Asia-pacific Journal of Convergent Research Interchange
Vol.6, No.8, August 31 (2020), pp.149-166
http://dx.doi.org/10.47116/apjcri.2020.08.14

Research on Online Detection Method of E-commerce False Comments Based on LBP

Kokula Krishna Hari Kunasekaran[1)], Lili Han[2)]

Abstract

With the rapid rise of e-commerce platforms, how to dig out false comments quickly and efficiently has become an urgent problem to be solved. This paper innovatively proposes a practical and effective online detection model for false reviews. Different from the previous offline detection methods, this model uses a sliding time window mechanism to realize the function of real-time monitoring and can dig out the relevant information of false comments in a short time. By extracting the review data in each time window, the measurement method of review similarity is defined, so that the relationship between reviews is quantified and the related review graph is established. Then the comment node is associated with random variables, and the comment graph is modeled as a Markov random field to generate the false comment online detection algorithm. The experimental results show that the accuracy rate and recall rate of the algorithm proposed in this paper have been improved.

Keywords: LBP, Markov Random Field, False Comments, E-commerce

1. Introduction

With the rapid development of Internet technology, more and more online shopping platforms and e-commerce websites come into people's lives. These online shopping platforms have their own comment systems. Users can not only publish their own use of products on the platform, but also evaluate the service attitude, product description and logistics quality of the store. At the same time, online reviews have become more and more popular as a kind of social media. The main performance is that consumers will take the historical review record of the commodity as an important reference index when they purchase a commodity. When consumers buy a product on a shopping website, they will pay attention to the overall evaluation information of the product. When consumers find that most of the products they

Received(April 11, 2020), Review Result(1st: June 8, 2020, 2nd: July 23, 2020), Accepted(July 27, 2020)

1) (Association of Scientist) Developers and Faculties (ASDF), United Kingdom
email: prm@kokulakrishnaharik.in
2) (Professor, Corresponding Author) College of Information Science Technology, Chengdu University of Technology, Chengdu, China
email: yuwang87@163.com

ISSN: 2508-9080 APJCRI

want to buy are highly praised, they will enhance their desire to buy and increase the possibility of consumers to buy products. If we find that the product evaluation information tends to be poor, it will reduce the purchase desire and reduce the possibility of consumers to buy products[1].

Driven by interests, every business tries to improve the praise of their products to attract more consumers. Some businesses will hire some people to make false comments on their products in order to improve the popularity of the store and the praise of the products. Some businesses may employ some people to suppress the products of their peers maliciously to improve the competitiveness of their stores. Similarly, driven by the interests, the part-time job with false orders will be produced. A fake order maker can complete an order quickly and conveniently at any time and anywhere just by using a computer or mobile phone. This kind of short time, paid false order work, attracted a large number of people who want to make money part-time.

Therefore, in order to maintain the good operation mechanism of e-commerce platform and ensure fair competition between businesses. In order to provide consumers with a good shopping experience and real information shopping environment. False comment detection has become an important research direction in the field of e-commerce.

2. Related Research

The concept of false comment detection was first proposed by famous American professor Bing Liu and his research team in 2008[2]. Many professors and scholars have devoted themselves to the research of false comment detection, which makes false comment detection become a hot research direction in the field of data mining at home and abroad. In recent years, with the application of artificial intelligence and deep learning in various research fields, false comment detection based on machine learning has become one of the hot research issues. A lot of research work has been done on the problem of false comment detection at home and abroad. Next, for different false comment detection methods, this paper classifies them from different angles, and analyzes the classical methods of each classification in detail.

2.1 Main Technology

2.1.1 False Comment Detection

False comment detection refers to the wrong, false, irrelevant and unreal comments found in the comment dataset. Before the detection of false comments, due to the huge amount of comment data in the data set, the diversity of comment languages and the unclear intention of comments, we can not quickly and intuitively distinguish the false comments from the real comments in the data sets. Therefore, it is an urgent problem to identify the false comments by using machine learning methods and evaluate and detect them with relevant algorithms[1].

2.1.2 False Reviewer Detection

Fake reviewer detection is based on the user's historical shopping product ratings and review text to dig out the user's shopping behavior. The user's historical shopping behavior is analyzed to identify false reviewers. Then based on "all comments made by false reviewers" The principle of "all false comments" achieves the purpose of detecting false comments. Lim et al.[3] make full use of review scoring data to construct fraud characteristics of false reviews from the three aspects of single product, product group, and review scoring deviation. And select the optimal feature according to the experimental results obtained from the corresponding characteristics. Combine, calculate the fraud probability of the reviewer. However, the disadvantage of this strategy is that only false reviewers with obvious fraudulent behavior can be found, and it is not easy to detect false reviewers with strong disguise and dynamic deception methods. Paper[4] also clustered according to the abnormal behavior of reviewers, and used the behavior characteristics of reviewers to mine false reviewers.

2.1.3 False Comment Group Detection

The false comment group considers the group cheating among reviewers, and digs out the false comments according to the characteristics of the group cheating. For example, the phenomenon of "false order", a large number of reviewers publish false comments on the specified products under the organized and clear purpose. The detection of false comment group is to find out the similarities among multiple reviewers by analyzing a large number of similar comment texts, the number of comments on the same product increases suddenly, and a large number of comment data with great difference from the existing reviews under the same product. Then, the corresponding algorithm is developed to mine the relevant false comments.

Mukherjee et al.[5] used false comment groups to mine false comments. This method uses the deviation degree between the comment score and the average score of the existing comments and the similarity degree of the comment content to measure the detection of false comment groups. Wang et al.[6] proposed a GSLDA algorithm model based on group fraud

detection. According to the growth of comment data in a short period of time and the similarity of comment data in a short period of time, the corresponding reviewer graph is generated, so as to effectively mine the groups of false reviews.

2.2 The Main Problems

Since the concept of false comment was first proposed, false comment detection has been developed and improved rapidly through the efforts of researchers in the field of machine learning. Researchers continue to propose new false comment detection methods to improve the accuracy and universality of false comment detection. In order to be able to promote to the highly economic value of the business field, let the development of e-commerce platform more equitable, more benefit to the majority of consumers. It makes consumers have more objective review reference and better shopping experience. However, the existing false comment detection methods have not achieved significant detection results, the main reasons are as follows:

2.2.1 Limitations of Existing Methods for Detecting False Comments

Most of the existing false comment detection technologies are offline detection. As we all know, offline detection requires research on a large amount of historical data. The existence of data has a long time span, which causes problems such as low real-time and low practicality for mining false comments. At the same time, offline detection is difficult to promote in practical applications because it cannot detect false comments in time and take timely measures.

2.2.2 Manual Tagging Requires a Lot of Work

To make the machine learning method have a good fitting effect and avoid the occurrence of overfitting due to the small data set, it is particularly important to obtain a data set with sufficient data. Generally, data sets that can achieve research purposes often have tens of thousands or even hundreds of thousands of original data. In addition, false comment detection is a supervised learning task, which requires the label of each piece of training data in order to use machine learning methods to achieve the purpose of classification. In summary, in order to meet the huge amount of data in the data set and the need for labeling, if you only rely on manual labeling of data labels, the time and energy spent on each researcher is unpredictable[7].

2.2.3 The Criteria for Judging False Comments are not Clear

Compared with other classification detection problems, the false comment detection task has

great subjective arbitrariness. Every consumer's comment is judged as false or true, only depending on the subjective thinking of the judge and his personal experience. Given a data set based on reviews of products and with review text for manual evaluation, the judges can easily identify those with obvious advertisements, information that has nothing to do with the reviewed products, and highly repetitive review information. However, other review information is different from the subjective feelings of each reviewer or the criteria for defining false reviews. The resulting evaluation results are also different from person to person, and there is no clear distinction between right and wrong.

2.2.4 Fake Comments Cover a Wide Range of Fields

False comments involve many fields, such as clothing, cosmetics, food and so on. In many fields, the language features, comment methods and fraud methods used by false comments have different fraud features[8]. The diversity of fraud features is closely related to specific fields, which makes the implementation of false comment detection very difficult.

2.2.5 False Comment Fraud Strategy is Complex and Changeable

The comment methods and deception methods of false commentators are varied. False commentators will dynamically adjust their fraud strategies according to the anti fraud detection strategies of specific e-commerce platforms, so as to avoid the detection and screening of their false comments by e-commerce platforms[9][10]. Driven by interests, it is difficult to achieve a dynamic balance between fraud and anti-fraud. This phenomenon will continue to develop and evolve, resulting in the situation that false commentators and anti-fraud system of e-commerce platform are fighting wits and courage. This makes the anti-fraud system need to quickly grasp the behavior of false commentators.

In summary, online false comment detection has gained more and more attention, but there is still a lack of effective models and methods for online false comment detection[11]. The effect of existing online false comment detection strategies It has not yet reached the expectations and recognition of consumers and e-commerce platforms. With the continuous evolution of false comment fraud strategies and the continuous expansion of the field of false comments, the task of detecting online false comments will become more difficult.

3. Theoretical model

3.1 The Basic Concept of Markov Random Field

Markov Random Field (MRF) is a generative model, that is to say, the core of the model is the joint probability distribution of variables. As a graphical model described using undirected graphs, MRF is a method of describing the joint probability distribution on a variable X, representing a set of conditional independent relationships. Therefore, MRF can be regarded as a set of independence assumptions that define a series of graph structures.

Conditional Independence defines the graphical semantic representation of probability distribution, so the conditional independence hypothesis can be determined by dividing a single graph. Suppose that there are three groups of nodes in an undirected graph, which are represented as a, B and C respectively. Conditional independence can be considered according to formula (1). In order to determine whether the probability distribution defined by graph satisfies this property, all possible paths connecting nodes of set a and node of set B need to be considered. If all these paths pass through one or more nodes in set C, then all paths are "blocked", so conditional independence is true.

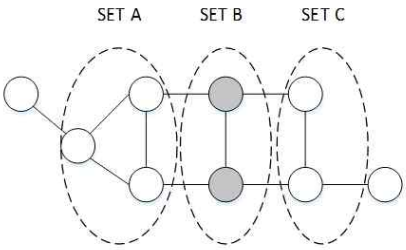
$$A \perp B | C \tag{1}$$

Given the current state and all the past states, the conditional probability distribution of the future state of a stochastic process only depends on the current state. In other words, when the current state is known, the stochastic process and the historical path of the process are conditionally independent, then the process can be called a stochastic process with Markov property.

Here we need to introduce the concept of Markov blanket. In an undirected graph, a collection of multiple adjacent nodes can form a Markov blanket of node. Its properties are mainly manifested as: the conditional probability distribution of is only related to the variables in the Markov blanket. This conclusion is based on the premise of all other remaining variables in the given graph. In other words, a node only has a conditional dependence relationship with its neighboring nodes, and is conditionally independent from any other nodes.

Suppose there is a node set C in undirected graph G, which separates any two node sets a and B, as shown in [Fig. 1]. The set of random variables corresponds to the node set A, B, C respectively. Global Markov means that the random variables and are conditionally independent. This assumption is based on the premise that the set of random variables is given. Its expression is shown in formula (2).

Asia-pacific Journal of Convergent Research Interchange
Vol.6, No.8, August 31 (2020), pp.149-166
http://dx.doi.org/10.47116/apjcri.2020.08.14



[Fig. 1] Set Partition of Glob Al Markov Property

$$p(x_A,\ x_B|x_C)= p(x_A|x_C)p(x_B|x_C) \tag{2}$$

Given an undirected graph, a set of random variables indexed by, if it satisfies the local Markov property, it can form a Markov random field about G.

3.2 Markov Random Field

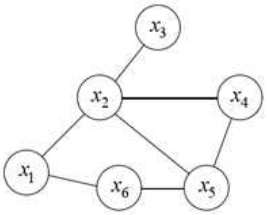
In the field of physics and probability, Markov random field (MRF) or Markov network is a group of random variables with Markov characteristics described by undirected graph. In other words, the necessary and sufficient condition for a random field to be a Markov random field is that the random field satisfies the Markov property.

Markov random field is similar to Bayesian network in the representation of correlation. The difference is that Bayesian network is directed and acyclic, while Markov network is undirected and may have rings. Therefore, Markov network can represent certain dependencies that Bayesian networks cannot represent (such as circular dependencies); on the other hand, it cannot represent certain dependencies that Bayesian networks can represent (such as induced dependencies). The basic graph of Markov random field can be finite or infinite.

In a Markov random field, all nodes satisfy the probability distribution characteristics of a node and do not involve any node outside its neighborhood, but are only closely related to the Markov property of nodes in its neighborhood. In this neighborhood model, each node in the random field only interacts with its directly adjacent nodes.

Markov random field is a probabilistic undirected graph model in which each node in the graph represents one or a group of variables, and the edge between nodes represents the dependence between two variables. Among them, the nodes in the figure can be divided into two types, one is the hidden node used to represent unknown information, and the other is the observable node used to represent known information. Next, use the likelihood function to represent the influence coefficient between the hidden node containing unknown information

and the observable node containing known information, and use the potential function represents the influence coefficient between two adjacent hidden nodes connected by an edge. [Fig. 2] shows a simple Markov random field. The edges in the figure indicate that the nodes have a mutual relationship, which is bidirectional and symmetric. For example, there is an edge connection between and which means that and have a correlation, and this correlation can be measured by a potential function. For example, the potential function can be defined according to formula (3), indicating that the model preference variable and have the same value, in other words, in this model, the values of and are positively correlated. Potential function describes the correlation between local variables, mainly used to define the probability distribution function, and is non-negative. In order to satisfy strict non-negativity, the potential function is usually defined as an exponential function as shown in formula (4). H(x) is called an energy function, and it usually exists as a real-valued function form based on the variable x. The common form is shown in equation (5). The joint probability distribution is defined as the product of the potential function, so the total energy can be obtained by adding the energy of each largest clique. Among them, and are the parameters that need to be learned, called parameter estimation.



[Fig. 2] Simple Markov Random Field

$$\Psi(x_2,x_3)=\begin{cases}1.5 & \text{if } x_2=x_3\\ 0.1 & \text{if } otherwise\end{cases} \tag{3}$$

$$\Psi(x)=e^{-H(x)} \tag{4}$$

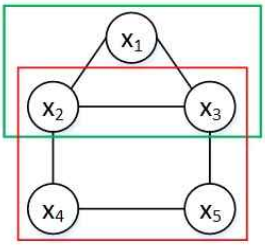
$$H(x)=\sum_{u,v\in x,u^{'}=v}\alpha_{uv}x_u x_v+\sum_{v\in x}\beta_v x_v \tag{5}$$

3.3 The Traditional Solution of Markov Random Field

For the solution process of undirected graph model with given probability, a better solution is

to decompose the joint probability of the whole into factors. In other words, the joint probability of several subsets is multiplied to represent the overall probability distribution, which reduces the complexity and facilitates the model learning and probability calculation. In fact, the biggest characteristic of probability undirected graph model is the decomposability of probability distribution. According to pairwise Markov property, if there are two nodes and , they are not connected, and the two nodes must be conditionally independent. The premise of this conclusion is that all the other nodes in the graph are given. Therefore, in order to realize that all possible probability distributions in the graph meet the conditional independence assumption, it is necessary to factor decompose the joint probability distribution so that and appear in different factors.

The concept of clique is introduced to prevent non-adjacent nodes from appearing in the same factor. It is defined as a subset of the nodes in the graph, and there are interconnected edges between any pair of nodes in this subset, in other words, each node in the set of nodes that is a factor of the clique All are fully connected. The largest clump (or extremely large clump) means that there is no node outside the clump that can be added to the clump without destroying the nature of the clump. This means that the maximal clump is as good as possible to satisfy that there are interconnected edges between any nodes inside the clump.



[Fig. 3] Nodal Undirected Graph and Clique

The schematic diagram of clique and maximal clique is shown in [Fig. 3]. Among them, the green circle is a large group, and the red circle is a group. The graph contains six cliques with two nodes, namely and a maximal clique, while the set is not a clique because there is no edge connection and Obviously the simplest clique is two nodes and one edge. In the previous section, the potential function is defined for the correlation between two nodes (each edge). Therefore, in Markov random field, the joint probability distribution of multiple variables can be decomposed into the product of multiple potential functions according to clique, and each clique corresponds to a potential function. The expression of joint probability distribution of probability undirected model is shown in equation 6.

$$\begin{cases} p(x)=\frac{1}{Z}\prod_{c}\Psi_c(x_c) \\ Z=\sum_{x}\prod_{c}\Psi_c(x_c) \end{cases} \tag{6}$$

Among them, assuming that C is a maximal clique of an undirected graph, the potential function (random function) corresponding to the node of the clique C is defined as, and the clique C defines as a strictly positive function, and is based on all the undirected graphs The maximal clique realizes the corresponding potential function multiplication. And Z is the normalization constant to ensure that is the correctly defined probability. However, if a potential function is defined for each edge in the Markov random field, the model will have too many potential functions, which will inevitably increase the computational burden. For example, in [Fig. 2] need to define three potential functions respectively, but from the figure we can see that are connected by edges, and their values will affect each other, x as a maximal clique, if any function is defined on the maximal clique, then the other factors defined on a subset of these variables are redundant, so the overall Consider, define a potential functionrepresent the preference of the three values. The joint probability distribution can be decomposed into the product of the potential functions on the maximal clique according to formula (7). Among them, C∗ is a set of extremely large cliques. For example, in [Fig. 2], and its joint probability distribution p(x) is defined as shown in formula (8).

$$\begin{cases} p(x)=\frac{1}{Z^*}\prod_{Q\in C^*}\Psi_Q(x_Q) \\ Z=\sum_{x}\prod_{cQ\in C^*}\Psi_Q(x_Q) \end{cases}\# \tag{7}$$

$$\begin{cases} p(x)=\frac{1}{Z}T_{12}*T_{16}*T_{23}*T_{56}*T_{245} \\ T_{ij}=\Psi_{ij}(x_i,x_j) \end{cases} \tag{8}$$

However, for most small Markov random fields, in order to solve the edge probability distribution of each hidden node with unknown information, the total probability formula can be directly calculated by summation or integration. However, if this method is directly applied to the Markov random field with a large number of nodes, the computational complexity of the probabilistic solution will increase exponentially. Therefore, in order to reduce the computational complexity of probability solution, we must introduce some calculation methods

which can speed up the solution.

3.4 The main idea of LBP algorithm

For each node in the Markov random field, the probability distribution state of the node can be transmitted to the adjacent node through the message propagation mechanism, and then affect the probability distribution state of the adjacent node. In this way, the probability distribution of each node will converge to the stable state after a certain training period. In other words, a node determines the final belief distribution by listening to the opinions of its neighbors. The message enters the network at the observed node and propagates throughout the network. Adjacent nodes exchange messages to tell each other how to update beliefs according to prior and conditional probabilities. Through this operation, information is continuously transmitted until the belief state is as stable as possible.

In Markov random field, the edge probability distribution of a node is defined as the confidence degree of the point obtained by the confidence propagation algorithm. The probability value obtained in this way is very approximate, if not very accurate. The theory shows that in Markov random field, only the number of iterations equivalent to the distance between two boundaries can be passed. The information of one boundary point can reach another boundary. Therefore, with the increase of Markov random field, the amount of calculation required to calculate the probability distribution of nodes only shows a linear growth trend, and the calculation amount required by this method is far less than that of brute force algorithm[12-14].

LBP is a widely used MRF approximate inference algorithm. In order to use LBP algorithm to realize message propagation in Markov random field, we need to define the concepts of message and confidence[15][16].

In LBP, a node sends messages to its neighbors iteratively until all messages become static and tend to be stable. Let denote the message that the label passes from node to node. The message passing formula or message update rule is shown in equation (9). Where is a set of adjacent nodes of and is a normalization constant. Equation (9) means that the message passed from node i to node j is proportional to the sum of the product of the node potential of node i, the joint potential between and, and all messages passed from the neighbors of node i other than node j to node i for each. The above message update equation is called sum product LBP.

$$m_{i\to j}(x_j)=\frac{1}{Z}\sum_{x_i\in L}\Psi(x_i)\Psi(x_i,x_j)\prod_{X_k\in N_i}m_{k\to i}(x_i) \tag{9}$$

In the initial stage of LBP, all messages are set to 1. Then LBP uses formula (10) to iteratively update each node. When LBP converges, the final confidence level of node with label can be calculated by formula (10). Where is the normalization constant.

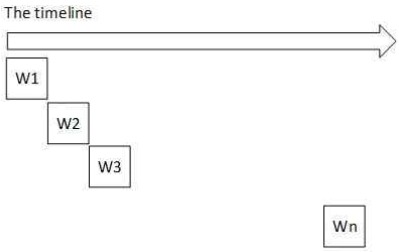
$$b_i(x_i)=\frac{1}{Z_2}\Psi(x_i)\prod_{X_k\in N_i}m_{k\to i}(x_i) \tag{10}$$

4. False Comment Detection Model

4.1 Principle of Sliding Window

In the existing research on false reviews, most of the detection algorithms are based on the entire data set. Normally, a data set that can achieve research purposes needs to have hundreds of thousands of data, spanning decades. In this time span of experimental data, even if the detection method has a good effect on mining false comments, due to the low real-time detection, the false comments and their false commenters cannot be mined in time, which cannot satisfy the requirements of today's e-commerce platforms. improve dramatically. This article uses the time window to segment the experimental data set, so that the time window slides along the time axis of the data set. Each time only the data in the time window is studied and analyzed, and only the comment related information in the time window is considered. To a certain extent, the influence of historical data on the current time window is avoided, and the real-time nature of mining false comments is greatly improved.

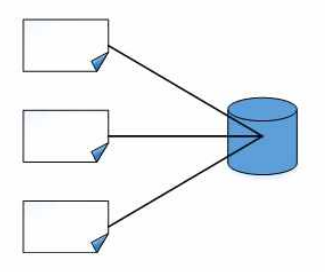
This article uses a non-overlapping sliding time window, as shown in [Fig. 4]. The figure shows an example of the sliding time window used by the comment data set with a timestamp of 2000-3000, where the time window size is set to T=100, then the comment data contained in the time window w1 is all within the timestamp 2000-2100 After that, only the comment data with a time span of 100 will be modeled and calculated. The time window w2 means that only all the comment data within the timestamp 2100-2200 will be modeled and calculated, and so on, until the The entire timeline is executed.



[Fig. 4] Time Window Mechanism

4.2 The Concept of Bipartite Graphs

In the field of false reviews detection, existing research methods generally use graph structures to visually describe the connection between the review collection and the product collection[17][18]. Among them, the most widely used structure is the bipartite graph. It is usually assumed that the undirected graph G = (V, E), where vertex V can be divided into two subsets of vertices (A, B) without any intersection, and the two disjoint vertex sets A and B are extracted respectively Vertices I and j make up each edge (I, j) of the graph. Usually a type of graph with the shape of G is called a bipartite graph.



[Fig. 5] Bipartite Graph Structure

In the review data within the time window, the relationship between the review and the product will form a bipartite graph structure through the connection of the reviewer, as shown in [Fig. 5] The online detection method for false product reviews proposed in this paper also uses a bipartite graph structure, but the bipartite graph structure is generated by review data within a time window, rather than the entire review data set.

4.3 Generate Dynamic Comment Graph

The process of generating comment graph is essentially the process of determining the edges

between nodes. If there is an edge between a node and other nodes, then there is a dependency relationship between the two nodes. On the contrary, if there is no edge between the node and all other nodes, that is, the node is an isolated node, then it can be ignored in the subsequent modeling process. This paper is to model the comments, so the node is the comment node in the time window. The first task is to determine whether the edge exists between the two comment nodes. Referring to the relevant information in the past, there are two main modeling methods of comment graph: GSBP model and GSBC model.

The GSBP model uses reviewers as nodes, and the determination of the edges between nodes is based on the number of products reviewed by two reviewer nodes as a measurement index to measure the intimacy of two reviewers, and then generate a review graph G = (V, E), where V is the set of reviewer nodes, and E is the set of edges between reviewer nodes. The higher the calculated similarity between two reviewers, the greater the intimacy between the two reviewers and the stronger the dependence.

4.4 Build Markov Random Field

Markov random field is a probabilistic undirected graph model, which is usually used to model a group of random variables with Markov attributes described by undirected graph. Each node is associated with a random variable, which can be in one of a limited number of States, that is, classification category. Local MRF is a random field which satisfies the local Markov property, that is, it is assumed that the random variable only depends on its neighbors and is independent of other nodes except neighbors.

By associating the comment nodes with random variables, the comment graph is modeled as Markov random field. Let L = {1, − 1} denote a node label, 1 denote that the node is a fake comment, and - 1 means that the node is a real comment. Therefore, given the observable comment graph G and the prior of the corresponding comment nodes, the problem of detecting false comments can be reduced to the classification of Markov random fields.

In the process of establishing Markov Random Field, how to determine the potential of each node is very important. One is the no prior method, which sets the prior value of each node to the same value. The other is a priori method that uses certain characteristics of reviews to predict the cheating degree of each review, so that each review node has a different potential. Due to the differences of reviews, this paper adopts a priori method and combines the following features to determine the priori of the reviewer node in Markov Random Field.

Asia-pacific Journal of Convergent Research Interchange
Vol.6, No.8, August 31 (2020), pp.149-166
http://dx.doi.org/10.47116/apjcri.2020.08.14

5. Experiments and Results

In order to ensure the authenticity and validity of the experimental data, a data set from the actual scene is needed to detect the false comments. This paper uses yelp NYC, a shared dataset shared by Dr. shebuti rayana of the United States. The data set is from the famous merchant review website in the United States Yelp.com It was founded in 2004. Consumers in the site after consumption, can be directly for consumer goods and businesses to score. The data set contains the review information of businesses located in the New York area of the website, including restaurants, hotels, tourist attractions, amusement parks, shopping centers and other fields.

Each line in the dataset represents a comment, and the information contained in it is the reviewer number, product number, rating, label, and comment date. Among them, reviewer number, product number and comment time are in the form of string.

After preprocessing the data set, we conduct a brief understanding and analysis of it so that we can have a deeper understanding of our process at each step during the experiment. The Yelp NYC dataset contains a total of 359,052 reviews, including 160,225 reviewers, and 926 products. The time stamp ranges from 200 to 4200 and the time span is about ten years (the earliest review date is October 2004, and the latest review is The time is January 2015).

False comment detection is a two category problem, so confusion matrix is used to evaluate its effect. In this dataset, comments are marked as "false" or "true," and reviewers are not marked. Before the analysis of the experimental results, the comments are sorted according to the final probability of cheating. The higher the ranking, the more likely the comments are to cheat, that is to say, the more likely the comments are to be false; the lower the ranking, the less likely they are to cheat. In the result analysis, this paper assumes that all nodes are false comments, and then compares them with the real tags of the corresponding comments according to the order of the sorted false comments.

This paper combines the features of these six fraud features to improve the algorithm's ability to detect false reviews. Among them, the feature of "Rating deviation" is selected because false reviewers usually give the same or similar scores to the target product; the feature of "Extremity of rating" is selected because false reviews usually give a higher score to the product; select "Thresholder" The "rating deviation of a review" feature is due to the fact that the scoring deviation of false reviews is usually higher than that of real reviews; the "Rating variance" feature is selected because false reviews usually have a large number of

consistent ratings in a short period of time; The "Early time frame" feature is due to the fact that fake reviewers usually make reviews in the early stages of the product to increase the impact on later reviewers; choosing "Is singleton review" is a feature because fake reviewers are usually not long-term users.

Since the above six features have strong identification ability for the authenticity and falsity of comments, this paper uses formula (11) to integrate the six features as cheating priori of each comment. Among them, F is the number of integrated features, and is the cheating prior obtained by the i-th feature. The larger the value of the formula, that is, the greater the cheating prior of the comment, the greater the probability that the comment is a cheating comment, otherwise, the smaller the probability of cheating. What needs to be pointed out is that the value is only the initial value given by its characteristic behavior, and the optimal value of cheating probability can be obtained after training optimization and continuous adjustment.

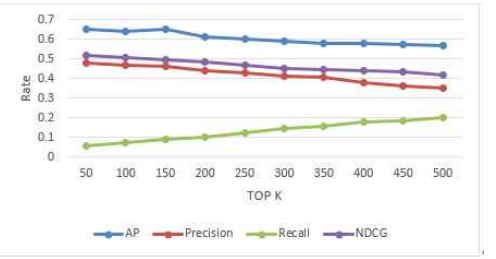
$$P(x_i) = 1 - \sqrt{\frac{\sum_{l=1}^{F} f(x_{li})^2}{F}} \tag{11}$$

The optimal values of the parameters involved in the false comment online detection algorithm in this article are: sliding time window step size TW_STEP=100, similarity threshold MIN_SIM=0.02, and similarity calculation parameters, [Fig. 6] shows the average accuracy (AP), precision (Precision) and recall of the false comment online detection algorithm when the time window start time TW_FROM=2000, the sliding time window step size TW_STEP=100, and the time window end time TW_END=2100 The change curve of metrics such as Recall and Normalized Loss Cumulative Gain. The horizontal axis represents the top K reviews, and the vertical axis represents the value of each metric. It can be seen from the results that the false comment detection algorithm has achieved better results through online detection methods. Compared with the offline detection algorithm, the model performance of the online detection algorithm has been improved, and the real-time performance of the algorithm far exceeds the existing offline detection algorithm.

6. Conclusion

In this paper, a practical and effective online detection model of false comments is proposed. Different from the previous off-line detection methods, this model uses sliding time window mechanism to realize the function of real-time monitoring, and can mine the relevant

information of false comments in a short time. By extracting the comment data in each time window, the measurement method of comment similarity is defined, which makes the relationship between comments numerical and establishes the relevant comment graph. Then, the comment graph is modeled as Markov random field by associating the comment nodes with random variables, and the online detection algorithm of false comments is generated.



[Fig. 6] Results of Algorithm Evaluation

This paper conducts experiments on the YelpNYC data set, and optimizes the performance of the experimental parameters through several indicators such as average precision, precision rate, recall rate, etc., and comprehensively analyzes the experimental results. The results show that the average precision of the algorithm can be More than most offline detection algorithms, especially in real-time detection of false comments. However, there is still room for improvement in this method, such as whether the priori selection of review nodes can more accurately locate false reviews in advance, thereby improving the accuracy of the entire model.

References

[1] Shengli D., Fenfen W., Research progress of fake review information detection: a perspective of internet governance, Journal of Information Resources Management, **(2019)**, Vol.9. No.3, pp.73-81.

[2] Jindal N., Liu B., Opinion spam and analysis, Proceedings of the 2008, International Conference on Web Search and Data Mining, ACM, pp.219-230, **(2008)**

[3] Lim E. P., Nguyen V. A., Jindal N., Detecting product review spammers using rating behaviors, ACM International Conference on Information and Knowledge Management, ACM, pp.939-948, **(2010)**

[4] Wu Y., Ngai E. W. T., Wu P., Fake online reviews: Literature review, synthesis, and directions for future research, Decision Support Systems, **(2020)**, p.113280

[5] Mukherjee A., Liu B., Glance N., Spotting fake reviewer groups in consumer reviews, International Conference on World Wide Web, ACM, pp.191-200, **(2012)**

[6] Wang Z., Gu S., Xu X., GSLDA: LDA-based group spamming detection in product reviews, Applied Intelligence, **(2018)**, No.1, pp.1-14.

[7] Ott M., Choi Y., Cardie C., Finding deceptive opinion spam by any stretch of the imagination, Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, Association for Computational Linguistics, **(2011)**, Vol.1, pp.309-319.

[8] Li J., Ott M., Cardie C., Towards a general rule for identifying deceptive opinion spam, Meeting of the Association for Computational Linguistics, **(2014)**, pp.1566-1576.

[9] Hooi B., Song H., Beutel A., FRAUDAR: Bounding Graph Fraud in the Face of Camouflage, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, **(2016),** pp.895-904.

[10] Fei G., Mukherjee A., Liu B., Exploiting burstiness in reviews for review spammer detection, International Conference on Weblogs and Social Media, **(2013)**, pp.175-184.

[11] Ye J., Kumar S., Akoglu L., Temporal opinion spam detection by multivariate indicative signals, Tenth International AAAI Conference on Web and Social Media, **(2016)**, pp.743-746.

[12] Yasuda M., Kataoka S., Tanaka K., Statistical analysis of loopy belief propagation in random fields, Physical Review, **(2015)**, Vol.92, No.4, p.42120

[13] Ford Lumban Gaol, Erik Chen, The Influence of Security, Trust, Service Quality and Risk Perception in B2C E-Commerce againts People's Online Purchasing Decisions (Survey on Customers of Tokopedia), International Journal of Advanced Science and Technology, **(2019)**, Vol.124, March, pp.103-110.

[14] Keum-Joo Kim, The Effects of Originality of E-Commerce Website Design On Site Reliability and Loyalty, International Journal of Advanced Science and Technology, **(2019)**, Vol.124, March, pp.33-46.

[15] Jiyoung Yoon, Soonhee Joung, A Study of Purchase Intention of Eco-friendly Products: A Cross-Cultural Investigation between Korea and China, International Journal of Smart Business and Technology, **(2019)**, Vol.7, No.2, pp.19-24.

[16] Jung Hyeon Jang, Ju Im Jung, Jee Young Kim, The Future Coping Strategies in the Field of Beauty in the Age of the Fourth Industrial Revolution, International Journal of Smart Business and Technology, **(2019)**, Vol.7, No.2, pp.7-12.

[17] Feng Jing, Chae-Kwan Lim, An Empirical Study on the Effect of Management Service Quality of High-rise Apartment on Residential Satisfaction: Focused on High-rise Apartment in China, International Journal of IT-based Management for Smart Business, **(2020)**, Vol.7, No.1, pp.23-30.

[18] Liu Yan, Zhou Wanting, Meng Lingyue, Lu Ying, Fan Zhipeng, Research on the Influencing Factors of Customer Experience of Retail Enterprises Based on AHP in the Background of New Retail, International Journal of IT-based Business Strategy Management, **(2019)**, Vol.5, No.1, pp.13-22.