Original Research Article: full paper

(2020), «EUREKA: Physics and Engineering»
Number 5

METHOD FOR DETECTING SHILLING ATTACKS BASED ON IMPLICIT FEEDBACK IN RECOMMENDER SYSTEMS

Oksana Chala
Department of Information Control Systems[1]
oksana.chala@nure.ua

Lyudmyla Novikova
Department of International Relations
International Information and Security[2]
l.novikova@karazin.ua

Larysa Chernyshova
Department of International Relations
International Information and Security[2]
lchernyshova@karazin.ua

Angelika Kalnitskaya
Department of Information Control Systems[1]
angelika.kalnitskaya@nure.ua

[1]*Kharkiv National University of Radio Electronics*
14 Nauka ave., Kharkiv, Ukraine, 61166
[2]*V. N. Karazin Kharkiv National University*
6 Svobody sq., Kharkiv, Ukraine, 61022

Abstract

The problem of identifying shilling attacks, which are aimed at forming false ratings of objects in the recommender system, is considered. The purpose of such attacks is to include in the recommended list of items the goods specified by the attacking user. The recommendations obtained as a result of the attack will not correspond to customers' real preferences, which can lead to distrust of the recommender system and a drop in sales. The existing methods for detecting shilling attacks use explicit feedback from the user and are focused primarily on building patterns that describe the key characteristics of the attack. However, such patterns only partially take into account the dynamics of user interests. A method for detecting shilling attacks using implicit feedback is proposed by comparing the temporal description of user selection processes and ratings. Models of such processes are formed using a set of weighted temporal rules that define the relationship in time between the moments when users select a given object. The method uses time-ordered input data. The method includes the stages of forming sets of weighted temporal rules for describing sales processes and creating ratings, calculating a set of ratings for these processes, and forming attack indicators based on a comparison of the ratings obtained. The resulting signs make it possible to distinguish between nuke and push attacks. The method is designed to identify discrepancies in the dynamics of purchases and ratings, even in the absence of rating values at certain time intervals. The technique makes it possible to identify an approach to masking an attack based on a comparison of the rating values and the received attack indicators. When applied iteratively, the method allows to refine the list of profiles of potential attackers. The technique can be used in conjunction with pattern-oriented approaches to identifying shilling attacks.

Keywords: e-commerce, recommendation system, temporal rules, shilling attack, feedback.

DOI: 10.21303/2461-4262.2020.001394

1. Introduction

Recommender systems are one of the key elements of e-commerce systems. They are designed to personalize the offered goods and services in accordance with the interests of a particular user [1]. For example, the streaming service Netflix receives a significant number of views based on personal recommendations [2]. The recommendation system “predicts” the interests of the target user based on information about the similarity of goods, the choice of users with similar interests, and ratings of objects received from clients. Recommendations building algorithms use

implicit and explicit feedback from the user. Implicit feedback is presented by purchases and confirmed by the user's financial spending. Explicit feedback is given by ratings and reflects the user's impressions of the offered goods [3].

If the consumer behaves in bad faith, the use of explicit feedback leads to the vulnerability of the recommendation system. A malicious user can distort the ratings of objects to change the sales of the target group. For example, to increase the demand for a specific manufacturer's goods or reduce the interest in the goods of its competitors. Therefore, when constructing recommendations, it is necessary to take into account user attacks or shilling attacks [4]. Such attacks are based on the ability to influence the user's interests using recommendations [5]. A shilling attack creates a set of fake user profiles. These profiles are used to generate artificial product ratings in the recommender system. As a result, clients of such systems receive a personalized list of objects that reflect attackers' interests. Shilling attacks provide a short-term boost in sales of goods and services important to cybercriminals. However, distorted ratings undermine the recommendations' credibility as a whole, which can lead to the long-term sales decline.

When conducting an attack, a malicious user solves two problems: falsifying ratings and masking an attack. Within the framework of the first task, the maximum possible rating without disclosing the attack is set for the target objects and the minimum rating for competing products [6]. The attack is masked by assigning ratings that are similar to those of other users. In works [1, 4], a classification of attacks is presented, taking into account the knowledge used about the recommender subsystem's work. According to the method of masking, it is advisable to combine these types of attacks into three groups: concealment based on data on existing ratings of objects (1); masking using information about popular items (2); the cover-up of an attack based on product segmentation (3).

The existing methods for detecting shilling attacks are focused mainly on the formation of attack patterns using the results of explicit feedback [7]. Statistical methods and machine learning are used to build such patterns. Statistical methods are primarily focused on identifying attacks of the first group. Such methods reveal the key characteristics of the attacker's profile that distinguish it from the average user. For example, uncharacteristic high or low ratings that do not match ratings from other users. To assess compliance, such indicators as the degree of similarity with the nearest neighbors [8], the deviation from the average rating value, taking into account the number of ratings and the number of users who posted these ratings [9] are used. Machine learning methods [10] make it possible to recognize attacks from all three groups, but they are very resource-intensive. Therefore, to identify complex attacks, work [11] uses a comparison of ratings at fixed time intervals. Work [12] proposes to dynamically change the duration of time intervals at which the attack profile will be detected. In general, the first group's approaches are focused on building patterns describing the differences between fake users and real ones, depending on the type of attack. The resulting templates only partially consider the changing interests of users over time. At the same time, in practice, the priorities of consumers change dynamically, for example, with a change in social status, workplace, or study. As a result, the patterns of their behavior change and, as a result, differences with fake users, which does not allow promptly detecting an attack on ratings.

In [13], when detecting attacks, it was proposed to take into account the temporal rules [14] to describe the dynamics of sales or ratings. Such rules can set both implicit constraints on the consumer's choice [15] and the conditions for this choice. The approach based on the comparison of individual temporal rules for purchases and ratings is focused primarily on shilling attacks of the second group since users are constantly rating such objects. However, during attacks of the first group, ratings may be set irregularly. In such cases, there are no temporal rules for ratings at individual time intervals, which do not allow comparing the change in sales and ratings and detecting an attack. In [16], it is proposed to present the dynamics of forming recommendations for a given period in the form of a multilayer temporal graph. However, such a graph model does not provide a comparison of the selection and presentation of ratings.

To take into account the results of implicit feedback within the selected time period, it is advisable to describe the processes of choosing a product and setting its ratings at several successive

intervals, linking them with temporal rules. Thus, to detect shilling attacks based on the identification of discrepancies between object selection processes and ratings, temporal models can be used to take into account the dynamics of the user's interests.

This study aimed to develop a method for the rapid detection of shilling attacks, taking into account the dynamics of user preferences.

To achieve the aim of research, the following objectives were set:

– develop a temporal model of the process of changing of the recommender system users' preferences;

– develop a method for detecting shilling attacks based on comparing the results of explicit and implicit feedback from the user;

– evaluate the effectiveness of the developed method for detecting shilling attacks.

2. A temporal model of the process of changing of the recommender system users' preferences

Changes in the users' preferences of the recommender system with a given item are reflected in the processes of purchasing this product and setting its ratings. In order to describe the temporal order of these processes' events, the adapted temporal rules of two types are proposed: "Next" and "Future". Each of these rules sets the order in time for a pair of facts Φ_m and Φ_s, reflecting the choice (purchase) of a product or setting its rating. The fact becomes true when given events occur, such as the choice of a given object at a given time τ; selection of multiple instances of an item on a given subset of purchases.

Each temporal rule $r_{m,s}^{(j)}$ specifies a relative temporal order of the early-later type. The rule $r_{m,s}^{(j)}$ determines that after the fact Φ_m of purchase of goods i_j on the interval $\Delta\tau_m$, the fact Φ_s purchase of goods i_j on the interval $\Delta\tau_s$ will be true. Therefore, such rules can specify temporal relationships between intervals or points in time and between subsets of facts ordered in time. The "Next" rule uses the temporal operator X, which links two successive selection/rating events [13]. When this rule is fulfilled, no true intermediate facts can exist between the facts Φ_m and Φ_s. The rule of type "Future" uses the temporal operator F, which connects two non-consecutive events. Between the facts Φ_m and Φ_s in the F-rule composition, there must be at least one intermediate fact. The generalized form of the rule $r_{m,s}^{(j)}$ is:

$$r_{m,s}^{(j)} = \Phi_m \left(X \vee F \right) \Phi_s . \tag{1}$$

The Π_R rules $r_{m,s}^{(j)}$ sequence describes the temporal ordering of purchases (ratings) of the object i_j for the period T:

$$\Pi_R = \left\langle r_{1,2}^{(j)}, r_{1,3}^{(j)}, ..., r_{1,S}^{(j)}, ..., r_{s,s+1}^{(j)}, ..., r_{S-1,S}^{(j)} : \forall s \Delta\tau_s \in \mathrm{T} \right\rangle. \tag{2}$$

The expression rule $r_{1,2}^{(j)}$ (2) contains the temporal operator X since it connects the facts Φ_1 and Φ_2 of purchases (or rating assignments) on two adjacent intervals $\Delta\tau_1$ and $\Delta\tau_2$. Dependency $r_{1,3}^{(j)}$ is an example of a rule with a temporal operator F linking facts Φ_1 Φ_3.

In addition to redefining the facts, the adaptation of temporal rules consists of setting their weights, taking into account the dynamics of the user's interests. The weight $w_{m,s}^{(j)}$ of the rule $r_{m,s}^{(j)}$ is set through the normalized difference between the number of purchases or the average value of the product's ratings i_j on the intervals $\Delta\tau_m$ and $\Delta\tau_s$.

The sequence of selection of goods or setting their ratings is represented by an ordered set of normalized weights of the Π_W rules:

$$\Pi_W = \left\langle w_{1,2}, ..., w_{1,S}, ..., w_{m,m+1}, ..., w_{S-1,S} : \forall m \forall s \, r_{m,s}^{(j)} = \mathrm{true} \right\rangle. \tag{3}$$

Based on the sequence of weights (3) for any time interval $\Delta\tau_s$, it is possible to evaluate $W_s^{(j)}$, the change in the user's interest in the subject i_j over time. This estimate combines the change in users' interests to the selected object from the first interval $\Delta\tau_1$ to the current interval $\Delta\tau_s$:

$$W_s^{(j)} = \frac{\sum_{m=1}^{s-1} w_{m,s}}{\max_s \left(\sum_{m=1}^{s-1} w_{m,s} \right)}. \quad (4)$$

An example of a set of rules $\left\{ r_{m,s}^{(j)} \right\}$, that are used to calculate the score $W_4^{(j)}$, is shown in **Fig. 1**.

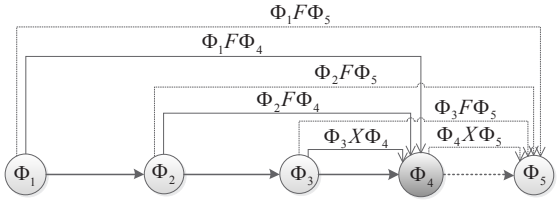

Fig. 1. Example of a subset of rules for calculating a score $W_4^{(j)}$

This example presents a sequence of facts ordered in $\langle \Phi_1, \Phi_2, \Phi_3, \Phi_4, \Phi_5 \rangle$. Such facts describe the process of choosing a given item (or setting its ratings) by various users on a continuous sequence of intervals

$$T = \langle \Delta\tau_1, \Delta\tau_2, \Delta\tau_3, \Delta\tau_4, \Delta\tau_5 \rangle.$$

In the example, the index j of the item is not specified for the sake of simplicity. The evaluation is formed for the current fact Φ_4. This fact is connected with the previous facts by F-rules $\Phi_1 F \Phi_4$ and $\Phi_2 F \Phi_4$, as well as the X-rule $\Phi_3 X \Phi_4$. The weights of these rules $w_{1,4}^{(j)}$, $w_{2,4}^{(j)}$ and $w_{3,4}^{(j)}$ are used to calculate the evaluation $W_4^{(j)}$. If there are no ratings values in the current interval, the previous range rules are used to construct the score $W_s^{(j)}$. For example, if on the interval $\Delta\tau_5$ for the one shown in **Fig. 1** of the example, there are no new rating values, then its average value remains the same as in the interval $\Delta\tau_4$. Then the weights of the $\Phi_1 F \Phi_5$, $\Phi_2 F \Phi_5$ and $\Phi_3 F \Phi_5$will be equal $w_{1,4}^{j}$, $w_{2,4}^{j}$ and $w_{3,4}^{j}$ respectively. The weight of the rule $\Phi_4 F \Phi_5$ is equal to zero since the average rating on the $\Delta\tau_5$ interval has not changed compared to the $\Delta\tau_4$ range due to the absence of new ratings.

Thus, estimate (4) on each current interval $\Delta\tau_s$ shows how users' preferences have changed to the product i_j compared with all previous time intervals.

Then the model $M^{(j)}$ of the process of changing the user's preferences for the subject i_j is a sequence of evaluations $W_s^{(j)}$ ordered by intervals $\Delta\tau_s$

$$M^{(j)} = \left\langle W_2^{(j)}, W_3^{(j)}, ..., W_s^{(j)} \right\rangle. \quad (5)$$

This model describes the change in sales or ratings for each interval $\Delta\tau_s$, which makes it possible to reveal the falsification of ratings by step-by-step comparison of the corresponding evaluations $W_s^{(j)}$.

3. Method for detecting shilling attacks based on comparing the results of explicit and implicit feedback from the user

The presented method, when detecting shilling attacks, compares the models of the processes of changing user preferences (5), obtained as a result of implicit (sales) and explicit (ratings) feedback. The method forms a quantitative assessment of the discrepancies between these processes. The initial data of the method are: sales list L; list of Q ratings; analysis period T; object of possible attack i_j; a subset of users – potential attackers $U=\{u_k\}$; the level of time detail (hour, day, week, month), represented by the length of the interval $\Delta\tau_s$. The original sales and rating lists include the following elements: u_k user id; the moment of choosing/setting the rating τ_s; the number n_k of goods i_j sold to the user n_k; rating ρ_k of product i_j, set by user n_k.

The method includes the following stages.

Stage 1. Preliminary processing of initial data. At this stage, data sets are formed for constructing sales facts, as well as the facts of assigning ratings at intervals $\Delta\tau_s$. The result of this stage is sets of facts of purchases $\left\{\Phi_s^{j,item}\right\}$ and ratings $\left\{\Phi_s^{j,rating}\right\}$. These facts contain information about the quantity of the purchased item $n_s^{(j)}$ and the average rating of this item $\rho_s^{(j)}$ at intervals $\Delta\tau_s$ for u_k users.

Stage 2. Construction in accordance with (2) sets of temporal rules for user selection $\Pi_R^{j,item}$ and rating assignment $\Pi_R^{j,rating}$. At this stage, Next-rules and Future-rules in the form (1) are formed from pairs of facts $\left(\Phi_m^{j,item}, \Phi_s^{j,items}\right)$, and $\left(\Phi_m^{j,rating}, \Phi_s^{j,rating}\right)$.

Stage 3. Formation, following expression (3), sets of weights of temporal rules for purchases $\Pi_W^{j,item}$ and $\Pi_W^{j,rating}$ ratings, respectively.

Step 3. 1. Construction of the set $\Pi_W^{j,item}$ is performed by normalizing the difference in the number of purchases n_s–n_m for all rules from the set $\Pi_R^{j,item}$.

Step 3. 2. Construction of the set $\Pi_W^{j,rating}$ is performed by normalizing the difference in ratings for the elements of the set $\Pi_R^{j,rating}$.

Stage 4. Building models of the processes of changing user preferences $M^{j,invoice}$ for sales and $M^{j,rating}$ for setting ratings.

Stage 5. Identification of intervals of a possible shilling attack. At this stage, both quantitative and qualitative differences between the corresponding elements of the sequences $M^{j,invoice}$ and $M^{j,rating}$ are taken into account.

Step 5. 1. Formation of a set of quantitative discrepancies $D^{(j)}$ between the purchasing and rating processes:

$$D^{(j)} = \left\{d_s^{(j)}\right\},$$

$$d_s^{(j)} = \begin{cases} \left|W_s^{j,item}\right| + \left|W_s^{j,rating}\right| \text{ if } \left(W_s^{j,item} \geq 0 \wedge W_s^{j,rating} < 0\right) \vee \left(W_s^{j,item} \leq 0 \wedge W_s^{j,rating} > 0\right). \\ 0 \text{ otherwise.} \end{cases} \quad (6)$$

At this step, the evaluations $W_s^{j,item}$ and $W_s^{j,rating}$ are summed up modulo in the case of oppositely directed changes in demand and rating, since such a multidirectionalness may indicate a possible attack.

Step 5. 2. Formation of the set of features $A^{(j)}$ of a possible attack of fake users based on the discrepancies $D^{(j)}$:

$$A^{(j)} = \left\{a_s^{(j)}\right\},$$

$$a_s^{(j)} = \begin{cases} -1 \text{ if } \left|d_s^{(j)}\right| \neq 0 \wedge W_s^{j,rating} < \\ < 0 \wedge W_s^{j,rating} \neq W_{s-1}^{j,rating}, \\ 1 \text{ if } \left|d_s^{(j)}\right| \neq 0 \wedge W_s^{j,rating} > \\ > 0 \wedge W_s^{j,rating} \neq W_{s-1}^{j,rating}, \\ 0 \text{ othewise.} \end{cases} \quad (7)$$

The condition of equality of the ratings $W_s^{j,rating} = W_{s-1}^{j,rating}$, represented by the "otherwise" clause in (7), blocks the formation of the attack attribute when the rating remains unchanged on a pair of successive intervals $\Delta\tau_{s-1}$ and $\Delta\tau_s$. This condition is fulfilled in the absence of information about the ratings on the interval $\Delta\tau_s$. Negative values $a_s^{(j)}$ indicate a possible attack to downgrade competitor products, while positive values indicate an attack aimed at increasing the rating of the target item.

The algorithm implementing this method is shown in **Fig. 2**. The algorithm includes the following steps.

Step 1. The input of initial data.

At this step, it is necessary to enter data on sales *L*, rating *Q*, period T, subject i_j; users *U*, and the granularity levels of the time. The granularity of time depends on the format of timestamps in sales and rating logs.

Step 2. Dividing the period *T* into time intervals $\Delta\tau_s$.

At this step, a time granularity level is selected that provides the highest value of the weights of the sales process rules

$$W^{j,item} = \sum_s W_s^{j,item}$$

for all intervals from period *T*. The result of this step is a set of intervals $\Delta\tau_s$ for period *T*.

Step 3. Implementation of the developed method for detecting shilling attacks.

Step 4. Clarification of the list of users. From the set *U* at *n* iterations, users $u_k^{(n)}$, are removed who did not set ratings at intervals $\Delta\tau_s$ with a sign $a_s^{(j)} \neq 0$, since these users did not take part in distorting ratings. The result of the step is a subset of users $U^{(n)}$, containing potential attackers.

Step 5. Checking the algorithm's termination condition $\left|U^{(n)}\right| = \left|U^{(n-1)}\right|$ according to which the number of users at the current *n* iteration has not changed compared to iteration *n*–1. When this condition is met, the operation of the algorithm ends. Otherwise, *step 6* is performed.

Step 6. Refinement of the set of *Q* ratings. User ratings are removed from this set since these users $u_k^{(n)}$, are not attackers. Next, the transition to the execution of the method at step 3 of the algorithm is performed.

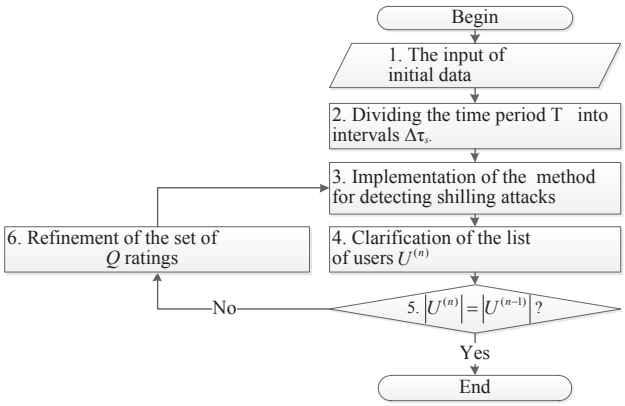

Fig. 2. Algorithm for the implementation of the proposed method

The algorithm execution results can be used for additional analysis of the method of concealing a shilling attack. This analysis is performed based on the comparison of deviations $d_s^{(j)}$ and characteristics $a_s^{(j)}$.

4. Results of experimental evaluation of the method for detecting shilling attacks

The experiment's goal was to test the effectiveness of the method for detecting shilling attacks on a time-ordered dataset of ratings and sales without information about absolute values of points in time. The basic assumption was that using a relative time scale and taking into account the use of F-rules, it is possible to distinguish intervals by the number of purchases of goods or services. In [17], it was shown that it is possible to form a description of the process by ordering events on the "earlier-later" time scale, taking into account their attributes, which makes it possible to justify this assumption.

The experiment included two phases. In the first phase, attacks were detected on the initial data set and an analysis of ways to conceal an attack. In the second phase, the proposed method's effectiveness is compared with the methods [11, 12], which, similarly to the proposed method, implement the partitioning of the initial data set into time intervals.

During the experiment, a dataset was used with information about the reading and ratings of several million books [18]. Read, and rating entries are ordered by time, but there are no absolute

timestamps in the original data. The facts $\Phi_m^{(j)}$ and $\Phi_s^{(j)}$ were formed on subsets of the input data from a fixed number of records.

The first phase looked at detecting attacks over long periods, represented by a large number of purchases. When constructing facts, subsets of 100,000 consecutive items (reading, ratings) were used. Such subsets correspond to the original intervals of the method and are denoted by Δ_s. The results of the method's key steps for the target i_{10000} (book with id=10000) are presented in **Table 1**.

Table 1

Method implementation results

Step number	The result of a step	Components of the result by subsets Δ_s								
		Δ_2	Δ_3	Δ_4	Δ_5	Δ_6	Δ_7	Δ_8	Δ_9	Δ_{10}
4. 1	$M^{10000,\,item}$	0.19	0.19	–0.41	0.04	–0.63	–0.52	0.07	–1.00	–0.59
4. 2	$M^{10000,\,rating}$	–0.10	–0.10	–0.30	**–0.30**	**–0.30**	–0.03	–0.03	–0.03	1.00
5. 1	$D^{(10000)}$	0.28	0.28	0.00	0.33	0.00	0.00	0.11	0.00	1.59
5. 2	$A^{(10000)}$	–1	–1	0	0	0	0	–1	0	1

The results of steps 4. 1 and 4. 2 in the **Table 1** contains a description of the process of selecting an i_{10000} object by users, as well as the process of forming ratings for this object. For example, according to the results of step 4.1, for the second subset Δ_2, only one temporal rule $r_{1,2}^{(10000)} = \Phi_1 X \Phi_2$, connecting it with the previous subset is valid. The score value $W_2^{10000,item} \in M^{10000,item}$ of 0.19 is the normalized weight of this rule and shows the increase in sales by Δ_2 compared to Δ_1. The indicator $W_4^{10000,item} = -0.41$ reflects the general drop in sales by Δ_4 compared to Δ_1, Δ_2, and Δ_3. The sign $a_2^{(10000)} \in A^{(10000)}$ has a negative value and therefore indicates a possible attack aimed at reducing sales. This sign $a_{10}^{(10000)}$, in turn, indicates a possible attack aimed at increasing the rating. The indicators of the rating dynamics for Δ_5 and Δ_6 were formed using the rules for Δ_4, since the rating for the i_{10000} object was not presented in the indicated subsets.

The experiment results with the selected object make it possible to analyze approaches to masking an attack. Revealing a method of concealing a possible attack is carried out based on a comparison of the signs of a shilling attack: $d_s^{(10000)}$ and $a_s^{(10000)}$, as shown in **Fig. 3.** On Δ_2 and Δ_3, there is a mismatch between the increase in sales and the simultaneous decrease in the rating compared to Δ_1. Since the magnitude of the discrepancy has not changed: $d_2^{(10000)} = d_3^{(10000)} = 0.28$ then the probable attack occurred in interval Δ_2. A disparity of less than 50 % of the maximum possible value indicates the possibility of concealing an attack, taking into account the average rating of existing items. The value $a_2^{(10000)} = -1$ indicates a possible shilling attack to downgrade the rating.

There is also a deviation $d_5^{(10000)}$ on the interval Δ_5. This deviation is not a sign of an attack $\left(a_5^{(10000)} = 0\right)$, since the estimates $W_4^{10000,item}$ and $W_5^{10000,item}$ coincide. The coincidence of ratings means that users did not put ratings in the subsequent interval Δ_5.

The discrepancy $d_8^{(10000)}$ describes the dynamics of the user's choice in the intervals from Δ_2 to Δ_8 inclusive. This deviation indicates a possible shilling attack at one of these intervals.

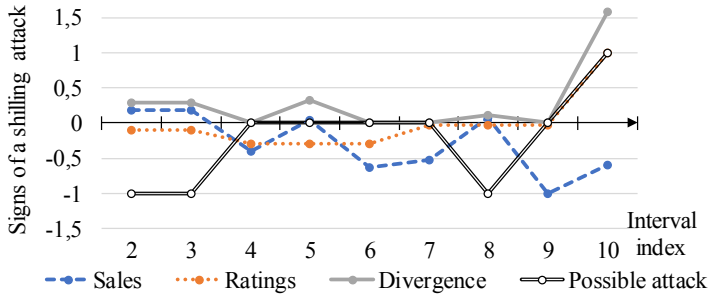

Fig. 3. Signs of a shilling attack

However, since $d_8^{(10000)}$ it covers all previous deviations and has a value less than $d_2^{(10000)}$, the probable attack occurred on interval Δ_2.

The discrepancy $d_{10}^{(10000)} = 1.59$ indicates a possible attack on the intervals from Δ_2 to Δ_{10}. Since this value is significantly higher than $d_2^{(10000)}$ and $d_8^{(10000)}$, the probable attack occurred on the interval Δ_{10}. The value $a_{10}^{(10000)} = 1$ indicates a possible shilling attack to increase the rating. A discrepancy $d_{10}^{(10000)}$ greater than 50 % of the maximum indicates possible concealment of an attack using popular items. A check was carried out on popular properties (with a large number of purchases) to confirm this. For example, the popular i_{10} has been selected over 91,000 times. The rating of this object increases on sets Δ_9 and Δ_{10}, which confirms the hypothesis about masking the attack using the rating of popular items.

Thus, this method allows to identify ways to mask an attack in conditions of incomplete rating data using data ordered on an "earlier-later" timeline.

The second phase of the experiment is devoted to comparing the proposed method's effectiveness with those of [11, 12]. These methods perform splitting into intervals under the condition of a sharp change in rating values over a limited period. The accuracy evaluation was used for comparing the methods. The accuracy evaluation is defined as the number of detected attacks to the number of all attacks. Attacks have been generated to downgrade (nuke attacks) and increase (push attacks) ratings of three targeted items. In the first case, the rating was set equal to zero, and in the second – equal to 5. These attacks were included in the set of Q ratings. Previously, in accordance with the algorithm in **Fig. 2**, the division into intervals was performed using evaluation of changes in user preferences for three books. The interval with the maximum weight value (4), taking into account rounding, was 9000 purchases. The results of the second phase of the experiment are shown in **Fig. 4, 5**.

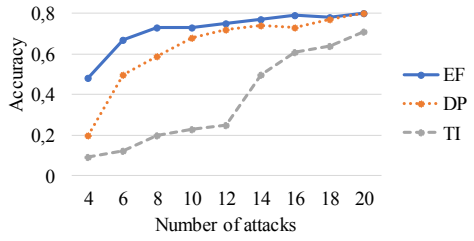

Fig. 4. Comparison of the detection accuracy of nuke attacks for different methods

The method [12] for detecting anomalies based on dynamic division for time series is represented by the abbreviation DP. The method [11] for detecting anomalous elements based on time intervals is represented by abbreviating TI. The developed method is represented by the abbreviation EF (Explicit Feedback).

At the initial stage of rating falsification, when forming up to 10 attacks to increase the rating (push attacks), the developed EF method allows increasing the accuracy from 8 % to 23 % compared to the DP method, and by more than 30 % compared to the TI method. In nuke (downgrade) attacks, the increase in accuracy at the initial stage ranged from 5 to 23 % compared to the DP method and over 30 % for the TI method. However, in the future, as the number of attacks increases, DP, TI methods show similar or higher accuracy compared to EF. Such characteristics determine that the scope of the developed method is the initial stage of actions of an attacking user.

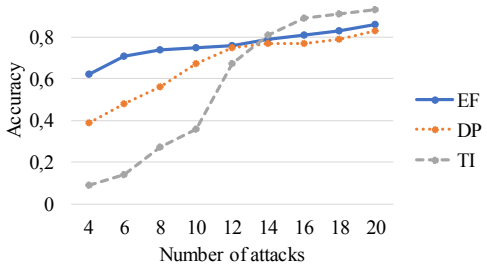

Fig. 5. Comparison of the accuracy of detecting push attacks for different methods

5. Discussion of the results of developing a method for detecting shilling attacks

The result of the work is a method for identifying shilling attacks of users based on a quantitative assessment of inconsistencies in the temporal characteristics of the purchasing processes and rating, reflecting the results of objective and subjective feedback from the user. The method uses temporal rules (1), which describe the change in sales or ratings of the same object for arbitrary pairs of time intervals. A set (2) of such rules describes the general temporal characteristics of the processes of selecting a target product and setting its ratings.

The method builds models (5) of the processes of changing user preferences, using weighted temporal rules. The difference from interval methods [11, 12] is that the signs of shilling attacks are revealed by comparing the elements (4) of the models of processes of purchasing and forming ratings. The result of the method is a list of possible attack intervals for a given list of users.

The difference between the proposed method and [13] in detecting a shilling attack consists in the use of a set of temporal rules that associate all previous intervals with the current one (**Fig. 1**), which makes it possible to obtain a generalized estimate of the discrepancies between sales and rating.

The advantage of the method is the ability to identify discrepancies between the dynamics of purchases and ratings at the initial stages of the attack (**Fig. 4**). Also, in the case of the incompleteness of the latter (**Table 1**). When implemented iteratively, the method allows a detailed list of potential attackers. Comparative analysis of the attack indicators obtained as a result of the method execution allows to classify the used approaches to attack masking (**Fig. 2**).

The method allows increasing the accuracy by at least 8 % for attacks to increase the rating (**Fig. 5**) and 5 % in attacks to downgrade (**Fig. 4**) at the very beginning of the actions of attacking users, with a small number of attacks. Therefore, it is advisable to use the method online in order to identify changes in the behavior of attacking users quickly.

The disadvantage of this method is that the detection of an attack does not take into account the possible partial time shift between purchases and ratings. The bias is due to the fact that some users give a rating after a purchase with a delay. The effect of this bias depends on the granularity of time. The longer the length of the time intervals, the less the impact of the rating lag.

The method imposes a time ordering constraint on the input. The additional requirement for timestamps in the input allows to identify the intervals of the attack more accurately.

The developed method is intended for iterative refinement of both the time intervals for carrying out shilling attacks and the list of possible attackers under conditions of periodic changes in real users' interests, including online. The method forms a process description of user actions and, therefore, can be applied to identify abnormal situations in process-oriented systems after adaptation.

Further development of the method is associated with the use of rules with the temporal operator "Until," which will allow formalizing the change of interests of user groups as a result of external events, for example, large presentations of new goods and services. Combining the considered temporal rules will make it possible to take into account both periodic and seasonal changes in user requirements when detecting shilling attacks.

6. Conclusions

1. A temporal model of changing the preferences of the users of the recommender system has been developed. The model is characterized by using a time-ordered sequence of numerical evaluations of changes in user preferences for the target item. Each evaluation specifies the change for the current interval concerning the previous time intervals. The model allows to estimate the increase or decrease in the user's interest in the target subject in the current range compared to all previous time intervals. A comparison of models obtained based on explicit and implicit feedback makes it possible to identify attacks aimed at artificially changing items' ratings.

2. A method for detecting shilling attacks based on a comparison of the processes of changing user preferences in sales and rating assignments for a given period is proposed. This method differs from the existing ones by using a temporally ordered set of generalized evaluations of the discrepancies between these processes to detect shilling attacks.

The proposed method allows for detecting shilling attacks at the initial stage of rating falsification in the absence of ratings at certain intervals within a given period. The method allows

identifying approaches to masking an attack taking into account the current combination of rating values and attack indicators, as well as iteratively refining the list of fake user profiles.

3. Experimental evaluation of the method was carried out on data sets obtained as a result of implicit and explicit communication from users. The experimental results showed an increase in the detection accuracy at the initial stages of a push attack by at least 5 %, and a nuke attack by at least 8 %, which makes it possible to apply the method in the online mode of the recommender system.

References

[1] Aggarwal, C. (2016). Recommender Systems. Springer, 498. doi: https://doi.org/10.1007/978-3-319-29659-3

[2] Hallinan, B., Striphas, T. (2014). Recommended for you: The Netflix Prize and the production of algorithmic culture. New Media & Society, 18 (1), 117–137. doi: https://doi.org/10.1177/1461444814538646

[3] Adomavicius, G., Bockstedt, J., Curley, S., Zhang, J. (2014). De-Biasing User Preference Ratings in Recommender Systems. Proceedings of the Joint Workshop on Interfaces and Human Decision Making for Recommender Systems co-located with ACM Conference on Recommender Systems, 2–9. Available at: http://ceur-ws.org/Vol-1253/paper1.pdf

[4] Gunes, I., Kaleli, C., Bilge, A., Polat, H. (2012). Shilling attacks against recommender systems: a comprehensive survey. Artificial Intelligence Review, 42 (4), 767–799. doi: https://doi.org/10.1007/s10462-012-9364-9

[5] Adomavicius, G., Bockstedt, J. C., Curley, S. P., Zhang, J. (2013). Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects. Information Systems Research, 24 (4), 956–975. doi: https://doi.org/10.1287/isre.2013.0497

[6] Mobasher, B., Burke, R., Bhaumik, R., Williams, C. (2007). Toward trustworthy recommender systems: an analysis of attack models and algorithm robustness. ACM Transactions on Internet Technology, 7 (4), 23. doi: https://doi.org/10.1145/1278366.1278372

[7] Wang, Y., Qian, L., Li, F., Zhang, L. (2018). A Comparative Study on Shilling Detection Methods for Trustworthy Recommendations. Journal of Systems Science and Systems Engineering, 27 (4), 458–478. doi: https://doi.org/10.1007/s11518-018-5374-8

[8] Patel, K., Thakkar, A., Shah, C., Makvana, K. (2016). A State of Art Survey on Shilling Attack in Collaborative Filtering Based Recommendation System. Smart Innovation, Systems and Technologies, 377–385. doi: https://doi.org/10.1007/978-3-319-30933-0_38

[9] Zhou, W., Wen, J., Gao, M., Ren, H., Li, P. (2015). Abnormal Profiles Detection Based on Time Series and Target Item Analysis for Recommender Systems. Mathematical Problems in Engineering, 2015, 1–9. doi: https://doi.org/10.1155/2015/490261

[10] Wang, Y., Qian, L., Li, F., Zhang, L. (2018). A Comparative Study on Shilling Detection Methods for Trustworthy Recommendations. Journal of Systems Science and Systems Engineering, 27 (4), 458–478. doi: https://doi.org/10.1007/s11518-018-5374-8

[11] Gao, M., Yuan, Q., Ling, B., Xiong, Q. (2014). Detection of Abnormal Item Based on Time Intervals for Recommender Systems. The Scientific World Journal, 2014, 1–8. doi: https://doi.org/10.1155/2014/845897

[12] Gao, M., Tian, R., Wen, J., Xiong, Q., Ling, B., Yang, L. (2015). Item Anomaly Detection Based on Dynamic Partition for Time Series in Recommender Systems. PLOS ONE, 10 (8), e0135155. doi: https://doi.org/10.1371/journal.pone.0135155

[13] Chala, O., Novikova, L., Chernyshova, L. (2019). Method for detecting shilling attacks in e-commerce systems using weighted temporal rules. EUREKA: Physics and Engineering, 5, 29–36. doi: https://doi.org/10.21303/2461-4262.2019.00983

[14] Levykin, V., Chala, O. (2018). Method of determining weights of temporal rules in Markov logic network for building knowledge base in information control systems. EUREKA: Physics and Engineering, 5, 3–10. doi: https://doi.org/10.21303/2461-4262.2018.00713

[15] Chalyi, S., Leshchynskyi, V., Leshchynska, I. (2019). Method of forming recommendations using temporal constraints in a situation of cyclic cold start of the recommender system. EUREKA: Physics and Engineering, 4, 34–40. doi: https://doi.org/10.21303/2461-4262.2019.00952

[16] Chalyi, S., Pribylnova, I. (2019). The method of constructing recommendations online on the temporal dynamics of user interests using multilayer graph. EUREKA: Physics and Engineering, 3, 13–19. doi: https://doi.org/10.21303/2461-4262.2019.00894

[17] Sergii, C., Ihor, L., Aleksandr, P., Ievgen, B. (2018). Causality-based model checking in business process management tasks. 2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT). doi: https://doi.org/10.1109/dessert.2018.8409176

[18] Zajac, Z. (2017). Goodbooks-10k: a new dataset for book recommendations. FastML. Available at: http://fastml.com/goodbooks-10k-a-new-dataset-for-book-recommendations/

Received date 10.03.2020
Accepted date 13.08.2020
Published date 30.09.2020