Evaluation of Valuable User Generated Content on Social News Web Sites

Yana Volkovich*
yana.volkovich@barcelonamedia.org

Andreas Kaltenbrunner
andreas.kaltenbrunner@barcelonamedia.org

Barcelona Media - Centre d'Innovació
Av. Diagonal, 177, planta 9
Barcelona, Spain

ABSTRACT

Social news websites have gained significant popularity over the last few years. The participants of such websites are not only allowed to share news links but also to annotate, to evaluate and to comment them. To quantify interestingness and attractiveness of the user generated content in respect to the original link source we introduce the User Generated Content add-on (UGC_+) index. Based on the definition of UGC_+ we also propose a concept for comparing groups of links filtered by different properties, e.g. authorship or topic-categories. We apply the proposed measure on the Spanish Digg-clone Menéame.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Measurement

Keywords

User Generated Content, News Aggregator, Social Media

1. INTRODUCTION

Nowadays a large part of social communications has moved to social media. In contrast to one-to-many communication structure of traditional mass media, social media allows the emergence of many-to-many communication, and gives a rise to mass self-communication [1]. Such self-communication generates new information and knowledge on the base of the original content. In particular, social participants of news websites may provide new angles and add dimensions to the original news content by commenting and annotating it.

In this work we study whether user generated content (UGC) adds any value to the original content shared on some social news website. To this end we define a measure of attractiveness of the user generated content in comparison to the original one. To the best of our knowledge such a study has not been undertaken before. We focus our work on social news websites, in particular on the Spanish link-sharing website Menéame. However, the proposed concept can easily be extended to any other link sharing website.

2. EVALUATION OF UGC

The purpose of a social news sharing platform is to allow users to publish annotated links (stories) to the news they consider relevant, to vote and to comment on the stories published by other users. The idea of our *User Generated Content add-on index* (UGC_+) is based on the assumption that a person who cites some news story will rather refer to the news sharing website than to the original content if the corresponding story on the news sharing website contains any new valuable information (e.g. an interesting discussion within the comments of the post).

2.1 UGC add-on index

Let $l(i)$ be a link to an original page i published on some social news website. The difference between the number of citations of the original and referring stories can be used to define the *UGC add-on index* as:

$$UGC_+(l(i)) = \deg(l(i))\ \mathrm{sgn}[\deg(l(i)) - \alpha\ \deg(i)],$$

where $\deg(i)$ is the in-degree of a web page i, and $\alpha \in (0, \infty)$ is a parameter. The UGC_+ can take positive and negative values. The positive indexes indicate the preference of the annotated link over the original content. The negative numbers with the large absolute values also refer to meaningful social news posts, however, these posts are less interesting than the original links. An UGC_+ index close to zero reveals no significant interest of the shared news story for the readers.

The meaning of the parameter α is illustrated in Figure 1(left). The selected area, bounded by the line $y = \alpha x$, corresponds to the positive values of $UGC_+(\alpha)$. By adjusting the parameter we can change the sign of some stories (red squares). Thus, by choosing $\alpha \leq 1$ we could ensure that stories with large but smaller numbers of links than the numbers of references to the original content, still obtain positive values. In this way we can encode the general popularity of the news sharing platform, e. g. to account for a smaller target audience of the website due language barriers.

The definition of UGC_+ has several advantages. First, the index can be easily calculated for any user-generated page.

*This work has been partially funded by the Torres Quevedo Program from the Spanish Ministry of Science and Innovation, co-funded by the European Social Fund.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

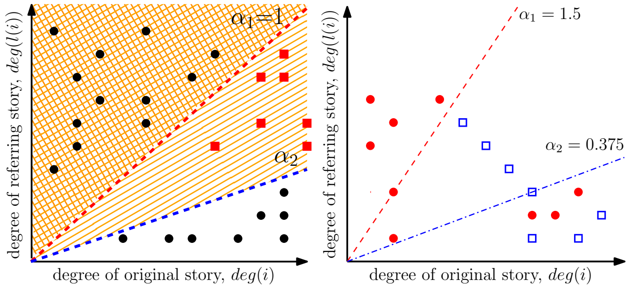


Figure 1: The meaning of α parameter

Second, to obtain a high UGC_+ it is not only important to have a prevalence of links to the story on the social news website in contrast to the original content (with respect to α) but also to have a large number of these references.

2.2 Comparison of groups of links

The parameter α can also be used for comparison of groups of links, e.g. for the analysis of user successfulness based on the set of links that the user has published, or for the study of what link categories provoke user generated content that is further referred to from other web pages.

For every group we define α such that it equilibrates the fraction of the positive and negative UGC_+ indexes for the links in the group. Formally, if $\{l_1^k(i_1), \ldots, l_n^k(i_n)\}$ is the set of all published stories with a certain property, then the corresponding α_k is defined as

$$\alpha_k = \text{median}\left(\left\{\frac{\deg(l_1^k(i_1))}{\deg(i_1)}, \ldots, \frac{\deg(l_n^k(i_n))}{\deg(i_n)}\right\}\right).$$

Since all original pages have been cited at least on the news sharing website, then $\deg(i) \neq 0$ for every i. In Figure 1(right) the value of α_k corresponds to the slope of the line that separates the stories of the same group into two equal sets. If there are more links with large numbers of citations to the social news website's entries, then the value of α is larger and vice versa.

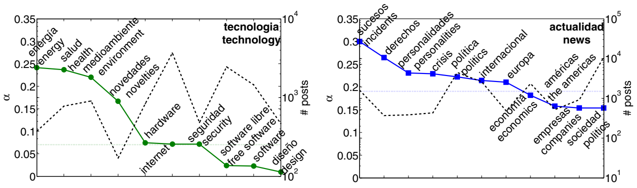
2.3 Comparison of individual users.

To rank website's users individually we also introduce a h_+*-index* (a variant of the h-index [2]) that measures both, productivity and impact of a user. We denote as $\{UGC_+^j\}_{j=1}^n$ the sorted (in descending order) list of the UGC_+ indexes of the n published links of a user and define the h_+-index of the user as the maximum rank number j such that $UGC_+^j \geq j$, or in other words, the user has h_+ stories with a UGC_+-index greater than h_+, but not $h_+ + 1$ stories that fulfil this condition.

Here we apply the analogy between scientific papers and links on social media websites. The advantage of h_+-index is that it is not easily influenced by the number of stories published by a user and by one lucky shot, meaning that there is a story that by chance became very popular.

3. NUMERICAL ANALYSIS

Menéame is the most successful Spanish social news website. We use a dataset which covers the time span from Dec. 7th, 2005 (release date) till July 14th, 2009. The data set contains 625 995 stories, 59 303 of which got published (received enough votes). For these published stories we used `Yahoo!Search BOSS` to obtain both, the number of links from other web pages to the Menéame story and to the original web page the Menéame story refers to.



(a) technology: $\alpha = 0.07$

(b) news: $\alpha = 0.19$

Figure 2: Distributions of α (continuous, in color) per topic categories vs the number of stories (posts) per category (dashed lines) for two parent categories. Dotted lines indicates α of parent category.

In Figure 2 we present the obtained α's for the groups of published stories assigned to selected categories. The shown categories belong to two parent categories: 'news' and 'technology' (dotted lines indicate the corresponding α's of the parent categories). We also plot the distribution of the number of stories (posts) per category (the dashed lines). We note that in some cases a small number of stories within a group may lead to a larger value of α, however in general we cannot draw any conclusion about a dependency between these characteristics.

We observe that the most popular categories for publishing stories (about society, Internet and free software) do not attract a lot of attention from outside of the news sharing websites, whereas the Menéame stories about current news and events have a higher probability to be cited, e. g. to be commented valuably or to receive interesting comments. Thus, the users tend to generate more valuable content about current news links rather than about links about technology. In particular, links about software (*software* and *software libre*) obtain only $\alpha = 0.024$ and 0.023, respectively, while the category about incidents (*sucesos*) has the largest value of $\alpha = 0.3$. We suspect that UGC_+-index indicates the real preferences of the users of the social news website better than traditional measures (e.g. the number of posts per category).

4. CONCLUSIONS

In this paper we have introduced an approach to quantify the interestingness and attractiveness of user generated content in respect to the original content. We believe that this approach should be further investigated both, by defining UGC_+ through more sophisticated page measures (e.g. PageRank) and by applying the framework for social evaluation of the content on some other websites where participants are allowed to share and to comment external links.

5. REFERENCES

[1] M. Castells. *Communication power*. Oxford University Press, USA, 2009.

[2] J. E. Hirsch. An index to quantify an individual's scientific research output. *PNAS*, 102(46):16569, 2005.