When Opinion Request Meets Majority Search: Avoiding Fraud in On-line Review Systems*

(Extended Abstract)

Roberto Centeno
Dept. of Computer Systems and Languages
UNED
Madrid, Spain
rcenteno@lsi.uned.es

Ramón Hermoso
Dept. of Informatics and Systems Engineering
University of Zaragoza
Zaragoza, Spain
rhermoso@unizar.es

ABSTRACT

On-line review systems have been proved to be sensitive to fraud and have shown some shortcomings due to their reliance on using numerical ratings as reviews. For that reason, and supported by recent work on the field, we aim to address the problem of fraud in this type of systems, by designing a mechanism based on pairwise comparisons – encapsulated as opinion polls – coupled with an incentive policy attempting to foster the collection of majority opinions over individual experiences.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence - Multiagent Systems

General Terms

Algorithms

Keywords

On-line review systems, fraud, opinions, incentives

1. INTRODUCTION

In this work we focus on the On-line Review Systems (ORS onwards), where users may exchange opinions and check what others prefer. These systems represent the global view of the society about a set of entities, typically by means of reputation rankings. They usually capture users' opinion through numerical ratings that are aggregated building reputation rankings. However, three problems are identified: *i)* the difficulty for users of mapping an opinion to a single numerical value; *ii)* what is called as *selection bias problem* that defines a potential bias problem showing a heavily and usually positive skewed tendency in average users ratings caused by the use of numerical evaluations; and *iii)* fraudulent opinions may appear, understanding fraud as the fact of promoting bad entities, or damage the reputation of good entities [1], by giving an opinion contrary to what s/he actually believes about that entity.

All these potential problems may lead to capture imprecisely users' preferences and, then, build inaccurate reputation rankings, skewing the reputation of some entities. Thus, how to capture users' opinions accurately and how to deal with fraudulent opinions to avoid skewed reputation rankings, become important issues to study.

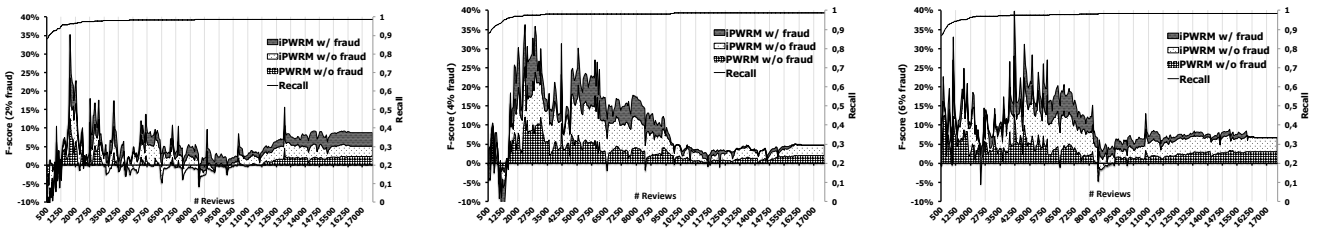
With all this in mind, the aim of this work is to give a solution to the aforementioned problems. Hence, we propose a solution for avoiding fraud in ORS based on capturing users' preferences through comparative reviews built as pairwise queries such as: *"which entity do you think users prefer, A or B?"*, where users agreeing the majority's choice are rewarded. The rationale behind this approach is that, as shown by several empirical studies, people expect to be "typical" and, therefore, overestimate the popularity of their own choices. Thus, assuming that fraud users are rational and minority in a ORS, they will be persuaded to be honest.

2. iPWRM FRAMEWORK

We tackle the problem formulated above by using a new mechanism, called iPWRM (incentive PairWise Reputation Mechanism). It is based on our previous works [3, 2] where we proposed the mechanism PWRM in charge of capturing ORS users' preferences through comparative opinions. The idea of iPWRM is to generate comparative queries, representing opinion polls over different entities of the system and, following an iterative process, to aggregate them building a reputation ranking. On the other hand, in order to obtain accurate replies, opinions polls are coupled with an incentive policy that rewards users whose replies agree the majority's choice, persuading them to give honest reviews. We take the concept of *match* to represent a question such as *"which entity do you think users prefer, A or B?"*, where users are requested to express their believe about the global opinion users in the system have over two particular entities. Therefore, a match can be understood as a poll conducted among different users in order to estimate the preference of the system - as a whole - about a particular entity over another. The outcome of a match reflects somehow a picture of how the system prefers an entity over the other.

*Research supported by eMadrid project S2013-ICE-2715, Spanish Ministry of Economy and Competitiveness (TIN2012-36586-C03-02-iHAS) and by the Autonomous Region of Madrid (P2013/ICE-3019-MOSI-AGIL-CM, co-funded by EU-FSE and FEDER")

Appears in: *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015), Bordini, Elkind, Weiss, Yolum (eds.), May 4–8, 2015, Istanbul, Turkey.*



(a) 2% fraudulent users (b) 4% fraudulent users (c) 6% fraudulent users

Figure 1: Experimental results regarding F-score improvement with Flixster dataset

The idea of iPWRM is to articulate an iterative process in which different matches are built and scheduled in knock-out tournaments allowing us to extract implicit relationships between entities, even when they are not compared directly each other. First, the mechanism selects a subset of entities to be evaluated. After that, the tournament is scheduled so forming initial matches. Then, the mechanism has to choose a set (possibly different for every match) of users to send the different matches to. Subsequently, the mechanism collects the reviews, in terms of votes, representing the solution of users for the match, and determines which entity wins. Once the match is solved, the mechanism applies its incentive policy and updates the empirical distribution of the current match. After that, it aggregates the results, updating the reputation ranking by using an adaptation of the Rank Centrality Algorithm [4] used by the PWRM. Then, the mechanism sets up the next round of the tournament with those entities that won in the previous one. This process is repeated until no matches are left in the tournament.

The mechanism provides users with partial global results of each match. They are encapsulated in a concept called *empirical distribution*. It is defined in terms of the win/loss ratio of votes received by an entity over the another. Each empirical distribution is publicly provided, so users may use this information in order to evaluate and select the reply they expect will maximize their utility.

As our main objective is to avoid the potential bias introduced by fraud opinions into an ORS, we couple the iPWRM framework with an incentive policy based on *rewarding those users that agree with the majority's choice*. Therefore, when rational users face a match and they have something to gain/lose, it is expected they will reply honestly. The idea is: *i)* to give a positive reward when all users agree in their reviews; *ii)* to give a positive reward (lower than previously) when a user agrees the majority's choice; and, *iii)* to give a low incentive - possibly null - when user's review does not match the majority.

We have conducted experiments with the aim of proving whether iPWRM is able to mitigate fraud effects. Rational users have been simulated by rational utility maximiser agents. User's decision for a match query reply is made according to his/her private opinion (extracted from Flixster dataset collected from the social movie site Flixster) and his/her belief that his/her opinion is also the preferred one by the majority of the system (modelled by a parameter τ enclosing the alignment of users with society regarding each particular pairwise comparison). After taking actions, users adapt their beliefs by taking into account how well their actions performed in terms of reward, as well as the empirical distributions collected by the mechanism.

We experiment with different percentage of fraudulent populations to test the performance and effectiveness of the mechanism. For each experiment, we show average results from 10 different runs where several subsets of 50 movies were randomly selected to be ranked. Besides, in each execution a different set of agents is selected to be set up as fraudulent users. We show the average improvement (positive or negative) of the F-score (combining precision and recall metrics), in terms of percentage regarding a reference ranking representing an ideal scenario – in which the mechanism is able to query all users, about all possible pairwise comparisons, with no fraud involved. In Figures 1(a), 1(b) and 1(c) we present how F-score evolves in executions with PWRM without fraud, iPWRM without fraud, as well as iPWRM with fraud, compared to PWRM with fraud, all of them with different fraud populations (2%, 4% and 6%). As we can observe, the iPWRM improves the PWRM mechanism in all cases, independently the rate of fraud. Therefore, we conclude the iPWRM is able to persuade users to avoid fraud reviews, also mitigating the potential bias problems.

3. CONCLUSIONS

In this paper we have put forward a mechanism, called iPWRM that builds reputation rankings modelling reviews as pairwise opinion polls instead of numerical ratings. This mechanism is also endowed with an incentive policy based on rewarding majority's choice for deterring rational users from providing fraudulent reviews. Moreover, the proposed mechanism has been tested by using agents as rational users whose preferences have been built from a real dataset obtained from Flixster site. The results show that our proposal is able to overcome the bias introduced as consequence of fraudulent reviews.

REFERENCES

[1] L. Akoglu, R. Chandy, and C. Faloutsos. Opinion fraud detection in online reviews by network effects. In *ICWSM*, 2013.

[2] R. Centeno, R. Hermoso, and M. Fasli. On the inaccuracy of numerical ratings: dealing with biased opinions in social networks. *Information Systems Frontiers*, pages 1–17, 2014.

[3] R. Hermoso, R. Centeno, and M. Fasli. From blurry numbers to clear preferences: A mechanism to extract reputation in social networks. *Expert Systems with Applications*, 41(5):2269–2285, 2013.

[4] S. Negahban, S. Oh, and D. Shah. Iterative ranking from pair-wise comparisons. In *NIPS*, 2012.