Detecting Group Review Spam

Arjun Mukherjee[†], Bing Liu[†], Junhui Wang[‡], Natalie Glance*, Nitin Jindal[*]
[†] Dept. of Computer Science, University of Illinois at Chicago
arjun4787@gmail.com, liub@cs.uic.edu
[‡]Dept. of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago
jwang@math.uic.edu
[*]Google Inc
nglance@google.com, nitin.jindal@gmail.com

ABSTRACT

It is well-known that many online reviews are not written by genuine users of products, but by spammers who write *fake reviews* to promote or demote some target products. Although some existing works have been done to detect fake reviews and individual spammers, to our knowledge, no work has been done on detecting spammer groups. This paper focuses on this task and proposes an effective technique to detect such groups.

Categories and Subject Descriptors

H.2.8 [Database Applications]: Data Mining

General Terms

Algorithms, Experimentation

Keywords

Review spam, spammer group detection, adversarial data mining.

1. INTRODUCTION

Opinion spamming refers to writing *fake reviews* that try to deliberately mislead human readers or automated opinion mining systems by giving undeserving positive opinions or unjust or false negative opinions to promote or demote some target products. The problem can be seen as a classification problem with two classes, spam and non-spam. However, to obtain training data for model building by manually labeling reviews is very hard, if not impossible, as a spammer can easily craft a fake review that is just like an genuine review [2]. Duplicate reviews were used in [2] as spam reviews for model building. However, many non-duplicate reviews can be spam too. Due to the labeling problem, studies have been made to find reviewers who behave in suspicious ways [3, 5]. For example, if a reviewer wrote all negative reviews about products of a brand but wrote all positive reviews about a competing brand, this reviewer is clearly a spam suspect.

In this work, we focus on *group spam*, which has not been studied so far. A *spammer group* refers to a group of reviewers who works together writing fake reviews to promote or demote a set of target products. Spammer groups are very damaging due to their sheer sizes. When a group is working collaboratively towards a product, it can take control of the sentiment for the product.

This paper proposes a method to detect such groups, which consists of pattern mining to find candidate groups, assessing them using criteria that indicate atypical behaviors of groups, and finally ranking the candidate groups. Our experiment is based on a large set of Amazon reviewers and their reviews. The user study shows that the proposed method is highly effective.

2. THE PROPOSED TECHNIQUE

If a group of reviewers who only worked together once to promote or to demote a single product, it can be hard to detect them. However, fake reviewers (especially those who get paid to write) cannot be just writing one review for a single product because they would not make enough money that way. Instead, they work on many products, i.e., write reviews for many products, which unfortunately gives them away. Frequent pattern mining can be used to find them working together on multiple products. Our proposed method works in three steps:

Step 1 - Frequent Pattern Mining to Find Candidate Groups: In this step, we extract the review data to produce a set of transactions. Each transaction represents a unique product and consists of all reviewers (their ids) who have reviewed that product. Using all the transactions, we can perform frequent pattern mining [1]. The resulting patterns (also called frequent itemsets) are candidate spammer groups.

Step 2 - Computing Spam Indicator Values: Many of the candidate groups may not be true spammer groups. This step tries to evaluate them based on a set of unusual behaviors to find out whether these groups behave strangely. We have designed 8 criteria. Due to space limitations, we are unable to give the computational details. Interested readers please refer to [7].

Time Window (TW): Reviewers in a spammer group are likely to work together to post fake reviews for a target product in a short time interval.

Group Deviation (GD): When members of a group work together to spam, they generally give either very high or very low ratings to the products. The same products typically are also reviewed by other genuine (non-spam) reviewers. Group spammers generally deviate in their ratings by a significant amount from the general review ratings that the product receives from other reviewers. So, the bigger the deviation the worse the group is.

Group Content Similarity (GCS): Group spammers may even know one another and copy reviews among themselves. So, the products which are victims of such group spamming can have many reviews with similar content.

Member Content Similarity (MCS): The members of a group may not know one another. Each of them just copy or modify his/her

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

own previous reviews. If multiple members of the group do this, the group is more likely to be a spammer group.

Early Time Frame (ETF): One damaging group spam activity is to strike right after a product is launched or is made available for reviewing. The purpose is to make a big impact and to take control of the sentiment on the product.

Ratio of Group Size (RGS): The ratio of the group size and the total number of reviewers for the product is also a good indicator of spamming. In one extreme (the worst case), the group members are the only reviewers of the product, which is very damaging.

Group Size (GS): The group size itself also tells something quite interesting. If a group is large, then the probability of members happening to be in the group by chance is small. Furthermore, the larger the group, the more devastating is its effect.

Support count (SC): Support count is the number of products for which the group has worked on together. If a group has a very high support count, it is clearly alarming.

Step 2 - Ranking Using SVM Rank: This step ranks the discovered candidate groups based on how likely they are true spammer groups using the above indicators or features. There are two options. First, we can design a custom formula to combine the feature/indicator values. This will need substantial trial and error. The second approach is to use learning to rank. This requires manually ranked examples as the training data. We took the second approach and used SVM rank [4] to perform the ranking task. Instead of producing manual rankings as training data, we produce them automatically.

It is easy to imagine that there exist many flavors of group spamming behaviors. We can then design ranking functions based on these flavors. In this work, we use three ranking functions to capture some alarming behaviors. These functions generate some preference rankings of the candidate groups based on their resulting scores. The three functions are as follows (G is a group):

$$h_1(G) : G \rightarrow \boldsymbol{R}^+,\ h_1(G) = \mathrm{GCS}(G) + \mathrm{MCS}(G)$$
$$h_2(G) : G \rightarrow \boldsymbol{R}^+,\ h_2(G) = \mathrm{GS}(G) + \mathrm{SC}(G) + \mathrm{TW}(G)$$
$$h_3(G) : G \rightarrow \boldsymbol{R}^+,\ h_3(G) = \mathrm{RGS}(G) + \mathrm{ETF}(G) + \mathrm{GD}(G)$$

$h_1(\cdot)$ ranks groups based on their content similarity across products and members. $h_2(\cdot)$ ranks groups based on scores obtained by features such as *group size*, *group support* and *time window*. Clearly, a group scoring high in this function is suspicious. $h_3(\cdot)$ captures the groups that review products when the products are just being launched in order to make a big impact. The ranking results of these three functions are then used by SVM^{rank} [4] to learn and produce the final single ranking of the candidate groups.

3. EXPERIMENTAL EVALUATION

Our experiment is conducted using a large number of reviewers and reviews of manufactured products from Amazon.com [2]. A user study is used to verify whether the ranking produced by our algorithm confirms to people's perceptions of spammer groups.

Frequent pattern mining and ranking: The number of candidate groups mined in step 1 was 2,273 with the minimum support of 3. Each group consists of at least two reviewers. SVM rank was then applied to produce the final ranking of the candidate spammer groups. This ranked result was employed in our user study.

User agreement study: This was conducted using three (3) independent human judges (raters). The judges were briefed with all individual indicators (or features) to make sure that they fully understand the task and the meaning of each indicator value. Due to a large number of candidate groups, it would have taken too much time for human judges to assess them all. We thus selected the following three types of groups for the user agreement study: top 100 groups, middle 100 groups and bottom 100 groups.

Table 1: Numbers of detected spam groups by three judges

	No. of Spam groups, J-1	No. of Spam groups, J-2	No. of Spam groups, J-3	Avg.
Top 100	100	98	94	97.3
Middle 100	19	12	2	11.0
Bottom 100	0	0	0	0.0

Table 2: Cohen's Kappa for pairwise inter-rater agreements

	Kappa (J-1, J-2)	Kappa (J-1, J-3)	Kappa (J-2, J-3)	Avg.
Middle 100	0.806	0.900	0.829	0.845
Total 300	0.918	0.932	0.912	0.921

User agreement results: The detailed user evaluation results are given in Table 1. In the table, J-1, J-2 and J-3 are the 3 judges. We can observe that for the top 100 ranked groups, almost all of them are considered as spam by all three judges. For the bottom 100 ranked groups, they are all considered as non-spam by the judges. For the middle 100 groups, their decisions vary, which is reasonable because the middle groups are much harder to judge. These results clearly show our ranking is effective and they reflect people's perceptions of spam and non-spam

Table 2 reports the level of user (inter-rater) agreement based on Cohen's Kappa. Since Kappa is not defined when some judges gave exactly the same values to all groups, it is thus not computed for the top 100 groups and the bottom 100 groups separately as one or more judges labeled them either all spam or all non-spam. From Table 2, we can see that the Kappa scores are all above 0.8 which indicates almost perfect agreements [6].

4. CONCLUSIONS

As individuals and businesses are increasingly using reviews for their decisions making, it is critical to detect spammers who write fake reviews. This paper proposed an effective technique to detect spammer groups who work together to write fake reviews. Our user-agreement study showed that the technique is promising.

5. Acknowledgement

This project was funded by a Google Faculty Research Award.

6. REFERENCES

[1] Agrawal, R. and Srikant, R. Fast algorithms for mining association rules. *VLDB*, 1994.

[2] Jindal, N, Liu, B, Opinion spam and analysis. *WSDM*, 2008.

[3] Jindal, N., Liu, B. and Lim, E.P. Finding unusual review patterns using unexpected rules. *CIKM*, 2010.

[4] Joachims, T. Optimizing search engines using clickthrough data, *KDD*, 2002.

[5] Lim, E.P., Nguyen, V.A., Jindal, N., Liu, B., and Lauw, H. Detecting product review spammers using rating behavior. *CIKM*, 2010.

[6] Landis, J.R. and Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33.

[7] Mukherjee, A., Liu, B, Wang, J., Glance, N, Jindal, N. Detecting Group Review Spam. Dept of CS. Technical Report, UIC, 2011.