Review Spam Detection via Time Series Pattern Discovery

Sihong Xie† Guan Wang† Shuyang Lin† Philip S. Yu† ‡
† Department of Computer Science, University of Illioins at Chicago, Chicago, IL
‡ Computer Science Department King Abdulaziz University Jeddah, Saudi Arabia
{sxie6, gwang26, slin38, psyu}@uic.edu

ABSTRACT

Online reviews play a crucial role in today's electronic commerce. Due to the pervasive spam reviews, customers can be misled to buy low-quality products, while decent stores can be defamed by malicious reviews. We observe that, in reality, a great portion ($> 90\%$ in the data we study) of the reviewers write only one review (singleton review). These reviews are so enormous in number that they can almost determine a store's rating and impression. However, existing methods ignore these reviewers. To address this problem, we observe that the normal reviewers' arrival pattern is stable and uncorrelated to their rating pattern temporally. In contrast, spam attacks are usually bursty and either positively or negatively correlated to the rating. Thus, we propose to detect such attacks via unusually correlated temporal patterns. We identify and construct multidimensional time series based on aggregate statistics, in order to depict and mine such correlation. Experimental results show that the proposed method is effective in detecting singleton review attacks. We discover that singleton review is a significant source of spam reviews and largely affects the ratings of online stores.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining

General Terms

Algorithms, Experimentation

Keywords

Review spam, time series, adversarial data mining

1. INTRODUCTION

Previous works use review contents and reviewers' behaviors [2, 3, 1] to detect spams. These methods only work in the situations where spammers write many reviews. In reality, however, most reviewers write only one review. This is due to the nature of spamming. In order to manipulate the rating quickly without being caught, it is highly desirable for a spammer to post many reviews in a short time under different names. This spamming strategy makes it appear that most reviewers contribute only one review (called a *singleton review*, SR for short). Most of the statistics adopted by previous works would not work on such singleton reviews. For example, the mean and standard deviation of ratings given by a reviewer [2] become meaningless if this reviewer has written only one review. Based on abnormaly correlated temporal pattern mining, we construct multidimensional time series and present a method to detect these singleton review attacks. We also use multiple resolutions to handle short term fluctuations in time series and locate the time of attacks.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

2. THE PROPOSED TECHNIQUES

To raise or lower the rating of a store safely and rapidly, spammers tend to post a large number of reviews with a high or low rating under different names. Therefore, if there is a sharp increase in the volume of (singleton) reviews while the rating also increases or decreases dramatically, it is highly likely that the rating is manipulated by the newly arrived reviews. Therefore, we can detect SR attacks by exploiting the correlation between the rating and the volume of (singleton) reviews.

2.1 Time Series Construction

Given all reviews for a store, their ratings are $R(s) = \{r_1, \ldots, r_{n_s}\}$ with posting time $TS(s) = \{ts_1, \ldots, ts_{n_s}\}$, where $ts_i < ts_j$ for all $1 \leq i < j \leq n_s$. Given a time window size Δt, we can split the time into consecutive time windows $I_n = [t_0 + (n-1)\Delta t, t_0 + n\Delta t]$, $I = \bigcup_{n=1}^{N} I_n$. For each time window I_n, we calculate three aggregate statistics using the SRs falling into that time window.

$$f_1(I_n) = \sum_{ts_j \in I_n} r_j / f_2(I_n), \qquad f_2(I_n) = |\{r_j : ts_j \in I_n\}|$$

$$f_3(I_n) = |\{r_j : ts_j \in I_n, r_j \text{ comes from an SR}\}| / f_2(I_n)$$

$f_1(I_n)$, $f_2(I_n)$ and $f_3(I_n)$ are the average rating, the number of reviews and the ratio of singleton reviews within the time window I_n, respectively.

2.2 Correlated Abnormal Patterns Detection in Mul- tidimensional Time Series

We use a three-step approach for the detection. First, on each dimension, we fit a curve using the time series. We then apply longest common substring (LCS) algorithm to match the fitted curves to a template representing bursty patterns. We slide the template across each time series. For each time point during the sliding, LCS gives the number of

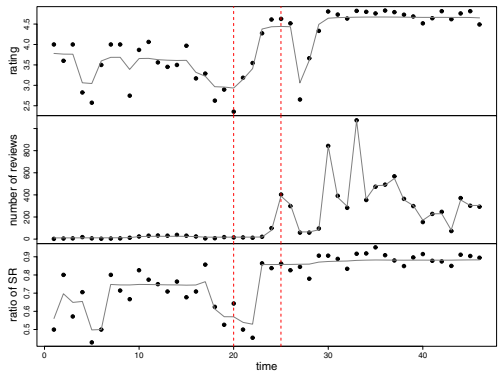

Figure 1: Bursty Patterns Detected in Store 24938

Table 1: Human evaluation results on stores

	Evaluator 1	Evaluator 2	Evaluator 3
Evaluator 1	17	14	16
Evaluator 2	-	20	19
Evaluator 3	-	-	24

matched points between the template and the current section of the time series overlapping the template. We also consider the range of the time series within the overlapping section. Putting the number of matches and the range together, we can measure the intensity of bursts for each time point on each dimension of time series. We retrieve top k bursty sections from each dimension. Lastly we slide a window of a certain size over the time axis. At each point, we find out how many top ranked locations in all dimensions are in the time frame specified by the current time window. A time window is reported if all three dimensions have bursty patterns fall into the window. One example of the output of the algorithm is shown in Figure 1.

2.2.1 *A Hierarchical Framework for Robust Singleton Review Spam Detection*

Given the review records of a store, one can construct the multidimensional time series with different time window sizes (resolutions). If the window size is set too small, the general trend of a time series would be buried in a large number of fluctuations, which might cause high false positive rate. We first smooth out short-term fluctuations using a larger window (lower resolution). We then fit curves out of these time series and detect any suspicious periods with correlated abnormal patterns, which indicate the high likelihood of SR spam attacks. A smaller window size (higher resolution) can be used to reveal more details (e.g. the exact time of the burst). This is accomplished by constructing new time series with a higher resolution on the detected periods, and detecting any finer suspicious period. This process continues until one reaches the desired resolution such that the time when the SR spam attack can be easily pinpointed.

3. EVALUATION

We evaluate the proposed algorithm on the review data crawled from a review website[1]. It contains 408,469 reviews written by 343,629 reviewers (identified by their id on the website) for 25,034 stores. 310,499 out of all reviewers ($> 90\%$) wrote only one reviews. We employ three human evaluators in this experiment.

Table 2: Human evaluation results on reviews

	Evaluator 1	Evaluator 2	Evaluator 3
Evaluator 1	59	20	28
Evaluator 2	-	41	38
Evaluator 3	-	-	72

In the evaluation we select 53 stores, each of which has more than 1,000 SRs. The proposed algorithm detects 33 stores out of them. For each detected store, we select reviews around the time point when the bursty patterns appear. These reviews are then examined by three evaluators and their opinions are recorded. If there are two or more evaluators believe that a store has ever committed an SR spam attack, we consider it to be a dishonest store. The precision of the proposed algorithm, when detecting such dishonest stores, is 75.86% (22/29). Table 1 shows the agreement between evaluators. The numbers on the diagonal show how many stores does each evaluator considers as dishonest. The other numbers give how many stores that both two evaluators identify as dishonest stores.

We also ask three human evaluators to examine 147 reviews contained in one of the detected bursts in one suspicious store. Each review is given a score (0-negative, 0.5-possibly, 1-positive) to indicate the degree of being regarded as a spam review by each evaluator. Among the 147 reviews, 43 reviews (38 are SR) have final score higher or equal to 2, and 12 reviews (11 are SR) have final score equal to 3. This indicates that the proposed algorithm can locate the period when singleton spams concentrate. Table 2 is similar to Table 1, except that the results are based on the human evaluations on the detected reviews.

4. CONCLUSION

This paper studies the problem of singleton review spam detection, which is both difficult and important to solve. We transform this problem to a temporal pattern discovery problem. We identify three aggregate statistics which are indicative of this type of spam attack, then we construct a multidimensional time series using these statistics. We design a multi-scale anomaly detection algorithm on multidimensional time series based on curve fitting. Experimental results show that the proposed algorithm is effective in detecting singleton review spams.

5. ACKNOWLEDGEMENTS

This work is supported in part by Google Mobile 2014 Program.

6. REFERENCES

[1] Mukherjee A, Liu B, Wang J, Glance N, and Jindal N. Detecting group review spam. WWW '11.

[2] Lim E-P, Nguyen V-A, Jindal N, Liu B, and Lauw H W. Detecting product review spammers using rating behaviors. CIKM '10.

[3] Jindal N, Liu B, and Lim E-P. Finding unusual review patterns using unexpected rules. CIKM '10.

[1] `www.resellerratings.com`