Distortion as a Validation Criterion in the Identification of Suspicious Reviews

Guangyu Wu, Derek Greene, Barry Smyth, Pádraig Cunningham
School of Computer Science and Informatics
University College Dublin, Ireland
{guangyu.wu,derek.greene,barry.smyth,padraig.cunningham}@ucd.ie

ABSTRACT

Assessing the trustworthiness of reviews is a key issue for the maintainers of opinion sites such as TripAdvisor. In this paper we propose a distortion criterion for assessing the impact of methods for uncovering suspicious hotel reviews in TripAdvisor. The principle is that dishonest reviews will distort the overall popularity ranking for a collection of hotels. Thus a mechanism that deletes dishonest reviews will distort the popularity ranking significantly, when compared with the removal of a similar set of reviews at random. This distortion can be quantified by comparing popularity rankings before and after deletion, using rank correlation. We present an evaluation of this strategy in the assessment of shill detection mechanisms on a dataset of hotel reviews collected from TripAdvisor.

Categories and Subject Descriptors

E.0 [**Data**]: General – Data quality; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Evaluation, Algorithms

Keywords

User-generated content, Credibility, Shilling

1. INTRODUCTION

Perhaps the greatest commercial success derived from user-generated content is the use of reviews and recommendations on sites such as Amazon and TripAdvisor [8, 12]. It is recognized that the fact that Amazon has a more extensive collection of user-generated reviews and recommendations than its competitors confers a significant sales advantage [12]. However, this reliance on user-generated content comes at a price. Tripadvisor claims to be the largest site for "unbiased travel reviews" on the internet [8] and if this unbiased claim is brought into question then it can be very damaging for them [6].

This vulnerability of recommender systems to 'shilling' attacks is widely recognized, and there is already an extensive literature on identifying such attacks and on making systems robust to malicious influence [9, 14]. While much of this work has addressed automatic collaborative filtering (ACF) systems (*e.g.* [9]), in the work described here we focus on identifying bogus reviews and ratings that are not necessarily being used in an ACF framework.

In this paper we explore the conjecture that shill reviews are likely to distort popularity rankings given that the objective is to improve the online reputation of a hotel. For instance, the Four Seasons Hotel in Las Vegas is ranked second of 286 hotels in Las Vegas based on 446 reviews. It would be difficult to influence this ranking because of the volume of reviews and ratings available, making it an unlikely target for shilling.

A major challenge for research in this area is the lack of annotated datasets for assessing the effectiveness of shill detection strategies. For this reason, we have gathered a dataset of approximately 30,000 TripAdvisor reviews covering Irish hotels, which we used in our evaluation. This evaluation assess the distortion impact of a number of review deletion policies and suggests that distortion is effective for separating true positives from false positives.

The paper proceeds as follows. In the next section we provide a brief overview of relevant research and in Section 3 we present the basic shill detection strategies that we use in our evaluation. In Section 4 we introduce the idea of using distortion as a principle for validating shill detection strategies, and in Section 5 we present an evaluation of this on the Irish TripAdvisor data. The paper concludes with a summary and some suggestions for future work.

2. RELATED WORK

We are concerned with the situation where an agent, in collusion with the seller of an asset or service, heaps praise on mediocre offerings. This practice, which has existed in the real world for centuries, has found its way into online opinion and recommendation sites [9, 14]. The proliferation of such practices can lead us to question whether the gap in quality and unreliability between user-generated content and expert editorial opinion could render the former valueless [1].

If we consider the identification of spam reviews as a subset of the larger problem of identifying reviews that are authoritative, credible or helpful, then there is some interesting research to draw on. Both O'Mahony & Smyth [11] and Hsu

et al. [5] cast the problem of ranking reviews in a supervised learning framework, and show impressive results. O'Mahony & Smyth use customer feedback on the helpfulness of reviews on Amazon to provide the *supervision*, while Hsu et al. use feedback provided from Digg. Unfortunately in the TripAdvisor scenario there is no user feedback to support a supervised learning approach.

There are many related or analogous problems that have received attention – in particular e-mail spam [3], link spam (search engine spam) [2], detecting attacks on recommender systems [10, 4], and assessing authoritativeness on sites such as Wikipedia [7].

The idea of using *distortion* to identify anomalous behavior is not new. For instance this general principle has been used to reveal link spam [2] and to identify untrustworthy participants in peer-to-peer search networks [13].

3. SHILL DETECTION

The focus of this paper is on distortion as a validation criterion in shill detection rather than on features that are predictive of shills so the features we employ are quite basic. The two features we consider are based on the idea of positive singletons as shown in Figure 1. Positive singletons are positive reviews from reviewers who have posted no other reviews.

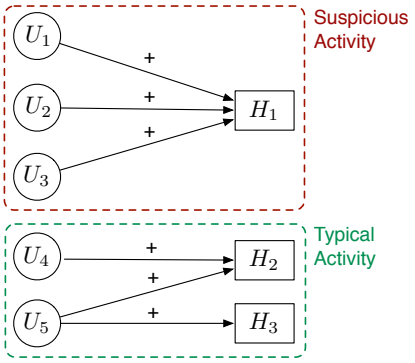

Figure 1: Bipartite graph representing a simple scenario involving five users and three hotels.

Proportion of Positive Singletons (PPS): The PPS score for hotel H is the proportion of reviews on that hotel that are positive singletons:

$$PPS(H) = \frac{N_{ps}}{N} \tag{1}$$

where N_{ps} is the number of positive singleton reviews, and N is the total review count for the hotel.

Concentration of Positive Singletons (CPS): Often multiple shill reviews will be injected in quick succession.[1] The greater the degree of temporal clustering between positive singletons, the more suspicious these reviews appear.

Given the list of of positive singleton reviews $\{r_1, \ldots, r_P\}$ for a hotel H arranged in ascending order by submission date, we define a score for H as a function of the average date distance D (*i.e.* number of days) between each review r_i and its temporally nearest neighbor:

$$CPS(H) = \frac{1}{P}\sum_{i=1}^{P} e^{-\lambda \times min(D(r_i, r_{i-1}), D(r_i, r_{i+1}))} \tag{2}$$

where λ is a bandwidth parameter that controls the influence of the proximity of reviews. We found that a value of $\lambda = 1$ was most effective on the TripAdvisor data.

[1]The review spam recently discovered on Apple's App Store had this characteristic `http://edition.cnn.com/2009/TECH/12/09/wired.apple.apps/index.html`

4. VALIDATION USING DISTORTION

Our proposal for using distortion to validate the filtering of suspicious reviews is based on the prominence given to user-based popularity rankings on many e-commerce sites. For instance, TripAdvisor assigns a ranking to each hotel in a given region (*e.g.* 2^{nd} of 446 hotels in Las Vegas). Our contention is that a common objective of shilling will be to influence this ranking. Deleting a set of reviews chosen at random should not overly disrupt the ranked list of hotels, while deleting shill reviews should significantly alter or distort the ranking of hotels to reveal the "true" ranking.

It is important to state that TripAdvisor do not disclose the details of their ranking algorithm. However, it is clear that the main component is the average reviewer rating, as their ranked lists are strongly correlated with lists ordered simply based on average rating. Since we can recalculate the average reviewer rating after review deletion, we use this to produce the popularity ranking used in our experiments. We first calculate a raw distortion score resulting from the deletion of suspect reviews. We subsequently calculate an adjusted distortion score which takes account of the impact of deleting a similar number of reviews chosen at random.

Raw Distortion: The raw distortion score simply quantifies the change in popularity ranking resulting from deleting a number of suspicious reviews. It is calculated as the rank correlation between the original popularity ranking and the popularity ranking after the suspicious reviews have been deleted. More formally, if P is the original popularity ranking where P_i is the rank of the i^{th} hotel and S is the ranking after deleting shills, then the raw distortion after deleting suspected shills is:

$$RD = SRC(P,S) = \frac{\sum_i (P_i - \bar{P})(S_i - \bar{S})}{\sqrt{\sum_i (P_i - \bar{P})^2 \sum_i (S_i - \bar{S})^2}} \tag{3}$$

where $SRC(P,S)$ is the Spearman rank correlation of the two rankings and $\bar{P}$ is the average rank in P. Lower values indicate a higher level of distortion.

Adjusted Distortion: To allow comparisons across hotels where different numbers of reviews may be deleted, it is useful to adjust the raw distortion score to account for this. This is done by assessing the impact of deleting a similar number of positive reviews from a hotel with a similar number of overall reviews. The adjusted distortion score is the difference between this expected distortion score and the raw distortion score. Significant adjusted distortion scores will be positive and insignificant scores will be close to zero. This signifies that there is no difference between deleting the suspected reviews and simply deleting reviews at random. The adjusted distortion score AD for S, which incorporates an expected distortion of ED based on a ranking R after random deletions, is given by:

$$AD = ED - RD = \overline{SRC(P,R)} - SRC(P,S) \tag{4}$$

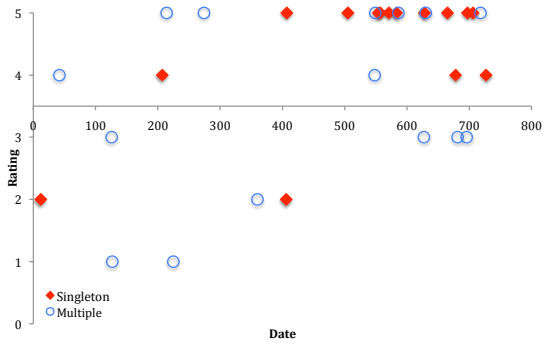

Figure 2: A time plot of the reviews for a typical hotel covering a two year period (730 days).

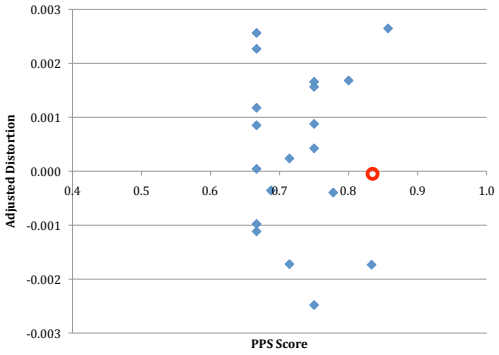

Figure 3: The top 20 hotels as ranked by the PPS score. The chart plots the PPS score against the adjusted distortion. The corresponding time-plot for the hotel marked by the circle is shown in Figure 4.

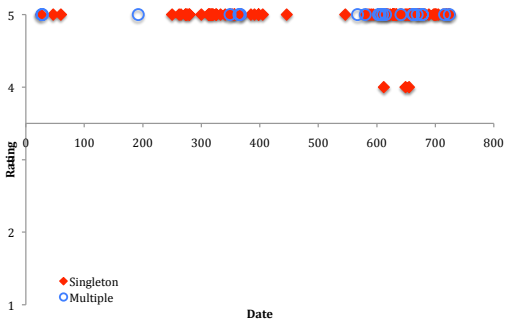

Figure 4: The time-plot of the hotel marked with the circle in Figure 3.

In practice, the expected distortion score ED is calculated by repeatedly choosing hotels at random of a similar size, removing positive reviews, and calculating the raw distortion – the expected is score is given by the average raw distortion over many runs.

5. EVALUATION

In our evaluation we explore whether the distortion in popularity rankings is an effective mechanism for validating the output of a shill detection processes. We examine the impact on distortion of review deletion based on the PPS and CPS scores described previously in Section 3.

The Irish TripAdvisor dataset[2] used here comprises 29,799 reviews from 2,1851 unique reviewers, covering hotels from all regions of Ireland over a two-year time window from September 2007 to September 2009. Note that we only consider a subset of 843 hotels which received four or more reviews during this time. Approximately two thirds of the reviews are positive – *i.e.* awarding at least four out of five stars. Of these roughly 30% are positive singletons as defined in section 3.

A time-plot of the reviews for a typical hotel from the TripAdvisor dataset is shown in Figure 2. For this hotel there is a reasonable balance between singleton reviews and reviews from users who have submitted multiple reviews. Other, perhaps more suspicious cases, are shown in the time-plots in Figures 4 and 6.

We have also conducted an evaluation on artificial data that is not presented here for space reasons. This evaluation entails the insertion of artificial shills into the TripAdvisor dataset. The details of this evaluation are available in a longer version of this paper that is available as a technical report [15].

5.1 Evaluation on TripAdvisor Data

The scatter plot in Figure 3 shows the top 20 most suspicious hotels as ranked by the PPS score, with PPS scores plotted against adjusted distortion in the popularity ranking. Half of the top hotels have negligible or negative distortion when the 'suspicious' reviews are deleted, suggesting that these reviews are unlikely to be shills.

An example time-plot for one of these hotels is shown in Figure 4. This hotel has a highly suspicious PPS score because an overwhelming 83% of the positive reviews are singletons (101 of 121 reviews). Remember that the baseline for the whole dataset is that ≈ 30% of positive reviews are singletons. However, it can be seen in Figure 3 that the distortion score for this hotel is close to zero suggesting that this is a *false positive*. This is because all the non-singleton reviews are also positive so deletion of the purported suspicious reviews does not distort the popularity ranking. Furthermore, an inspection of the text of the suspicious reviews suggests that they might be genuine. We speculate that there may be something more innocent than full-scale shilling going on here – perhaps the hotelier is soliciting reviews from satisfied customers?

In Figure 5 we show the scatter plot for the top 20 hotels as ordered by the CPS score. It is interesting to note that there is far less negative distortion in this plot. This is because the CPS score has no bias towards hotels with few reviews. Thus distortion will be positive or close to zero. This contrasts with the situation for the PPS score, which is inclined to select hotels with few reviews and thus can result in significant negative distortion when a large fraction of a small review set is deleted.

The time-plot for the hotel marked with the circle is shown in Figure 6. The reviews producing the high CPS score are the two shown in the top right of the plot. When these are deleted, the average rating goes from 4⋆ to 3.3⋆, resulting in a significant distortion. We feel this is valid as the two positive singletons look suspiciously like a management response to the strongly negative review that immediately precedes

[2]Available at: `http://mlg.ucd.ie/datasets/trip`.

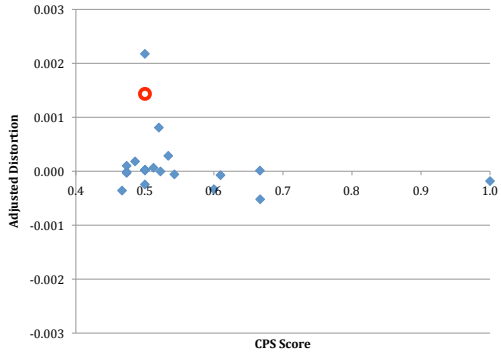

Figure 5: The scatter plot for the top 20 hotels based on the CPS score. The corresponding time-plot for the hotel marked by the circle is shown in Figure 6.

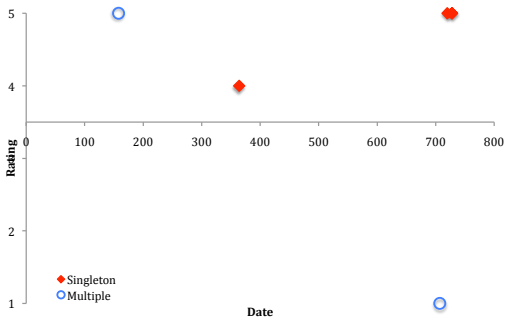

Figure 6: The time-plot of the hotel marked with the circle in Figure 5.

them. This view is supported by an inspection of the text of the reviews.

6. CONCLUSIONS

The objective of the work described in this paper is to explore distortion in popularity ranking as a measure for highlighting shilling. We have presented a preliminary evaluation on real data that supports this. We have used two scores based on the proportion of positive singleton reviews and the concentration of positive singletons to highlight suspicious behavior, and have then shown that distortion helps to separate out true positives (Figures 5 & 6) from false positives (Figures 3 & 4).

Clearly, if distortion is effective for validating other shill scoring mechanisms, then it would make sense to integrate it into a multi-variate shill detection mechanism. The difficulty with integrating the validation mechanism into the detection process is the problem of validating results. We plan to explore this issue in future work.

Acknowledgments: This research was supported by Science Foundation Ireland (SFI) Grant No. 08/SRC/I1407.

7. REFERENCES

[1] R. Baeza-Yates. User Generated Content: How Good Is It? In *3rd Workshop on Information Credibility on the Web (WICOW 2009)*, pages 1–2, 2009.

[2] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates. Using rank propagation and probabilistic counting for link-based spam detection. In *Proc. Workshop on Web Mining and Web Usage Analysis (WebKDD)*, 2006.

[3] P. O. Boykin and V. P. Roychowdhury. Leveraging social networks to fight spam. *IEEE Computer*, 38(4):61–68, 2005.

[4] K. Bryan, M. O'Mahony, and P. Cunningham. Unsupervised retrieval of attack profiles in collaborative recommender systems. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, pages 155–162, 2008.

[5] C. Hsu, E. Khabiri, and J. Caverlee. Ranking Comments on the Social Web. In *Proc. 2009 International Conference on Computational Science and Engineering-Volume 04*, pages 90–97, 2009.

[6] R. Jurca and B. Faltings. Mechanisms for making crowds truthful. *Journal of Artificial Intelligence Research*, 34(1):209–253, 2009.

[7] N. Korfiatis, M. Poulos, and G. Bokos. Evaluating authoritative sources using social networks: an insight from Wikipedia. *Online Information Review*, 30(3):252–262, 2006.

[8] S. Litvin, R. Goldsmith, and B. Pan. Electronic word-of-mouth in hospitality and tourism management. *Tourism management*, 29(3):458–468, 2008.

[9] M. O'Mahony, N. Hurley, N. Kushmerick, and G. Silvestre. Collaborative recommendation: A robustness analysis. *ACM Transactions on Internet Technology (TOIT)*, 4(4):344–377, 2004.

[10] M. P. O'Mahony, N. J. Hurley, and G. C. M. Silvestre. Recommender systems: Attack types and strategies. In M. M. Veloso and S. Kambhampati, editors, *AAAI*, pages 334–339. AAAI Press / The MIT Press, 2005.

[11] M. P. O'Mahony and B. Smyth. Learning to recommend helpful hotel reviews. In L. D. Bergman, A. Tuzhilin, R. D. Burke, A. Felfernig, and L. S-Thieme, editors, *RecSys*, pages 305–308, 2009.

[12] T. O'Reilly. What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. Technical report, Communications & Strategies, No. 1, p. 17, First Quarter 2007. Available at SSRN: http://ssrn.com/abstract=1008839, 2007.

[13] J. Parreira, D. Donato, C. Castillo, and G. Weikum. Computing trusted authority scores in peer-to-peer web search networks. In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, page 80. ACM, 2007.

[14] J. Staddon and R. Chow. Detecting reviewer bias through web-based association mining. In *2nd Workshop on Information Credibility on the Web (WICOW 2008) at ACM CIKM'08*, 2008.

[15] G. Wu, D. Greene, B. Smyth, and P. Cunningham. Distortion as a Validation Criterion in the Identification of Suspicious Reviews. Technical Report UCD-CSI-2010-4, University College Dublin, `http://www.csi.ucd.ie/biblio`, May 2010.