Decision Support Systems 79 (2015) 99–108

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Towards a highly effective and robust Web credibility evaluation system

Xin Liu [a,*], Radoslaw Nielek [b], Paulina Adamska [b], Adam Wierzbicki [b], Karl Aberer [c]

[a] Data Analytics Department, Institute for Infocomm Research, Singapore
[b] Department of Computer Networks, Polish Japanese Institute of Information Technology, Poland
[c] School of Computer Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

ARTICLE INFO

Article history:
Received 2 May 2014
Received in revised form 5 May 2015
Accepted 22 July 2015
Available online 31 July 2015

Keywords:
Web credibility
Recommendation
Robustness
Imitating attack

ABSTRACT

By leveraging crowdsourcing, Web credibility evaluation systems (WCESs) have become a promising tool to assess the credibility of Web content, e.g., Web pages. However, existing systems adopt a passive way to collect users' credibility ratings, which incurs two crucial challenges: (1) a considerable fraction of Web content have few or even no ratings, so the coverage (or effectiveness) of the system is low; (2) malicious users may submit fake ratings to damage the reliability of the system. In order to realize a highly effective and robust WCES, we propose to integrate recommendation functionality into the system. On the one hand, by fusing Matrix Factorization and Latent Dirichlet Allocation, a personalized Web content recommendation model is proposed to attract users to rate more Web pages, i.e., the coverage is increased. On the other hand, by analyzing a user's reaction to the recommended Web content, we detect imitating attackers, which have recently been recognized as a particular threat to WCES to make the system more robust. Moreover, an adaptive reputation system is designed to motivate users to more actively interact with the integrated recommendation functionality. We conduct experiments using both real datasets and synthetic data to demonstrate how our proposed recommendation components significantly improve the effectiveness and robustness of existing WCES.

1. Introduction

The Internet has become the primary information source to serve people's daily life. However, unlike traditional media such as television and newspapers, a considerable fraction of contents are posted online without being seriously fact-checked. This may incur serious consequences if non-credible Web information is used for decision making. It is thus important to assess the credibility of Web content [6,20,17].

Recently, by leveraging crowdsourcing, several Web credibility evaluation systems (WCESs) have emerged. For instance, MyWOT (www.mywot.com) aggregates users' ratings on two aspects of Web credibility: trustworthiness and child safety. In academia, similar systems have also been proposed, improving the commercial counterparts in various aspects [22,21,18].

Although WCES has become a promising tool to assess Web credibility, most existing systems adopt a passive way to collect users' ratings (i.e., they wait for users to submit ratings, but actually only a few users voluntarily provide Web credibility assessments), which incurs two crucial challenges. *First, given the huge volume of Web pages, a considerable fraction of them have no credibility information.* For instance, based on our analysis, among the top 1 million domains in Alexa traffic ranking (www.alexa.com), only around 42.67% of them (as of January 9th, 2013) are covered by the most representative credibility evaluation site MyWOT, not to mention the domains with low popularity.[1] Furthermore, most domains covered by MyWOT have a limited amount of ratings, i.e., the confidence of their credibility information is low.

Second, malicious users may submit fake ratings to attack certain Web content. Although existing reputation systems can handle average malicious users, they are not effective in detecting smart attackers. In [14], a new type of attack called imitating attack is identified and verified: an attacker queries the credibility of certain Web content, when he receives the rating from the system, he just copies and resubmits the same rating to the system again as his own contribution.[2] From the perspective of the system, this attacker behaves quite similarly to a highly reputable user because his ratings are always consistent with the system's ratings. In this way, the attacker can easily build high reputation by cheating the system and then effectively attack certain Web content. A machine learning based defense mechanism is proposed in [14], but its computational complexity analysis is missing, so its applicability in a large-scale system with millions of users has not been verified.

* Corresponding author. Tel.: +65 64082227.
E-mail addresses: liu-x@i2r.a-star.edu.sg (X. Liu), nielek@pjwstk.edu.pl (R. Nielek), tiia@pjwstk.edu.pl (P. Adamska), adamw@pjwstk.edu.pl (A. Wierzbicki), karl.aberer@epfl.ch (K. Aberer).

[1] Based on our study, the probability that a Website has credibility evaluation is strongly correlated with its popularity, i.e., ranking.

[2] Note that aggregating users' ratings to derive system's ratings is another yet challenging research question, particularly when a Web content is highly subjective [11]. However, this issue is beyond the scope of this paper and here we assume that system's ratings have been reliably generated.

http://dx.doi.org/10.1016/j.dss.2015.07.010

In order to address the two challenges mentioned above, we propose to integrate recommendation functionality into a WCES to (1) attract users to rate more Web pages to increase the coverage of the system and (2) defend against the imitating attack to make the system more robust. To achieve the first goal, we propose a recommendation model to learn users' interests in Web content by fusing Matrix Factorization (MF) and Latent Dirichlet Allocation (LDA) [4]. Specifically, each Web content is represented by a "bag-of-words". We apply LDA to extract a set of topics for the Web content, and assign a latent factor vector to each topic. A user's interest in a Web content is inferred by combining (1) the correlation between a user and the topics (by multiplying user-specific and topic-specific latent factors) and (2) the correlation between the Web content and the topics (by LDA).

In order to realize the second goal, we propose a recommendation based defense mechanism. The system first selects a set of active users[3] and recommends a set of interesting Web pages to them. Along with a recommended content, a fake credibility rating is shown and the system entices users to provide their ratings to make the system's rating more confident. Since imitating attackers simply copy system's ratings, they will copy (submit) the fake rating with a high probability. On the other hand, honest users often spend efforts in evaluating the recommended Web content, and submit a rating that is closer to the real credibility. We propose to model users' responses to the recommended Web content using beta distribution to detect imitating attackers.

To further enhance the effectiveness and efficacy of the integrated recommendation functionality, *we design an adaptive reputation system to motivate users to more actively rate the recommended Web content. The basic idea is to assign different reputation points to users when they rate Web content in different ways*. For instance, more reputation will be awarded if a user rates a recommended content than a self-selected content. Since rating the recommended Web content can boost reputation rapidly, both honest users and attackers are motivated to actively interact with recommendations thus further improving the system's coverage and robustness.

Although Web credibility, recommender systems and reputation systems have been well studied in their respective literature, to the best of our knowledge, *this work is the first one that seamlessly integrates diverse information sources by applying various information retrieval techniques to realize a highly effective and robust WCES*. The contributions of this work are summarized as follows: (1) In order to handle the challenges of coverage and robustness, we propose to integrate recommendation functionality into a WCES (see Section 3 System model). Note that such a system is independent and self-contained, i.e., it is not designed as a component of other information systems, e.g., Web search engines. (2) By combining MF and LDA, we propose a personalized recommendation model to motivate users to rate more Web pages (Section 4.1). (3) In order to defend against the imitating attack, we propose a recommendation based defense mechanism. The system recommends a set of Web pages with fake credibility ratings to users and analyzes their rating behavior in the presence of recommendations. Beta distribution is used to model a user's imitating behavior probability (Section 4.2). (4) For the purpose of enhancing the effectiveness of the integrated recommendation functionality, in Section 4.3, we design an adaptive reputation system to award more reputation points when a user rates the recommended Web content. This not only motivates normal users to rate more to increase system's coverage, but also entices malicious users to copy fake credibility rating (i.e., get detected). (5) We evaluate the effectiveness of the recommendation functionality using the data published by Wikimedia foundation as well as a simulated multi-agent system (Section 5).

[3] In order to quickly build high reputation, attackers typically densely submit ratings thus are highly active in certain periods of time.

2. Background and related work

2.1. Web credibility

Quite a few works have identified a variety of factors that may influence an individual's perception of Web credibility [7,6,24]. For instance, Schwarz et al. [20] showed that visualizations by considering features such as Webpage popularity, domain type and the PageRank metric, can improve a user's Web credibility assessment in Web search results.

Recently, WCES has become popular in both academia and industry. Sharifi et al. [22] proposed SmartNotes, a crowdsourcing system to detect Web security threats such as Internet scams. Machine learning and natural language processing are applied to analyze and integrate users' reports. In [18], Web credibility is assessed by a decentralized recommender system. A single credibility metric is derived by combining three components: (1) item-based collaborative filtering, which is based on features identified from pages' textual contents, (2) user-based collaborative filtering, which is based on users' social relationships and (3) Web search page ranking.

In practice, MyWOT is a real-world example of a WCES that can have a real economic impact on content providers. In MyWOT, entire domains receive credibility ratings, and these ratings are later used by a Web browser extension that displays them next to Google search results. A domain that has a very low rating could therefore experience a significant decrease of Web traffic, even if it would be ranked high by Google. Therefore, a competitor of some commercial company would have a real incentive to maliciously decrease his competitor's credibility rating. This could be achieved through several simple means, such as spamming the MyWOT system automatically with negative credibility ratings. However, such crude means could be equally easily detected. This motivates attackers to devise more sophisticated means, out of which an imitating attack is one of the least expensive and most effective.

Although WCES has been studied recently, its robustness issue is relatively unexplored. Most existing solutions apply reputation systems to constrain users' rating behavior, however, smart attackers can easily cheat for high reputation. Liu et al. [14] identified a new type of attack called imitating attack, where a malicious user simply copies and resubmits the system's ratings to pretend to be a reputable user to gain high reputation. By studying Web pages' characteristics and users' rating behavior patterns, a two-stage machine learning based defense mechanism is proposed. Experimental results demonstrate the effectiveness of the approach, but its computational complexity analysis is not provided. In this work, we try to defend against the imitating attack using a light-weight, recommendation based approach.

2.2. Recommender systems

Although no recommendation is provided in existing WCES, recommender systems have been widely studied in many other application scenarios such as e-commerce and online social networks. Traditional recommender systems rely on collaborative filtering [1], which predicts a user's interest in an item by mining rating information of other similar users and items. In particular, MF has been proved to be one of the most effective methods in terms of rating prediction [12].

In order to improve recommendation quality, a variety of types of information, e.g., social information and contextual information is integrated into MF. For instance, in [13], the authors applied random decision trees to combine diverse types of contexts. Social information is integrated as a MF regularization term, but different from previous work [15], contexts are considered when measuring user similarity. Recently, content based recommendation has been improved by fusing topic modeling and MF [2,16]. The basic idea is to assign a latent topic to each word of an item, and then generate item latent factors by averaging the topics of all words associated with this item. A missing rating