaborative filtering recommender systems. In *The Adaptive Web*, pages 291–324. 2007.

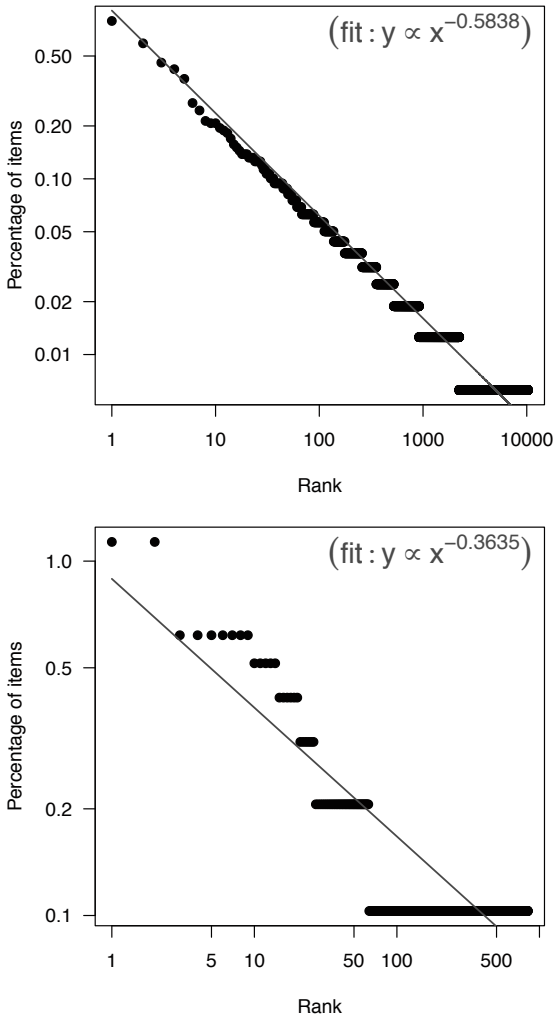

Figure 3: Studying the authorship concentration on `Reddit` for the the "new" list (top) and the default list (bottom) shows a good fit to Zipf's law (solid line). Of the two the concentration on the "new" list is the highest.