# Blog Owner Detection: User Reference Matrix

Roland Hou-Yin Hui Akihiro Miyata Harumi Kawashima Hidenori Okuda
NTT Cyber Solutions Laboratories, NTT Corporation
<roland.hui, miyata.akihiro, kawashima.harumi, okuda.hidenori>@lab.ntt.co.jp

## ABSTRACT

We present a behavioral method to detect the blog owner of a blog entry by evaluating the comment thread with a proposed method called the "User Reference Matrix". For each comment, we analyze the references, if any exist, made to another comment sender and tabulate the results into four separate factors which together, make up the User Matrix Score (UMS). The UMS determines the blog owner and an evaluation of both the accuracy and coverage, using our test data, yielded expected results. Furthermore, our method also compares favorably against two other blog owner detection methods, "URL Detection" and the "String Match" method.

## 1. INTRODUCTION

In recent years, weblogs, better known as blogs, have become increasingly popular among the online community. As of April 2007, Technorati, the largest blog search engine, tracked over 70 million blogs with about 120,000 new blogs being created each day [1]. Blog entries can range anywhere from personal diaries to journalistic articles and they also act as a catalyst for communication through comments. As a result, comments are integral components that help differentiate a blog from a simple webpage.

## 2. RELATED WORK AND BACKGROUND

Previous works involving the analysis of blog comments treat all comments alike and therefore evaluate each comment with the same algorithm. Beibei et al. used blog comments to enhance blog clustering [2]. Herring et al. considered the number of blog comments when analyzing blogs [3]. Miyata et al. also considered the number of blog comments as well as the time interval between comments and unique comment senders in their blog search algorithm [4]. Current comment analysis methods ignore the comment sender information which should not be overlooked. Comments from blog owners may be more valuable than those from ordinary senders, and therefore, should be evaluated differently. Humans are able to identify the blog owner through the use of context and reasoning. Automated blog owner detection, however, lacks these abilities and is therefore much more difficult. Our goal is to develop an automated method which has the ability to identify the blog owner by using the comment thread of a blog entry.

## 3. PRELIMINARY SURVEY

Our database consists of over 279000 crawled blog entries, 32% of the entries have at least one comment and 28% have at least two comments. This database, which we use to populate our dataset and calculate general statistics upon, is populated with various Japanese written blogs. Our "blog crawler," a system that finds, analyzes and stores blogs in a database, uses the goo blog search engine [5] to collect random blog entries; ordering the search results by relevancy as opposed to recent entry and collecting the first 500 blog entries returned from each keyword.

The dataset, which we use to perform most of our experiments on, consists of over 1600 commented blog entries randomly chosen from our database. Each blog entry in our dataset contains at least two comments. We decided to exclude blog entries with fewer than two comments from our dataset because our research approach evaluates the comment thread of a blog entry. More specifically, our research approach evaluates the conversation within a comment thread, which we define as having two or more comments. From our dataset, we manually evaluate each blog's owner and use this information to compare our method's results against two simpler methods.

Before we proposed any new approaches, we first surveyed two simple methods for blog owner detection. The first method compares the URL, taken from a comment sender's name's signature, with the actual blog URL. This method, appropriately named "URL Detection" method, proved to be extremely accurate. However, because a URL signature is optional, only 7% of blog owners from our dataset supplied a URL when making a comment on their own blog.

The second method, the "String Match" method, compares the account name and blog title to the comment sender's name. The account name is the blog host ID, in our case the goo ID, and is retrieved from the blog URL while the title is retrieved by parsing the blog HTML. Again, accuracy is high, but this method can only detect approximately 31% of our dataset. These two methods, although extremely accurate, did not quite meet our expectations, as they only cover (coverage) 31% of our dataset.

## 4. PROPOSAL

We now propose a new approach with enhanced coverage which detects comment referencing (described below) in comments. This idea is based on an observed behavior that blog comments are used primarily to communicate with the blog owner; although communication between non-blog owners also exists. Furthermore, this behavioral approach can easily be applied to blogs written in other languages as it is language independent.

Comment referencing is established when one sender's comment contains another sender's name. We detect comment referencing by comparing a comment entry with a list of all comment sender names. Comments that do not contain another sender's name are ignored. For each sender's comment(s) in the comment thread, we evaluate two important factors based on comment referencing, which we use to determine the blog owner: (1) the number of comments by sender 'X' which references to other comment sender(s) and (2) the number of comments sender 'X' is referenced by other comment sender(s). Tabulating the factors of each sender, we can identify the person that refers to other senders many times and is also referred by other senders many times. Conveniently, such a person is highly correlated with being the blog owner according to our observations.

However, we notice some inconsistencies in our results when we only consider the sheer quantity of comment referencing (Factors 1 and 2). These inconsistencies are attributed to "noise," defined as conversations amongst comment senders which do not involve the blog owner. Therefore, we evaluate two more factors aimed specifically to filter out this "noise": (3) the number of unique comment senders referenced by sender 'X' and (4) the number of unique comment senders which reference sender 'X'. With these two factors, we can detect the strength of comment sender involvement in every conversation. "Noise" will have relatively few comment senders and as such, those users will have a low factor 3 and 4 score.

The four factors can be represented visually in a matrix, appropriately named the "User Reference Matrix," see Figure 1 for an example. The matrix operates as follows: For every sender, we calculate the score of each factor. Looking at sender "B" from Figure 1, the number of comments by sender B which references any other comment sender(s) is just the sum of row "B", four. Similarly, the number of comments sender "B" is referenced by any other comment sender(s) is the sum of the column "B", three. The number of unique comment senders referenced by sender "B" is the number of non-zero cells in row "B", three. Likewise, the number of unique comment senders which reference sender "B" is the number of non-zero cells in column "B", two.

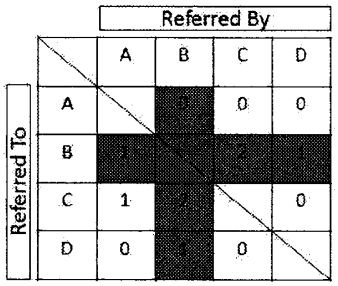


**Figure 1:** Example of a "User Reference Matrix"

We use these four factors to calculate a User Matrix Score (UMS) for each comment sender and we determine the blog owner as the sender with the highest UMS. The UMS is determined by the following linear algorithm:

$$UMS = w_1S_1 + w_2S_2 + w_3S_3 + w_4S_4 \quad (1)$$

Where $S_1 \rightarrow S_4$ are the scores from each of the four factors and $w_1 \rightarrow w_4$ are the weights applied to each score. We will describe the UMS in more detail in the following section.

# 4. EVALUATION

We evaluate the accuracy and coverage of our proposed solution, the "User Reference Matrix" method, to determine how our solution fares against the "URL Detection" and "String Match" methods. With our dataset defined above, we first separate our dataset into three smaller subsets as follows: the "Learning" set is used in conjunction with the "User Reference Matrix" to realize an effective way of combining the four factors together. The other two subsets form our "Testing" set, which we use to test the accuracy and coverage of our system. "Testing" set A can be approximated as a reduced corpus of the entire commented blogosphere due to the random retrieval algorithm used to obtain these blog entries. "Testing" set B is artificially inflated with blog entries containing only "large" number of comments (41+ comments) to compensate for "Testing" set A's lack of blog entries with a "large" number of comments. This set is essential to evaluate the accuracy and coverage of blog entries as comment threads increase in size. See Table 1 for specifications of each subset.

**Table 1:** Dataset specifications

| Dataset | Learning Set | Testing Set A | Testing Set B |
|---|---|---|---|
| # of Entries | 373 | 967 | 233 |
| # of Entries with 2-20 Cmts | 353 | 918 | 0 |
| # of Entries with 21-40 Cmts | 15 | 37 | 0 |
| # of Entries with 41+ Cmts | 5 | 12 | 233 |

As previously mentioned, our approach assigns a UMS to each comment sender and the blog owner is simply the sender with the highest UMS. The weights $w_1 \rightarrow w_4$ in formula (1) are derived from the "Learning" set. We calculated the accuracy of each factor by comparing the output of each individual factor to the list of blog owners we manually obtained. See Table 2 for results.

**Table 2:** Accuracy and weights of each factor

| | Accuracy | Weights |
|---|---|---|
| Factor 1 | 97% | $w_1 = 97/238 = 0.407$ |
| Factor 2 | 20% | $w_2 = 20/238 = 0.084$ |
| Factor 3 | 98% | $w_3 = 98/238 = 0.412$ |
| Factor 4 | 23% | $w_4 = 23/238 = 0.097$ |
| Total: | 238% | 1 |

To calculate the accuracy and coverage of the "User Reference Matrix" method, for each blog entry in our two "Testing" sets, we calculate the UMS for each comment sender. We then compare the calculated blog owner with the manually obtained blog owner. Furthermore, we compare our results from the "User Reference Matrix" against the "URL Detection" and "String Match" methods, tabulating the results in Table 3.

**Table 3:** Accuracy and coverage of each method

| Method: | User Reference Matrix | URL Detection | String Match |
|---|---|---|---|
| "Testing" Set A | Coverage: 80% Accuracy: 95% | Coverage: 7% Accuracy: 99% | Coverage: 31% Accuracy: 97% |
| "Testing" Set B | Coverage: 99% Accuracy: 69% | Coverage: 10% Accuracy: 83% | Coverage: 25% Accuracy: 92% |

From Table 3, we see that the "User Reference Matrix" method has a much greater coverage than the "URL Detection" and "String Match" methods. It also retains a high degree of accuracy for blog entries with relatively few comments. Taking into consideration "Testing" set B, we can see that the accuracy starts to suffer as the number of comments increase. However, a statistical evaluation using our blog database shows only 1.2% of blog entries, with two or more comments, having greater than forty comments. We believe this percentage is similar in the blogosphere as well due to the size of our database, and therefore, this decrease in accuracy is negligible.

Nonetheless, we propose several options to enhance our proposed method. We tested a system that combined the "URL Detection," "String Match," and "User Reference Matrix" methods by using a decision tree and witnessed an improvement in both accuracy and coverage, 96% and 86%, respectively, on "Testing" set A. In addition, utilizing previous blog entries from the same blog when applying the "User Reference Matrix" method increases the overall accuracy of the system.

# 5. CONCLUSION

We proposed a behavioral approach for blog owner detection, the "User Reference Matrix" method, which proved to be quite effective. We suggested four main factors to evaluate from each comment thread and combined these factors in the form of a UMS to realize a blog owner. One of the highlights of a behavioral based approach is that it is easily portable to any blog sites independent of language. This technology can aid in future blog data-mining or blog owner profiling. If this technology is applied to blogging communities, it has the potential to identify comments posted by one blog owner in the community to another blog owner's comment thread. Enabling researchers to evaluate where and how blog owners are participating in the blogosphere.

# 6. REFERENCES

[1] Sifry, D. Sifry's Alerts: The State of the Live Web, April 2007.: http://www.sifry.com/alerts/archives/000493.html

[2] Beibei, L., Shuting, X., and Jun. Z. Enhance clustering blog documents by utilizing author/reader comments. In *Proceedings of the 45th Annual Southeast Regional Conference*, 2007

[3] Herring, S. C., Scheidt, L. A., Bonus, S., and Wright, E. Bridging the gap: A genre analysis of weblogs. In *Proceedings of the 37th Hawaii International Conference on System Sciences*, 2004.

[4] Miyata, A., Matsuoka H., Okano, S., Yamada, S., Ishiuchi, S., Arakawa, N. and Kato, Y.: Blog Search Method Based on Analysis of Response to a Blog Entry. In *IPSJ Journal Vol.48, No.12*, 2007, 4041-4050

[5] goo blog.: http://blog.goo.ne.jp