Detection of Spammer with Review Behaviors Based Scoring Method

Chan Myae Aye
Universities of Computer Studies, Yangon.
cmaye84@gmail.com

Abstract

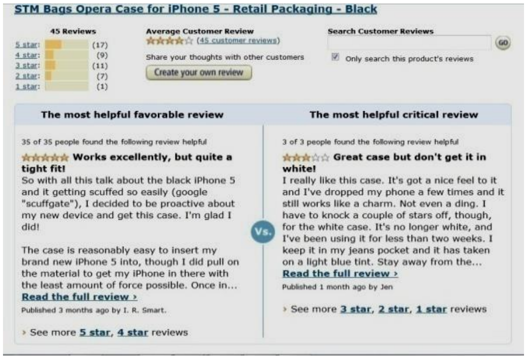
It is now quite common for online user to write reviews on websites and these reviews are read by customer before deciding to purchase a product. Trustworthiness of reviews is now a challenging research problem. There is not many published study on this topic although web spam and email spam has been investigated extensively. Spammer detection techniques that define spam score based on review behaviors of the reviewer are presented in this paper. The experiment showed that the presented technique has comparatively effective spammer detection than other techniques.

Key words: Spam detection, review behaviors, scoring method

1. Introduction

The Web has dramatically changed the way that people express themselves and interact with others. They can now post reviews of products at merchant sites (e.g., amazon.com) and express their views in blogs and forums. It is now well recognized that such *user generated contents* on the Web provide valuable information that can be exploited for many applications. [15] In Web 2.0 application, the user contributed comments offer the promise of a rich source of contextual information about Social Web content. Due to the fact that the quality is not control, anyone can write anything on the Web. This results in many low quality reviews, and worse still *review spam*. This spam review can mislead reader and detection of this spam is now a challenging research problem.

Typically, the reviews consist of an overall product score (often in the form of a star-rating) and some free-form review text to allow the reviewer to describe their experience with the product or service in question. Web user can post products reviews at merchant sites to express their views and interact with other users via blogs and forums. Reviewer gives review and also star rating on the product. Figure1 shows the most helpful favorable and the most helpful critical reviews on Amazon web site. It is now well recognized that the user generated content contains valuable information that can be exploited for many applications [3, 14].



STM Bags Opera Case for iPhone 5 - Retail Packaging - Black

45 Reviews	
5 star:	(17)
4 star:	(9)
3 star:	(11)
2 star:	(7)
1 star:	(1)

Average Customer Review
★★★★☆ (45 customer reviews)

Share your thoughts with other customers

Search Customer Reviews

☑ Only search this product's reviews

The most helpful favorable review

35 of 35 people found the following review helpful

★★★★★ **Works excellently, but quite a tight fit!**

So with all this talk about the black iPhone 5 and it getting scuffed so easily (google "scuffgate"), I decided to be proactive about my new device and get this case. I'm glad I did!

The case is reasonably easy to insert my brand new iPhone 5 into, though I did pull on the material to get my iPhone in there with the least amount of force possible. Once in...

Read the full review ›

Published 3 months ago by I. R. Smart.

› See more 5 star, 4 star reviews

The most helpful critical review

3 of 3 people found the following review helpful

★★★☆☆ **Great case but don't get it in white!**

I really like this case. It's got a nice feel to it and I've dropped my phone a few times and it still works like a charm. Not even a ding. I have to knock a couple of stars off, though, for the white case. It's no longer white, and I've been using it for less than two weeks. I keep it in my jeans pocket and it has taken on a light blue tint. Stay away from the...

Read the full review ›

Published 1 month ago by Jen

› See more 3 star, 2 star, 1 star reviews

Figure 1. Example of reviews on Amazon.com

The existing work has been mainly focused on extracting and summarizing opinions from reviews using natural language processing and data mining techniques [1, 4, 12, 14 and 17]. In the context of Web search, due to the economic and/or publicity value of the rank position of a page returned by a search engine, Web page spam is widespread. [5, 6, 7, 18, 20, 21] Web page spam refers to the use of “illegitimate means” to boost the rank positions of some target pages in search engines [2, 21]. In the context of reviews, the problem is similar, but also quite different.

Due to the openness of product review sites, spammers can pose as different contributing spammed reviews making them harder to eradicate completely. Spam reviews usually look perfectly normal until one compares them with other reviews of the same products to identify review comments not consistent with the latter. The efforts of additional comparisons by the users make the detection task tedious and non-trivial [11]. One approach taken by review site such as Amazon.com is to allow users to label or vote the reviews as helpful or not. Unfortunately, this still demands user efforts and is subject to abuse by spammers. Spam review is sometime with high helpfulness score so that helpfulness score is not a good indicator of spam.

Most review spam detection system focus on review behaviors and detect with some classification techniques. Language modeling technique and some similarity computation techniques on review text are also proposed to detect spam and this can have time consuming because of deeply analysis on opinion and text understanding. One should focus on detecting spammers based on their spamming behaviors instead of detecting spam reviews only. Subsequently, spam review can be removed to accelerate the interests of other review users.

The rest of the paper is organized as follows. Section 2 covers some related works. Section 3 presents spammer detection techniques with review based spam score methods. Experimental evaluation is described in Section 4 and Section 5 is devoted to conclusions.

2. Related Work

Analysis on online opinion becomes a popular research topic recently. Most research trends focus on opinion mining and opinion extraction. A preliminary study of opinion spam was reported in [14].

A spam activities analyzing and spam detection methods are presented in [16]. Three types of spam review such as untruthful opinion, review on brand only and non-review (e.g question and answer and random texts) are also discussed. From this analysis, spammer detection can use some features to detect behaviors of the spammer. For example, the analysis said that reviews with negative deviation (in rating) on the same brands give the highest lift curve (that is spam), only reviews are very likely to be candidates of spam [11].

The scoring methods to measure the degree of spam for each reviewer is presented in [11] and applies on an Amazon review dataset. Scoring method for spammer on targeting product and product groups is proposed and they focus on scoring methods of target based spamming and deviation based spamming score methods. Then highly suspicious reviewers is selected for further scrutiny by user evaluators with the help of a web based spammer evaluation software specially developed for user evaluation experiments. The results show that the ranking and supervised methods are effective in discovering spammers and outperform the baseline method based on helpfulness votes alone. To assign an overall numeric spam score to each user, the spam scores of the user's different spamming behaviors are combined by using linear weighted combination. The weights on the different component spam scores can be empirically defined or learnt automatically. Only rating deviation and early rating deviation was not much effective.

Rating behaviors are considered more detail to detect spammer. [9] This technique has comparatively effective spammer detection than other techniques. Deeply investigation is made in rating behaviors of reviewer for more efficient spamming scores. The presented technique based on rating score methods and review behaviors based score method is added for more accurate spammer detection.

A language modeling approach for consumer review spam detection is presented in [9]. They showed that Kullback-Leibler (KL) Divergence and the probabilistic language modeling based computational model. This model is effective for the detection of untruthful reviews to estimate the similarity between any pairs of reviews in terms of the likelihood of a review “generating” the contents of another review. Moreover, the

Support Vector Machine also called SVM -based method is also effective for the detection of non-reviews. The empirical study found that around 2% of the consumer reviews posted to a large e-Commerce Web site is spam. The computational model proposed spam detection based on review similarity and other behaviors of rating has not taken into account.

Review similarity is an important factor to detect spammer because most spammers are lazy so that they just copy and paste their review instead of writing new one. In this paper, review similarity spam score by using shingle method is proposed to detect spammer instead of deeply analysis on opinion mining. Other review behaviors such as posting date, the only review and the first review of the reviewer are also considered for effective detection. Spamming behaviors are complicated and not easily captured. So that, many researches about review spammer detection are required for improving web sites.

3. Review Behaviors based Spammer Detection

In this work, spammer detection is based on the review behaviors of the reviewer given to the product. The spamming score of the reviewer is calculated by scoring techniques. Table 1 describes notation and their descriptions used in this work.

The spam score methods based on the review features are presented in the next subsections.

3.1 Review-based Spamming Score

Spam reviews usually look perfectly normal until one compares them with other reviews of the same products. Most spammers are likely to spam the product with multiple review texts and these review texts are likely to be identical or look similar so as to conserve spamming efforts. So that it is importance to look at how reviews are equal to another review of the same user. Other review behaviors are also considered to detect spammer. The more spamming behaviors the system can detect for a reviewer, the more likely the reviewer is spammer. The next subsections described about spammer detection techniques based on review behaviors.

Table 1. Notations and descriptions

Notation	Description
U	$\{u_i\}$: set of users
O	$\{o_j\}$: set of products
V	{v}:set of review v
o(v)	product of review v
V_{i*}	$\{ v_k \backslash u(v_k)=u_i \ \wedge o(v_k)=o_j\}$: set of reviews on any products by user u_i

3.1.1 Review Similarity Spam Score

As described above, a user spam a product with multiple review texts and such reviews texts are likely to be identical or look similar to conserve their spamming efforts. Cosine similarity with bag of words is used in similarity computation of review but TFIDF of cosine has less similarity for rare terms. [10] So, the system use shingle method to measure reviews similarity. Single method is a good similarity computation method to detect similar reviews of reviewer.

The following equation is to calculate similarity between two reviews posted by the reviewer.

$$sim(v_k, v_k^{'}) = \frac{s(v_k,w) \cap s(v_k^{'},w)}{s(v_k,w) \cup s(v_k^{'},w)} \qquad (1)$$

$s(v_k, w)$, $s(v_k^{'}, w)$ is contiguous subsequences of tokens in review $v_k, v_k^{'}$ and w denotes the number of tokens. The similarity is a number in the range [0,1], where 1 indicates that two reviews are identical. The similarity score of all reviews given by the user u_i, can be calculated by the following equation:

$$sim(v) = \underset{v_k,v_k^{'} \in V_{i*}}{avg} sim(v_k, v_k^{'}) \qquad (2)$$

And then all the similarity scores of reviews are taken average for similarity spam score of user u_i.

$$SV_S(u_i) = \underset{v \in V_{i*}}{Avg} Sim(v) \qquad (3)$$

If similarity scores are high, the spamming scores are also high.

3.1.2 Other Review Spam Score

Reviews which are written early tend to get more user attention, and thus can have bigger impact on the sale of a product. The first reviews or the only reviews are more likely to be spam than other reviews. The position of the review date is also an important factor because early reviews are more concentrated by users. So the system considers on these factors for other review spam score by the following equation:

$$SV_B(v)= \frac{1}{3}\ [pos(v) + fpos(v) + only(v)] \qquad (4)$$

where, $SV_B(v)$ is other review spam score user u_i and pos(v), fpos(v) and only(v) can be obtained by the following equarion:

$$pos(v) \ = \begin{cases} 1, if\ review\ v\ is\ the\ first\ five\ review \\ 0, else \end{cases} \qquad (5)$$

$$fpos(v) = \begin{cases} 1, if\ review\ v\ is\ the\ first\ review \\ 0, else \end{cases} \qquad (6)$$

$$only(v)= \begin{cases} 1, if\ review\ v\ is\ the\ only\ review \\ 0, else \end{cases} \qquad (7)$$

3.1.3 Combined Review Spam Score

Combined review spam score is the spam score of user to the products by all of their review spamming behaviors. It achieves effective spam score by simply giving some weight of each score function. In this case, review similarity score get more weight because the similar review behavior is more likely to be spammer than other behaviors of the reviewer. The required review based spam score of user u_i is defined by:

$$SV(u_i)= \frac{2}{3}\ SV_S(u_i) + \frac{1}{3} SV_B(u_i) \qquad (8)$$

Where $SV(u_i)$ is review spam score of user u_i.

3.2 Combined Spam Score

Finally the system needs to compute combined spamming score of review and rating. The rating based spam score $SR(u_i)$ is achieved by the following equation:

$$SR(u_i) = \frac{1}{2}\ SR_S(u_i) + \frac{1}{4} SR_D(u_i) + \ \frac{1}{4}\ SR_B(u_i) \ (9)$$

Where SR_S is the rating similarity spam score, SR_D is the rating deviation spam score , SR_B is other rating behaviors spam score and all this scoring methods are presented in previous work. [9]

To achieve effective spammer detection system, it is needed to add the review spam score with rating spam score. In this work, all the respective behaviors of rating and review are considered and calculated by the following equation:

$$CS(u_i) = \frac{1}{2}[SV(u_i) + SR(u_i)] \qquad (10)$$

where, $CS(u_i)$ is spam score of user u_i.

4. Evaluation

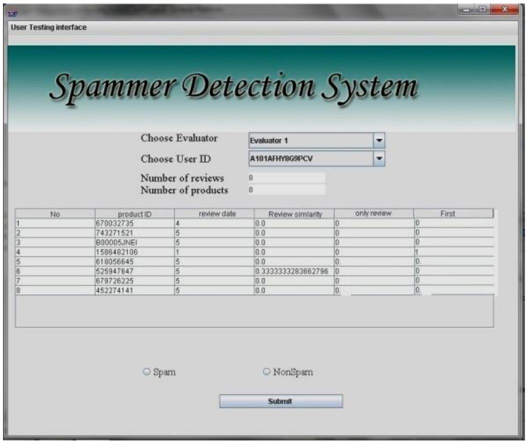
The presented methods are evaluated by using data from amazon.com. The reason for using this data set (http://131.193.40.52/data) is that it is large and covers a very wide range of products. Amazon.com is considered one of the most successful e-commerce Web sites with a relatively long history. This dataset gives the information such as Product ID, Reviewer ID, Rating, Date, Review Title, Review Body, Number of Helpful Feedbacks and Number of feedbacks. The statistics of the dataset are presented in the following table 2.

Spam score techniques of (i) review similarity $SV_S(u_i)$ (ii) other review behaviors $SV_B(v)$ (iii) combined review spam scores $SV(u_i)$ (iv) combined spam score $CS(u_i)$ are evaluated in this section.

Table 2. Dataset statistics

	Number
U	11,038
O	5,693
$\|V_{i*}\|$	48,894

There is still no standard dataset to test the accuracy and this is why detection of review spam has been neglected so far. So the system uses three human evaluators to check whether a reviewer is spammer or not by using spammer detection system that are developed and presented in figure 2. The evaluator can look at the feature of each reviewer's rating behaviors and can give feedback whether the reviewer is spammer or not.



Choose Evaluator: Evaluator 1

Choose User ID: A101AFHY8G9PCV

Number of reviews: 8

Number of products: 8

No	product ID	review date	Review similarity	only review	First
1	670032735	4	0.0	0	0
2	743271521	5	0.0	0	0
3	B00005JNEI	5	0.0	0	0
4	1586482106	1	0.0	0	1
5	618056645	5	0.0	0.	0.
6	525947647	5	0.3333333283662796	0	0
7	679726225	5	0.0	0	0
8	452274141	5	0.0	0.	0.

○ Spam ○ NonSpam

Figure 2. Spammer Detection System

4.1 Evaluation Setup

There are several challenges in conducting the user evaluation experiments. There are many reviewers and it is impossible for the human evaluators to judge everyone. Only small subsets of reviewers are evaluated to handle this issue. For each spammer detection method, select 10 top ranked reviewers and 10 bottom ranked reviewers. And then merge all the selected spammers into a pool which consists of 62 reviewers. These reviewers are then sorted by their combined spamming behavior scores. 25 top ranked reviewers and 25 bottom ranked reviewers are then selected for user evaluation. This number of reviewers is quite reasonable for a human evaluator to examine. The system further randomly order the reviewers so that there is no relationship between reviewers' order and their spammer scores

For each reviewer, select his/her ratings to be highlighted for human evaluator's attention. These selected ratings must be seen by the human evaluator before the latter makes judgment on whether the reviewer is a spammer. The reviews are selected based on their involvement in the spamming behaviors identified. Specifically, identical (or similar) review with other reviews, first review or not, position of review date and the only review or not.

Three human evaluators are recruited to examine the selected reviewers and ratings using the spammer detection system. For each reviewer, the labeled decision is either spammer" and non-spammer" and the evaluators are not informed about the number of spammers to be labeled.

4.2 Results

Given the results of the three evaluators, final label to each reviewer is assigned by using majority voting. A reviewer is assigned a final spam or non-spam label if the label is agreed by two or more evaluators. The number of top 10 reviewers with the final spammer labels, and the number of bottom 10 reviewers with non-spammer labels for different methods are shown in Table 3. Rating based spam score is a base line method and the presented techniques are significantly better than base line method.

Table 3: Results of Top 10 and Bottom 10 Ranked Reviewers

	Spam score methods			
	Rating based spam score (Base line)	**Review Similarity spam score**	**Review based spam score**	**Combined score(rating and review)**
#Spammer in top 10	9	7	9	9
#nonspammer in bottom 10	8	10	9	10

Table 4. Confusion matrix for the definition of the Effectiveness Measure

	Gold Standard – Human Classification		
		Spam	**Ham**
System's Classification	**Spam**	**a**	**b**
	Ham	**c**	**d**

Text Retrieval Conferences (TREC) spam track (Tormack and Lynam. 2005; Cormack 2007) is used to measure the performance of the presented techniques. They were widely used to evaluate other kinds Web spam. With reference to a confusion matrix depicted in table [4], the various effectiveness measures can be defined by:

$$hm= \frac{b}{b+d} \quad (11)$$

$$sm= \frac{c}{a+c} \quad (12)$$

$$lam= logit^{-1}\left(\frac{logit\ (hm)+logit(sm)}{2}\right) \quad (13)$$

$$tp= \frac{a}{a+C} \quad (14)$$

where a, b, c, and d refer to the number of reviews falling into each category presented in table [4]. The ham misclassification rate (*hm*) is the fraction of all ham misclassified as spam; the spam misclassification rate (*sm*) is the fraction of all spam misclassified as ham. It is desirable to have a single measure which combines both of the above measures. Therefore, the TREC Spam track also made use of the logistic average misclassification rate (*lam*) to measure the effectiveness of spam detection systems, where $logit^{-1}(x) = \frac{e^x}{1+e^x}$ and $logit(x) = \ln\left(\frac{x}{1-x}\right)$. Since hm, sm, and lam are the measures for failure rather than effectiveness, the lower scores imply a better detection performance. The true positive rate (*tp*) is the fraction all spam identified by the system. On the other hand, the common effectiveness measure $accuracy = \frac{a+d}{a+b+c+d}$ may be measured for spam detection System. The accuracy figure is report in table [5]. These values are calculated from results of the user evaluation tests on 50 reviewers. The presented techniques have more accuracy than base line method.

Table 5. Comparative performance of the spammer detection techniques

	Spam score methods			
	Rating based spam score (Base line)	**Review similarity spam score**	**Review based spam score**	**Combined score(rating and review)**
tp%	0.90	0.70	0.90	0.90
hm%	0.20	0.00	0.10	0.00
sm%	0.10	0.30	0.10	0.10
lam%	0.14	0.39	0.10	0.25
Accuracy	0.85	0.85	0.90	0.95

5. Conclusion and Future Work

This paper presents review spammer detection techniques based on review behaviors on similarity, other review behaviors and combination of these. This score is also combine with rating based spam score and this method is better than baseline method. The presented techniques focus on scoring methods and has not considered about deeply understanding of review text. So it is needed to add some machine learning techniques to achieve more effective spammer detection system.

References

[1] A-M. Popescu and O. Etzioni. Extracting Product Features and Opinions from Reviews. *EMNLP'2005*.

[2] A. Ntoulas, M. Najork, M. Manasse & D. Fetterly. Detecting Spam Web Pages through Content Analysis. WWW'2006.

[3] B. Liu. Web Data Mining: Exploring hyperlinks, contents and usage data. Springer, 2007.

[4] B. Pang, L. Lee & S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. *EMNLP'2002*.

[5] B. Wu and B. D. Davison. Identifying link farm spam pages. *WWW'06,* 2006.

[6] B. Wu, V. Goel & B. D. Davison. Topical TrustRank: using topicality to combat Web spam. *WWW'2006*.

[7] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, S. Vigna. A reference collection for web spam, *SIGIR Forum'06,* 2006.

[8] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, and L. Lee. How opinions are received by online communities: a case study on amazon.com helpfulness votes. In WWW, 2009.

[9] Chan Myae aye. Spammer Detection by using Rating behaviors. 10th International Conference on Computer Application, 2012.

[10] C.L. Lai, K.Q. Xu, Raymond Y.K. Lau ,Y. Li and L. Jing, A language modeling approach for consumer review spam detection, IEEE International Conference on E-Business Engineering , 2010.

[11] L . Ee-Peng, V. Nguyen, N. Jindal, B. Liu, H. W. Lauw. Detecting Product Review Spammers using Rating Behaviors, CIKM'10, Toronto, Ontario, Canada, October 26–30, 2010.

[12] K. Dave, S. Lawrence & D. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *WWW'2003*.

[13] B. Liu, Web Data Mining. Springer, 2007.

[14] M. Hu & B. Liu. Mining and summarizing customer reviews. KDD'2004.

[15] N. Jindal and B. Liu. Review spam detection. In WWW (poster), 2007.

[16] N. Jindal and B. Liu. Opinion Spam and Analysis. WSDM'08, Palo Alto, California, USA, February 11-12, 2008.

[17] P.Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *ACL'2002*.

[18] R. Baeza-Yates, C. Castillo & V. Lopez. PageRank increase under different collusion topologies. *AIRWeb*'*05*,

[19] K. Soo-Min, P. Pantel, T. Chklovski, M. Pennacchiotti. Automatically Assessing Review Helpfulness. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pages 423–430, Sydney, July 2006.

[20] Y. Wang, M. Ma, Y. Niu, H. Chen. Spam Double-Funnel: Connecting Web Spammers with Advertisers. *WWW2007.*

[21] Z. Gyongyi & H. Garcia-Molina. *Web Spam Taxonomy*. Technical Report, Stanford University, 2004.