Merging Multiple Criteria to Identify Suspicious Reviews

Guangyu Wu
School of Computer Science & Informatics
University College Dublin
guangyu.wu@ucd.ie

Derek Greene
School of Computer Science & Informatics
University College Dublin
derek.greene@ucd.ie

Pádraig Cunningham
School of Computer Science & Informatics
University College Dublin
padraig.cunningham@ucd.ie

Abstract

Assessing the trustworthiness of reviews is a key issue for the maintainers of opinion sites such as TripAdvisor, given the rewards that can be derived from posting false or biased reviews. In this paper we present a number of criteria that might be indicative of suspicious reviews and evaluate alternative methods for integrating these criteria to produce a unified 'suspiciousness' ranking. The criteria derive from characteristics of the network of reviewers and also from analysis of the content and impact of reviews and ratings. The integration methods that are evaluated are singular value decomposition and the unsupervised hedge algorithm. These alternatives are evaluated in a user study on TripAdvisor reviews, where volunteers were asked to rate the suspiciousness of reviews that have been highlighted by the criteria.

1 Introduction

A significant challenge in the administration of review and rating sites such as Amazon and TripAdvisor is the management and presentation of the user-generated content. For popular items, there will often be many more reviews available than a user might read, so identifying helpful or informative reviews is an interesting challenge [6]. At the other end of the spectrum, it is vital for the credibility of rating sites to identify and filter fraudulent "shill" reviews – that is the subject of this paper.

In the work described here, we present a number of criteria that might be indicative of hotels with suspicious reviews (see Section 3). This is not the central contribution of the paper however. The main contribution is the comparison of methods for combining good sets of features to produce useful overall suspiciousness rankings. These methods are compared in a user study that identifies a good subset of features for identifying suspicious reviews. Our evaluation shows that reviews highlighted as suspicious correspond with the opinions of real users.

The two aggregation methods evaluated are: singular value decomposition (SVD), and the unsupervised hedge algorithm (UH) [8]. Our evaluation demonstrates that these aggregation methods behave quite differently, resulting in rankings that do not correlate strongly. Notably, the top-ten suspicious hotels identified by both SVD and UH agree on just two of ten hotels. The assessment of these top-ten sets in the user trial suggests that SVD provides a more effective means of aggregating the various review filtering criteria.

In the next section, we present a very brief review of some relevant research, before presenting details of the suspiciousness criteria in Section 3. Details on the two aggregation methods are described in Section 4, where the details of the evaluation are also presented.

2 Related Work

The related research relevant to this work falls into two categories: the work that directly addresses this problem, and the work on spam detection and authoritativeness in related domains. This second category is extensive, as there are many related or analogous problems that have received attention – in particular e-mail spam [2], link spam (search engine spam) [1], detecting attacks on recommender systems [5], and assessing authoritativeness on sites such as Wikipedia [4].

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 [6] and Hsu et al. [3] 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. This presents a particular problem for feature subset selection and aggregation. However, the features employed by O'Mahony & Smyth to identify helpful reviews have influenced the selection of the criteria described in Section 3.

The most relevant work we have found on aggregation is that by Tan & Jin [8]. They modify the original Hedge algorithm for rank/score aggregation so that it can operate in an unsupervised setting (see Section 4.1). They find this is better for opinion aggregation than a number of alternative techniques, including SVD. It is interesting that in our evaluation SVD proves to be more effective.

3 Suspiciousness Criteria

In this section, we define a number of criteria (features) that we expect will be predictive of suspicious reviews. These criteria are influenced by some of the criteria used in [6] but also by the expectation that shill reviewers will have unusual social network characteristics – we think of the bipartite network of reviews and hotels as a simple social network.

Following the practice in [6], we split hotel reviews into *positive* and *negative* categories according to the overall ratings where a positive review is one assigned 4 or 5 stars and reviews with less than 4 stars are negative. For all criteria, a higher value is more indicative of potential shilling activity – we rank hotels in descending order, where those at the top of the list are deemed to be most suspicious.

3.1 Proportion of Positive Singletons (PPS)

This criterion is motivated by two principles, that reviewers who have posted many reviews have more authority, and that reviewer profiles with may reviews take a lot of effort to create. Based on this a *positive singleton* is a positive review from a reviewer who has posted no other reviews. Thus, 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.

3.2 Concentration of Positive Singletons (CPS)

We expect that, typically multiple shill reviews will be injected to boost a hotel's popularity. We suggest that this activity may occur over a short time period, in which multiple user accounts are created and strongly-positive reviews for the target hotel are submitted in quick succession.[1] The

[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`

greater the degree of temporal clustering between a batch of positive reviews, 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 neighbour:

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

For r_1 and r_P the time to the beginning and end of the evaluation period is considered. This score is based on a Gaussian kernel where λ is a bandwidth parameter that controls the influence of the proximity of reviews. A higher value of λ will emphasise pairs of reviews that are very close in time. We examined a range of values for this parameter, but found that a value of $\lambda = 1$ was most effective on the TripAdvisor data.

3.3 Reactive Positive Singletons (RPS)

In an attempt to recover from negative reviews the management of a hotel may react by posting some positive shill reviews. These will show up as positive singletons that closely follow genuine negative reviews. The strength of evidence for these can be quantified as $\frac{T-t_i}{T}$ where T is the length of the entire time period covered by the dataset, and t_i is the *reaction time* associated with shill i (*i.e.* the number of days between the negative review and the subsequent shill).

If a hotel has n such reactive positive singletons, then we can accumulate this evidence into an RPS score as follows:

$$RPS(H) = \frac{1}{T_H}\left(1 - \prod_{i=1}^{n}\left(1 - \frac{T-t_i}{T}\right)\right) \tag{3}$$

$$= \frac{1}{T_H}\left(1 - \prod_{i=1}^{n}\left(\frac{t_i}{T}\right)\right) \tag{4}$$

where T_H is a normalisation factor for each hotel which is the elapsed time before the 1^{st} and n^{th} RPS. This employs what Shafer calls Hooper's rule of concurrent testimony [7] to accumulate the evidence. The evaluation in Section 4.3 suggests that RPS is a strong indicator of suspicious activity.

3.4 Review Weighted Rating (RWR)

With this criterion we seek to assess the impact on the hotel's average star rating of reviewers with little track record. To do this, we produce an alternative average rating where each reviewer's contribution is weighted by the number of reviews they posted (n_r). The difference between the unweighted and weighted rating is the RWR score:

$$RWR(H) = \frac{1}{|R_H|}\sum_{r \in R_H} r - \frac{\sum_{r \in R_H} r \times n_r}{|R_H| \sum n_r} \tag{5}$$

where R_H is the set of ratings for hotel H. The mean difference will be greater for hotels that have received high scores from members who have infrequently reviewed hotels.

3.5 Contribution Weighted Rating (CWR)

This criterion elaborates on the RWR score by considering all contributions by the reviews. TripAdvisor allows members to make contributions other than reviews: this includes photos, videos, forum posts, articles, and travel itineraries. We extract the total numbers of contributions for members, and define CWR as follows:

$$CWR(H) = \frac{1}{|R_H|}\sum_{r \in R_H} r - \frac{\sum_{r \in R_H} r \times c_r}{|R_H| \sum c_r} \tag{6}$$

where c_r is the number of contributions from the reviewer who produced the rating. It is interesting that neither CWR and RWR fare well in the evaluation in Section 4.3.

3.6 Truncated Rating (TR)

The idea here is to remove a portion of the most positive reviews for a hotel and recalculate the average star rating to see if it deviates much from the simple average. This TR score is calculated as follows:

$$TR(H) = \frac{1}{|R_H|} \sum_{r \in R_H} r - \frac{1}{|R_H^{tr}|} \sum_{r \in R_H^{tr}} r \tag{7}$$

where R_H^{tr} is the truncated rating set. In our evaluation the top 20% of ratings were removed. Hotels where the average rating falls more than average after deleting the top 20% of reviews are suspicious, they would have a high TR score.

3.7 Sentiment Shift (SS)

An alternative strategy is to look at the change in hotel ratings over time. We split each hotel's reviews into "early" and "late" sets. These correspond to the first and second year of our dataset. SS is a measure of the change in popularity between the first and second period:

$$SS(H) = \frac{1}{|R_l|} \sum_{r \in R_l} r - \frac{1}{|R_e|} \sum_{r \in R_e} r \tag{8}$$

where R_e and R_l are the early and late rating/review sets. A positive SS score may indicate shilling in the second period.

3.8 Positive Review Length Difference (PRLD)

In our attempts to identify fraudulent reviews, we also wish to consider the text content of the reviews. Creating long reviews that appear to be genuine is a time-consuming process, so we expect that spam reviews might be shorter than normal. The PRLD feature identifies hotels with reviews that deviate a lot from the mean. The PRLD score for a hotel H is the average absolute difference between the length of its positive reviews and the mean length.

$$PRLD(H) = \frac{1}{|PosRev_H|} \sum_{p \in PosRev_H} |len_p - \overline{len}| \tag{9}$$

where $PosRev_H$ is the set of positive reviews, len_p is length of review p, and $\overline{len}$ is the mean length of positive reviews.

4 Evaluation

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 gathered a dataset of 26,903 reviews from 21,440 unique reviewers, covering hotels from all regions of Ireland over a two-year time window from September 2007 to September 2009. A total of 741 hotels are covered in our evaluation – we have eight criteria to score these hotels (PPS, CPS, RPS, RWR, CWR, TR, SS and PRLD). This gives us a 741 × 8 score matrix where the rows are the hotels and columns are the features.

4.1 Aggregation Methods

Given the score matrix, we need a means of aggregating the 8 columns to produce a single suspiciousness ranking. To do this we consider two alternative techniques:

Singular Value Decomposition: SVD is a well established technique for projecting high-dimensional data into a lower dimension space. The standard form is $\mathbf{X}_{n\times m} \simeq \mathbf{T}_{n\times k}\mathbf{S}_{k\times k}\mathbf{V}_{k\times m}$ where $\mathbf{X}$ is a matrix describing n items in terms of m features and $\mathbf{S}$ is a diagonal matrix of k singular values. In order to produce an aggregated ranking we use just one singular value so the decomposition is $\mathbf{X}_{n\times m} \simeq \mathbf{T}_{n\times 1}\mathbf{S}_{1\times 1}\mathbf{V}_{1\times m}$ so the scores in $\mathbf{T}_{n\times 1}$ give us a ranking of the hotels.

Unsupervised Hedge Algorithm: This is an unsupervised variant of an older supervised rank aggregation algorithm, adapted by Tan & Jin [8]. In the absence of supervision, it sets out to produce a ranking with maximal agreement with the component rankings. To do this it produces an aggregate score that is a weighted sum of the component scores $f_{comb}(x) = \sum_{i=1}^{m} w_i f_i(x)$ where w_i is the weight assigned to the i^{th} score. The algorithm works iteratively with the weights updated at each step.

$$w_i^{t+1} = \frac{w_i^t \beta^{\text{Loss}(f_j, f_{comb}^t)}}{\sum_{j=1}^{m} w_j^t \beta^{\text{Loss}(f_j, f_{comb}^t)}} \tag{10}$$

The parameter β is effectively a learning rate that controls convergence, and 'Loss' is a simple function that quantifies the disagreement between the aggregate score and the component score - see [8] for details. This weight update strategy has the effect of de-emphasising component scores that disagree with the aggregate.

4.2 Comparison of Aggregation Methods

For each of the aggregation methods describe above, we evaluate two variants, since aggregation can be performed either on the raw feature scores or on the rankings derived from these scores. The rank correlations between the four resulting rankings of the 741 hotels are listed in Table 1. It is interesting to observe that both UH variants are very strongly correlated. This is because, in both cases, the aggregation is dominated by a single feature, RWR.

Table 1: The rank correlations between the rankings from the four aggregation alternatives.

	SVD$_s$	SVD$_r$	UH$_s$	UH$_r$
SVD$_s$	1.00	0.90	0.56	0.56
SVD$_r$	0.90	1.00	0.66	0.66
UH$_s$	0.56	0.66	1.00	1.00
UH$_r$	0.56	0.66	1.00	1.00

To produce a ground truth for our dataset, we conducted a user study. We firstly selected 41 hotels corresponding to the union of the top-ten sets of hotels selected by a variety of aggregation and feature selection alternatives. Five unsuspicious hotels were added to this set to act as a control. Users were presented with a random selection of six of these hotels, and asked to mark any reviews that might appear suspicious. Based on the judgements provided by 55 users who completed the task, we calculated a *suspiciousness* score for each of the 46 hotels – the mean of the fraction of reviews marked as suspicious by each user.

The chart in Figure 1 shows that the five control hotels have the lowest suspiciousness score, as expected. The top-ten lists produced by the two SVD-based aggregations scored equally well, while the UH results are disappointing. Note that the lists for the ranking and scored-based UH methods were identical. The poor performance of UH seems to be because the consensus is dominated by the RWR feature, a feature that is not very informative – see next sub-section.

4.3 Comparison of Suspiciousness Criteria

It is interesting to assess which of the features are most predictive of suspiciousness as determined by the annotators. To do this, we compare the feature-based rankings for the 46 hotels with the suspiciousness ranking according to the annotators – see Figure 2. This analysis shows that RPS and TR are the strongest features while review length (PRLD) and the weight of established reviewers (RWR and CWR) seem to provide little information. The poor performance of RWR also explains why UH does not perform well in the aggregation as the UH ranking is dominated by RWR.

To investigate how useful the strongest features alone proved in identifying suspicious reviews, we applied the SVD scores aggregation method based on subsets of the top 3 and 5 features from Figure 2. The results in Figure 3 show that using the 3 best features (RPS,TR,PPS) leads to a significant improvement over the full set of 8 features. This indicates that this combination of criteria provides an effective means of identifying shill hotel reviews.

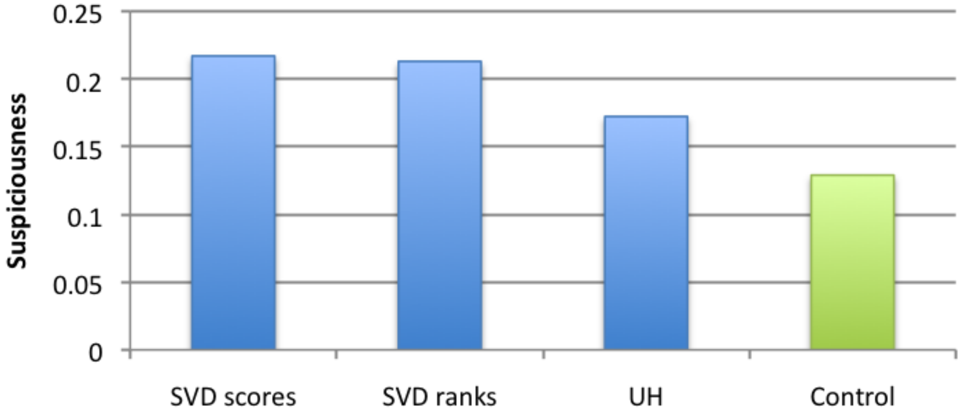

Figure 1: Suspiciousness scores for the top-ten lists produced by three aggregation alternatives.

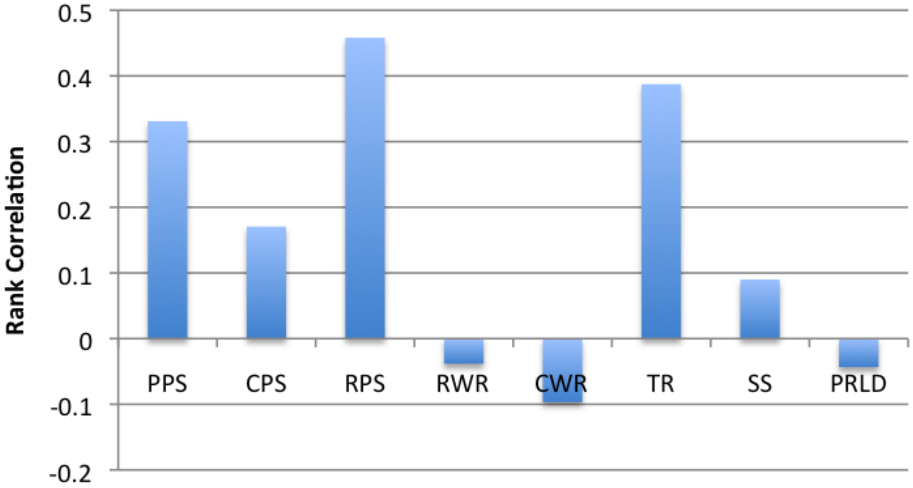

Figure 2: The correlation of feature-based rankings and the suspiciousness ranking by annotators.

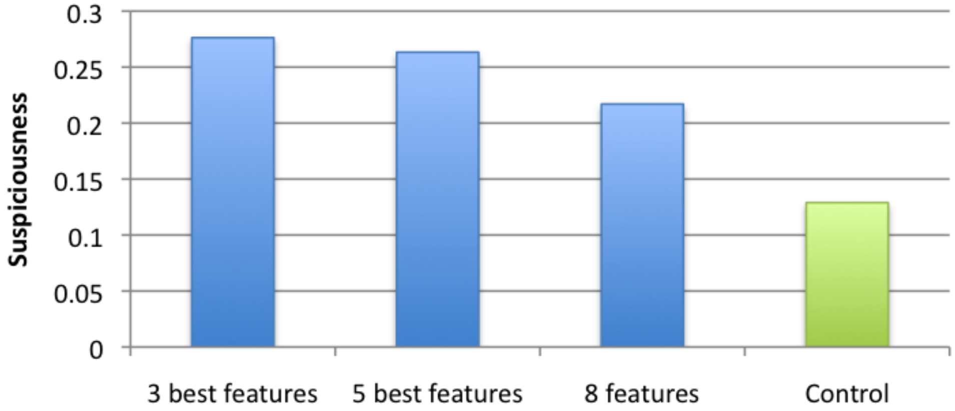

Figure 3: Suspiciousness scores for the top-ten lists produced by SVD on the strongest features.

5 Conclusions

We have presented a number of criteria for highlighting suspicious hotel reviews on TripAdvisor. A natural question arises regarding how best to integrate their outputs. We have evaluated two alternative strategies for aggregating the criteria into a single ranking. Surprisingly, we have found that SVD outperforms UH in this aggregation – this is the opposite of the findings in [8].

Our evaluation suggests that positive reviews that quickly follow negative reviews are suspicious, as are hotels whose ranking deteriorated dramatically when the 20% most positive reviews are removed. Both of these criteria will focus on hotels with significant variance in their review sets.

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

References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] G. Shafer. The combination of evidence. *International Journal of Intelligent Systems*, 1(3):155–179, 1986.

[8] P. Tan and R. Jin. Ordering patterns by combining opinions from multiple sources. In *Proc. 10th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data mining*, page 700. ACM, 2004.