AN ANALYSIS OF POST CONTRIBUTIONS IN THE HACKER COMMUNITY

Alex Tsang
Department of Information Systems
City University of Hong Kong
xzeng9@alumni.cityu.edu.hk

Wei Thoo Yue
Associate Professor
Department of Information Systems
City University of Hong Kong
wei.t.yue@cityu.edu.hk

Michael Chau
Associate Professor
School of Business
The University of Hong Kong
mchau@business.hku.hk

Abstract

We are interested in how the posters in the hacker community contribute and exchange information. Text-mining techniques have been used to learn about the quality nature of the posts. We found that the knowledge exchanges in the hacker community are both interesting and complex. We uncover some interesting knowledge exchange behavioral patterns between initial post contributor and post repliers (initiators vs. followers). Namely, popular threads, i.e., threads with more replies, actually generate lower quality discussions in replies. On the other hand, we observe a higher percentage of quality replies with less popular threads. The results show that thread popularity does not immediately imply valuable discussions. In fact, threads with lower initial post quality are often associated with higher quality yet less popularity discussions

Keywords: Text mining, hacker community, supervised classification, contribution analysis

1. Introduction

Web forum has emerged as a crucial venue for many to share hacking knowledge. Apart from a lot of non-consequential conversational patterns, we are more interested in the high value information conveyed and shared in forums. Such information represents the "essence" of the knowledge accumulated in forums, and also reflects the value of forums in the eyes of hackers. In this study, we analyze the post contents to measure the overall value exhibited by hacker forums, and investigate the structural characteristics exhibited in the knowledge exchange and discovery processes. To facilitate the analysis, a text-mining based classification methodology was adopted to distinguish the quality levels for the two integral elements of a forum thread - header and reply. By integrating all of the classification results, some discoveries were made, which showed that the proportional distribution of valuable information contributed in the threads varies with the initiating header of different quality levels, while the overall contributions by the summation of quality posts in all of their threads are nevertheless basically balanced.

2. Related Work

Plenty of work concerning contribution analysis and quality measurement techniques for technical web forums had been reported. Chai et al. [**2**] proposed an automated approach to perform a similar quality measurement task for forum posts, which relies on a conceptual model and a classification approach to monitor the users "consuming forum posts." Aumayr et al. [**1**] and Wang et al. [**8**] raised their respective methodologies to re-construct the thread structures for content-centric analysis, both aiming at facilitating the positioning of high-value information in the thread-level interaction activities. Focusing on user-centric analysis, Chai et al. [**3**] established a User Contribution Measurement (UCM) model for web-based discussion forums, which took into account a lot of quantified post features for classification; and Lui et al. [**6**] proposed a more comprehensive model for user classification based on their defined user-level attributes to generate weighted scores, finally leading to a conclusive classification in terms of user competency levels.

3. Data Collection and Sampling

The hacker forum under study has more than 6 years of history and gathered over 145,000 active users by 2012, which had accumulatively generated approximately 3 million posts organized in 354,286 threads. Their topics extensively cover hacking-supportive technologies, such as programming, scripting, sniffing, scanning, remote administration tool; as well as direct hacking techniques such as website defacement, SQL injection, wireless hacking, keylogging, etc. Due to their large amount, sampling was done by selecting 50,000 headers randomly and indiscriminately along with their 357,721 replies to collectively constitute our data set for study.

4. Model and Implementation

4.1 Classification Basis

As a basic constituent unit for any web forums, each thread consists of a header and various following replies, which vary in terms of evaluation criteria for their significance to the viewers or the contribution to the forum. To facilitate the employment of expert judgment, we define the terms "quality" and "useful" to indicate their respective contributing nature as follows:

Quality: *the content of the header in question is inspiring and relevant to the hacking topics, such as teaching, introduction, disclosure, explanation, demonstration, troubleshooting, or any other forms presenting some kind of knowledge or information.*

Useful: *the content of the reply in question is inspiring and relevant to the corresponding header, such as strengthening, complement, disproving, proliferation or any other forms capable of introducing or stimulating more knowledge or information to share or reveal.*

These two dimensions have thus served as a basis throughout the classification process of headers and replies. Given the text-heavy nature of forum posts, we used Joachim's SVM^{light} utility [**5**] for its superior performance on text categorization tasks.

4.2 Features for Extraction

By expert judgment, we observed there were a total of 9 data fields in each post relevant to the classification. The usefulness of a reply is closely related to the title of the entire thread, making it one of the influential factors. Both headers and replies had therefore shared an identical list of classification features as follows:

Fields for features	Description
postTitle	The title of the entire thread.
postContent	The content of the first post in the thread, i.e., the header post.
imageCount	Number of images appearing in the content.
linkCount	Number of hyperlinks appearing in the content.
postView	Times for which this post has been viewed.
postReply	Times for which this post has been replied to.
userLevel	The user's level when posting the post.
userPostNumber	Quantity of posts by this user when posting the post.
postLength	Word count of the *postContent*.

[*Table 1: Classification Features*]

Owing to the requirement of the SVM^{light} learner, the text contents of *postTitle* and *postContent* are converted into standard tf-idf values and indexed with ordered IDs, and the rest of other plain numerical features are appended and ordered accordingly.

4.3 Keyword Extraction

As the classification is performed on the headers and their replies respectively, two SVM^{light} models are required and three sets of keyword lists are needed for all the *postTitle* and *postContent* for the headers and for the replies. To avoid any biases, each list of keywords is extracted automatically from the respective sources via an uninterfered process, in which a standard Porter stemming algorithm [7] is performed at first to strip any suffix of each single word, then all these words are aggregated again as a space-delimited text for N-gram extraction, where N ranges from 1 to 5. With regard to each candidate, it becomes a keyword for *postTitle* if included therein for no less than twice, or becomes a keyword for *postContent* if included for no less than 5 times. Given the 50,000 headers and 357,721 replies in our corpus, we finally extracted 137 keywords for titles, 685 for header contents and 1,282 for reply contents.

	postTitle	postContent
Header	crack defac password question revers tut	download obvious tutori http www spoiler click youtub com/watch
Reply	fud hack hash help keylogg password problem site server	Admin comput crypter exactli recommend try version work net/showthread php

[*Table 2: Keyword Samples for Quality Headers and Useful Replies*]

With the intent to boost the performance and shorten the classification time later, an IDF (Inverse Document Frequency) value for each keyword is calculated and reserved at this stage:

, where *D* is the entire corpus of headers or replies, *d* is a single post in the corresponding *D*, *t* for "*term*" is the current keyword in question; so the numerator and denominator are actually the total number of posts in the current corpus and the number of posts containing the current keyword, respectively.

4.4 Model Creation

Training sets and test sets were created for both headers and repliers to generate their respective models used by the SVM classifier. To ensure comprehensiveness, 11 headers were randomly selected across each of the 19 sub-boards of the forum to construct the training set for headers, with another 6 from each of the same sub-boards to construct the test set for headers. Replies following these headers were correspondingly grouped into training and test sets for replies by nature. Each of these entries selected passed through a manual classification to determine its significance for either model generation or test result verification.

	Training set		**Test set**	
Headers	209	Quality: 63 (30.1%)	114	Quality: 34 (29.8%)
Replies	1,922	Useful: 358(18.6%)	920	Useful: 273(29.7%)

[*Table 3*: *Training sets and test sets with their quantities*]

Each post in both sets was then quantified as a list of feature items, largely comprised of tf/idf values converted from its title and content, which were scanned for any keyword falling in the keyword lists extracted beforehand. And each identified keyword was counted for the number of occurrence in its residing place to calculate its Term Frequency, and then multiplied by its corresponding IDF value reserved during their extraction:

, where *tf(t,d)* is obtained by dividing the number of occurrences of the keyword (*t*) by the total number of words of the current post (*d*).

Supplemented with the other 7 numerical features and a few rounds of parameter tunings, we finally obtained 2 effective classification models for both headers and replies with the following accuracy, precision, and recall values:

Headers:

Accuracy on test set: 88.60% (101 correct, 13 incorrect, 114 total)
Precision/recall on test set: 80.00%/82.35%

Replies:

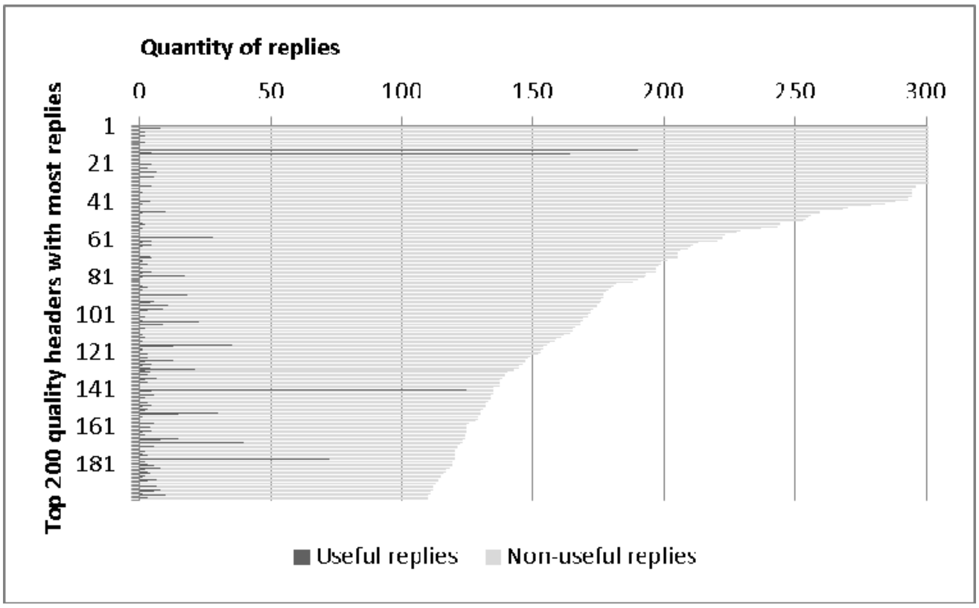
Accuracy on test set: 88.70% (816 correct, 104 incorrect, 920 total)
Precision/recall on test set: 80.73%/81.32%

4.5 Header and Reply Classification

With both models ready, contribution classification on headers and replies were thereby enabled. Each header was tagged as "quality/non-quality" while each reply as "useful/non-useful" to its followed header. All classification results of both groups were then integrated and analyzed from different perspectives to unveil any meaningful facts.

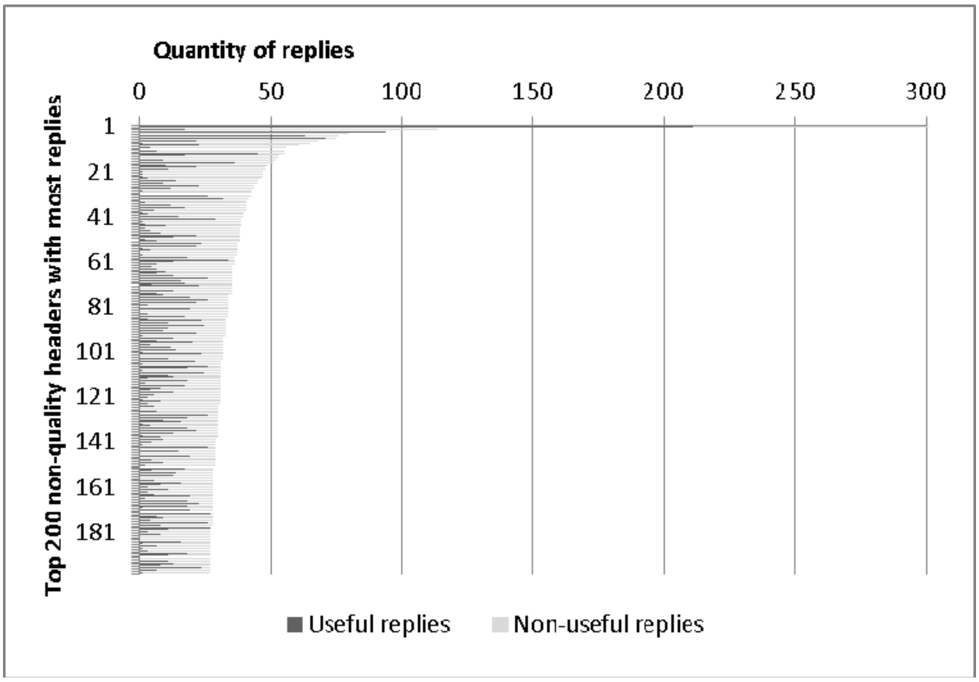
5. Results and Analyses

The classification results for both headers and replies exhibit some kind of distinct and obvious phenomenon, especially with their integration as appearing in the thread structure. The raw headers were first organized into two groups according to their classification results. Next, these headers were attached by all of their respective replies to re-establish the structure of the original threads. All the headers in each group were then compiled in descending order by their "visible influence", i.e., the quantity of replies, generating the following two charts showing the differentiated distribution between useful and non-useful replies following headers of different quality natures: (*see Figure 1 and Figure 2*)



[*Figure 1: Threads Initiated by Quality Headers (Group 1)*]

For brevity, in both groups, only the top 200 threads (headers) were shown. By comparison, the quality headers represented in Figure 1 distinctly attract more replies than the non-quality ones represented in Figure 2; however, most of them were also made of non-useful ones. On the contrary, despite of the fewer replies following the non-quality headers, the useful ones appeared much more actively and take higher percentage. This may reflect a fact that, header of high value in terms of information or knowledge sharing are naturally receiving more acknowledgement, but such commendations in the form of replies are less inclined to present more of this value as the header did. Meanwhile, headers in the form of questions or some immature or even incorrect statements undoubtedly attract lower popularity by their nature, but those people who are willing to reply are more inclined to provide the values lacked by the header itself, and are more likely to expand the topics or light up the sparkles among the viewers. These two figures had also highlighted the different positions of contribution weights concentrated in threads initiated with quality and non-quality headers.



[*Figure 2: Threads Initiated by Non-quality Headers (Group 2)*]

When summed up in a table, the quantities of classified headers and replies further had drawn the following observations:

		Reply		
		Useful	**Non-useful**	**subtotal**
Header	**Quality**: 16,908 (33.8% of headers)	58,754 (27.6% of subtotal)	153,848 (72.4% of subtotal)	212,602 (59.4%)
	Non-quality: 33,093 (66.2% of headers)	71,908 (49.5% of subtotal)	73,219 (50.5% of subtotal)	145,127 (40.6%)

[*Table 4*: *Quantity of classified headers and replies*]

Observation 1: Reading horizontally in the table, we can see the quantity of quality headers effectively attracted the majority of replies. Also, due to the larger quantity of replies, each quality header stimulates around 3.47 useful replies on average, compared to 2.17 useful replies following the each non-quality header.

Observation 2: Reading vertically in the table, however, we can see that despite the fewer replies attracted by the higher quantity of non-quality headers, they nevertheless gather a distinctively higher percentage (49.5%) of useful replies than the quality headers (27.6%), recapping the implication that repliers are more stimulated to contribute knowledge and information to those non-quality headers.

Observation 3: Intuitively, since the thread as the basic structural unit of a web forum is typically viewed as a whole, the quality of the header and the usefulness of all its replies collectively constitute an overall "contribution degree" on behalf of the thread to all of their viewers as well as the forum itself. In such a sense, the total contribution weights from either of the groups of classified threads initiated by quality and non-quality headers can be simply regarded as the summation of quality header plus useful replies, and this sum value of Group 1 (16,908 + 58,754 = 75,662) is actually very close to that of Group 2 (71,908).

6. Conclusion and Future Work

By analyzing the hacker forum posts, we have derived some basic facts concerning the identifying high contribution posts in a hacker forum, and how different type of posts initiate the different thread initiators and the followers exchange pattern. With more replies and fewer useful ones, quality headers may falsely convey the impression of stimulating valuable exchange. In reality, such initial posts may attract more valuable replies on average, but the concentration of quality discussions in actually lower. Finally, we see an equal amount of quality and non-quality header posts in the forum. For the viewers, the useful information is apparently much easier to locate and find in the replies of non-quality headers due to their smaller number and higher concentration degree. These conclusions can help us further understand the knowledge exchange and discovery patterns exhibited in the hacker forum, which is associated with high degree of learning patterns. Using information diffusion models [**4**], we can further analyze how knowledge is disseminated from knowledge leaders to other knowledge learners in future studies.

References

1. Aumayr, E., Chan, J., and Hayes, C. (2011). Reconstruction of Threaded Conversations in Online Discussion Forums. In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, AAAI 2011*.
2. Chai, K., Hayati, P., Potdar, V., Wu, C., and Talevski, A. (2010). Assessing Post Usage for Measuring the Quality of Forum Posts. In: *IEEE International Conference on Digital Ecosystems and Technologies (DEST 2010)*, pp. 233-238
3. Chai, K., Potdar, V., and Chang, E. (2009). User Contribution Measurement Model for Web-based Discussion Forums. In: *3rd IEEE International Conference on Digital Ecosystems and Technologies, DEST '09*, pp. 347-352.
4. Chau, M. and Xu, J. (2012). Business Intelligence in Blogs: Understanding Consumer Interactions and Communities. *MIS Quarterly* 36(4), pp. 1189-1216.
5. Joachims, T. (2002). Learning to Classify Text Using Support Vector Machines. Dissertation, Kluwer
6. Lui, M. and Baldwin, T. (2010). Classifying User Forum Participants: Separating the Gurus from the Hacks, and Other Tales of the Internet. In: *Proceedings of Australasian Language Technology Association Workshop*, pp. 49-57
7. Porter, M. (1980). An algorithm for suffix stripping, in Program, 14(3) pp. 130−137.
8. Wang, L., Kim, S.N., and Baldwin, T. (2010). Thread-level Analysis over Technical User Forum Data. In: *Proceedings of Australasian Language Technology Association Workshop*, pp. 27-31