

Characterizing High-impact Features for Content Retention in Social Web Applications

Kaweh Djafari Naini, Ricardo Kawase, Nattiya Kanhabua and Claudia Niederée

L3S Research Center / Leibniz Universität Hannover, Germany
{naini, kawase, kanhabua, niederee}@L3S.de

ABSTRACT

One of the core challenges of automatically creating Social Web summaries is to decide which posts to *remember*, i.e., to consider for summary inclusion and which to *forget*. Keeping everything would overwhelm the user and would also neglect the often intentionally ephemeral nature of Social Web posts. In this paper, we analyze high-impact features that characterize memorable posts as a first step for this selection process. Our work is based on a user evaluation for discovering human expectations towards content retention.

Categories and Subject Descriptors

H.5.m. [Information Interfaces and Presentation]: Miscellaneous

General Terms

Classification, Experimentation, Aggregation

Keywords

Social Web Application, Managed Forgetting, Content Retention, Feature Analysis

1. INTRODUCTION

Social web applications, e.g., Facebook and Twitter clearly have a focus on the current situation and interactions as well as on what is coming up. Over time, there is, however, a lot of information created by individual users and some of the popular Social Web applications are now already existing for many years. This raises the issue of what to do with the accumulating information and what can be gained from it - beyond enabling more effective advertisement.

By documenting personal life [2] this information clearly constitutes an asset. For example, the large volumes of digital photo- and videography created and shared by individuals today are considered a valuable part of personal remembrance [4, 5]. However, an ever-growing information space tends to clutter. Recent work [2, 6] has shown the interest of users in also using social media content for reminiscence and self-reflection and the potentials of social media content for this task. In [6] for example a study with Facebook users

has discovered a considerable interest in managing a *personal region* for personal reminiscence and reflection about oneself. Automatically identifying real-world, public events in social media for browsing and search has been studied in previous work [1]. However, there is little work on identifying personal events, or autobiographical events (a subjective sense of times, places, associated emotions, and other contextual knowledge) in social media, which is considered a first step towards long-term summarization.

Therefore, we think of added value for the user and his/her social network by capturing personal history, harvesting the daily activities in an intelligent way and condensing them into different types of life summaries. Such summaries are not only useful for personal remembering. They also can provide an important source for catching up with what happened in the lives of not-so-close friends (e.g., former class mates), whose activities we do not have the time or interest to follow on a day-to-day basis.

Creating Social Web summaries, which are compact and meet human expectations on what to remember (and what to forget) is, however, a challenging task [3, 6]. In this light, this work provides important contributions towards a model for content retention preferences in social media and social networks. In more detail: (i) We design a user study, which enables insights into retention preferences, while respecting users privacy; and (ii) We present an analysis that provides us with evidences of the most remarkable features.

2. EVALUATION

Methodology. The goal of this evaluation was to collect participants' opinions regarding the retention preferences for their own Facebook posts, shares and updates. In our user study, each participant had to judge their own posts on a 5-point Likert scale answering the following question: *How relevant is your post for future reference?* We asked participants to judge at least 100 of their posts. Participants were recruited through research communities on Facebook where we posted the link to the evaluation and kindly asked for their contribution. Extra care was taken to respect user's privacy and to comply with Facebook's Platform Policies¹. First of all, the collected data will of course not be disclosed to third parties. Furthermore, the actual content of the posts (text or photos) is not analysed in our experiments. In general, the data cached represent the minimal amount of information needed in order to perform the experiments.

Results and Data Analysis. In total, we had 20 participants, 15 males and 5 females ranging from age 25 to 37.

¹<https://developers.facebook.com/policy/>

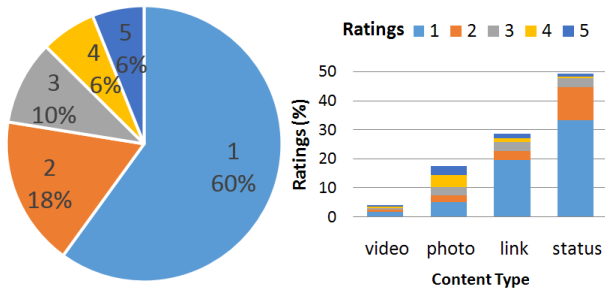


Figure 1: On the left, the overall distribution of ratings (average 1.81, standard deviation 1.21, variance 1.47). On the right, the distribution of ratings by content types.

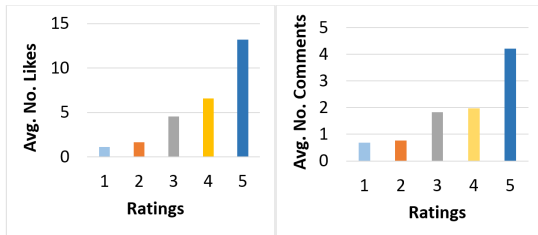


Figure 2: Average numbers of likes and comments by ratings (scores from 1 to 5).

Together they evaluated 3,330 posts. Additionally, once the user provided us authorization to access their data, we were able to collect general numbers that help us to study the general use of Facebook Social network.

Figure 1 (left) depicts the distribution of the ratings. We clearly identify the dominance (60%) of irrelevant posts (1 star). Further, Facebook defines seven types of posts, namely: link, checkin, offer, photo, question, swf and video. These basically describe the type of content that is attached to a post. Figure 1 (right) shows the distribution of posts among these categories. 49% of the evaluated posts consist of status updates, followed by shared links (28%), photos (17%) and videos (4%).

Disregarding *swf* and *offer* types (which do not have sufficient occurrences to be significant), we found out that *links* and *status* updates have the highest irrelevance rates with 69% and 67% 1-star votes. In contrast, *photo* updates seem to be the most memorable type. Only 30% of photos were rated as irrelevant, moreover, this type of posts leads the 5-starts votes with 18%.

On Facebook, users are able to comment or to like a particular post. In the most common case (default setup), a user's post is visible for his network of friends, and those are the ones that are able to comment and to like. A first analysis of the results suggests that the average number of comments and average number of likes for the evaluated posts has a considerable impact on the rating of the post (Figure 2): the higher the number of likes/comments, the higher the rating.

Additionally, as depicted in Figure 3, we see the average ratings over time (January 2009 - November 2013). The image shows a clear trend where participants in the evaluation assigned higher ratings to more recent posts. This is in line with the idea of a decay function underlying the content retention model. From this statistics, one can deduce first ideas

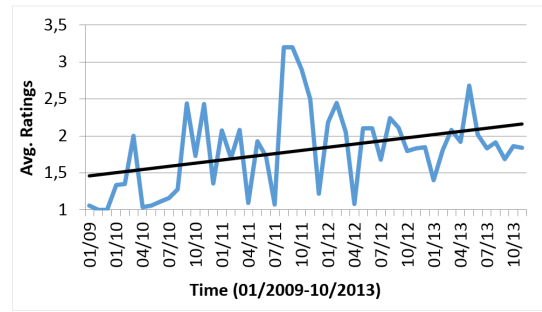


Figure 3: Average ratings over time.

for the features that have a higher impact in the detection of memorable posts. In summary, the type of the post and its number of likes and comments have a high impact, while the age of the post should be considered as well.

We have observed that other types of features based on the metadata of each post (e.g. *privacy*, *appType*, *hasLink*), as well as network measures, (e.g. *clustercoefficient*, *connected-components*, *density*), computed on likes and comments network of the post, have a considerably lower impact on the memorability of a post.

3. CONCLUSION

In this work, we have roughly identified features for classifying memorable posts. The data analysis presented in this paper is just a first step for the envisioned life summaries. As future work, we plan to increase the reach and the participation in our evaluation to collect a data set that can better represent the average social network users. Based on this, we plan to empirically identify the most important features that characterize memorable posts. In addition, we are planning research work for gaining a deeper understanding of *personal* retention preferences, since not all users use a social web application in the same way.

Acknowledgments This work was partially funded by the European Commission Seventh Framework Program under grant agreement No.600826 for the ForgetIT project.

4. REFERENCES

- [1] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *Proceedings of WSDM '2010*, 2010.
- [2] S. Greengard. Digitally possessed. *Commun. ACM*, 55(5):14–16, May 2012.
- [3] N. Kanhabua, C. Niederée, and W. Siberski. Towards concise preservation by managed forgetting: Research issues and case study. In *Proceedings of iPres '2013*, 2013.
- [4] D. S. Kirk and A. Sellen. On human remains: Values and practice in the home archiving of cherished objects. *ACM Trans. Comput.-Hum. Interact.*, 17(3):10:1–10:43, July 2010.
- [5] S. E. Lindley, C. C. Marshall, R. Banks, A. Sellen, and T. Regan. Rethinking the web as a personal archive. In *Proceedings of WWW '2013*, 2013.
- [6] X. Zhao, N. Salehi, S. Naranjit, S. Alwaalan, S. Volda, and D. Cosley. The many faces of facebook: Experiencing social media as performance, exhibition, and personal archive. In *Proceedings of CHI '2013*, 2013.