Postprint: Heterogeneous Information Network Embedding for Public Security Weibo Dissemination Prediction

Authors: Sun Ran, An Lu

Date: 2023-04-01T00:00:00+00:00

Abstract

[Purpose/Significance] Predicting whether users will repost or comment on wanted notices on Weibo, and researching and evaluating important features that influence the dissemination of such notices, can help public security Weibo accounts improve their operational performance and enhance communication and cooperation between the police and the public.

[Method/Process] Targeting the characteristics of wanted notices on Weibo, we extract case features—including case location keywords, temporal keywords, wanted notice level, and reward availability—from wanted notices, building upon user features, temporal features, and Weibo text structure features. We utilize the XGBoost algorithm to compute the importance of different features for repost and comment prediction. Additionally, by integrating dissemination network features and node attributes, we construct a public security Weibo dissemination prediction model based on feature-attributed heterogeneous information network embedding, and perform model training and evaluation.

[Results/Conclusion] The prediction model achieves AUC values of 0.737 and 0.799 on repost and comment datasets, respectively. As the model integrates network structure features and diverse node attributes, it more closely approximates real-world heterogeneous information networks and demonstrates higher accuracy compared to traditional link prediction models. Additionally, experimental results on feature importance indicate that the proposed case keyword features rank highest in importance among all features influencing Weibo repost and comment prediction.

ChinaRxiv 中X

Full Text

Preamble

Research on Propagation Prediction of Police Microblog Entries from the Perspective of Heterogeneous Information Network Embedding

Sun Ran[1]**, An Lu**[1,2] [1]School of Information Management, Wuhan University, Wuhan 430072 [2]Center for Studies of Information Resources, Wuhan University, Wuhan 430072

Abstract: [Purpose/Significance] Predicting whether users will retweet or comment on wanted notices on microblogs, and evaluating the important features affecting their propagation, can help public security departments improve their operational performance and enhance communication and cooperation between police and citizens. [Method/Process] Focusing on the characteristics of wanted microblog entries, we extracted case-related features including location keywords, time keywords, wanted notice level, and reward information, in addition to user features, temporal features, and microblog text structure features. The XGBoost algorithm was employed to calculate the importance of different features for retweet and comment prediction. Combining network propagation features with node attributes, we constructed a propagation prediction model for police microblogs based on feature-attributed heterogeneous information network embedding, which was then trained and evaluated. [Result/Conclusion] The prediction model achieved AUC values of 0.737 and 0.799 on retweet and comment datasets, respectively. By integrating network structural features with diverse node attributes, the model more closely approximates real-world heterogeneous information networks and demonstrates higher accuracy than traditional link prediction models. Furthermore, feature importance experiments revealed that the proposed case keyword features exhibited the highest importance among all features influencing retweet and comment predictions.

Keywords: Information dissemination; Police microblog; Link prediction; Graph representation learning; Heterogeneous information network

Classification Number: G203

With the advent of the big data era, information dissemination channels have become increasingly diversified, and social media platforms such as Sina Weibo and WeChat have emerged as important new platforms for government affairs media in China. According to the "44th Statistical Report on China' s Internet Development" released by the China Internet Network Information Center (CNNIC), as of June 2019, the scale of online government service users in China reached 509 million, accounting for 59.6% of all internet users, with 297 prefecture-level administrative regions having opened "two microblogs and one app" and other new media channels, achieving an overall coverage rate of 88.9%. On June 24, 2019, a video of a woman being violently assaulted on the street went viral online. Subsequently, police departments across the country intervened in the investigation. Unable to locate the source of the nighttime video, China Police

Online posted on Weibo soliciting relevant clues from the public. On the 25th, netizens began providing leads to the police, and by 22:00 that evening, the suspect was apprehended. This fully demonstrates the powerful function of mass participation in information acquisition within online communities. This form of citizen mutual assistance and cooperation provides new ideas for police to collect investigation clues and solve cases. How to conduct mass work in virtual online communities and leverage the broad masses to help public security organs punish evil and combat crime is a question that China' s criminal investigation work needs to consider.

From the user perspective, information dissemination in cyberspace is a diversified interactive diffusion process between individuals. In research on online public opinion information dissemination, scholars have primarily considered retweet relationships between users [1], while comments have been studied mainly from perspectives of text analysis and sentiment analysis. However, user interactions also involve following and commenting relationships, and considering only retweet relationships cannot accurately capture the interaction networks of users during public opinion event propagation. This study focuses on wanted microblog entries, employing the XGBoost algorithm to investigate the performance of different features—including content features and microblog text structure features—in retweet and comment prediction. Using users and microblogs as nodes, we construct heterogeneous information networks based on relationships between users and microblogs, as well as among users themselves. By combining propagation network structure with node attribute features, we build an attribute-specific heterogeneous information network embedding model to perform link prediction for police microblog propagation activities such as retweeting, commenting, and following, thereby improving prediction accuracy. Studying users' various information dissemination behaviors helps relevant departments understand information propagation mechanisms and is significant for government departments to conduct public opinion monitoring and personalized microblog recommendations. The research aims to reveal the influencing factors and mechanisms of police microblog information dissemination, as well as public retweeting and commenting behavior patterns, providing suggestions for the operation and construction of police microblogs and helping government departments enhance communication, cooperation, and interaction between police and citizens.

2 Related Research

2.1 Research on Microblog Propagation Patterns in Specific Domains

Currently, scholars have employed content analysis [2], social network analysis [3], neural networks [4], and other methods to explore microblog propagation patterns in specific domains such as government affairs, disasters, and health. From a macro perspective, based on retweet levels and frequencies, government microblog information propagation patterns can be categorized into two-level propagation, ordinary multi-level propagation, and satellite propagation models [5]. From a micro individual perspective, numerous studies have examined user interaction patterns on microblogs, revealing that user interaction patterns in disaster information propagation networks exhibit certain differences across different themes [6].

A wanted notice is an investigative measure by which public security organs issue a warrant to apprehend criminal suspects, defendants, or criminals who should be arrested but are at large, typically conducted through the release of wanted notices. In recent years, police microblogs have provided a new platform for public security organs to release information, making online wanted notices more widely disseminated and faster, thereby overcoming the geographical limitations of traditional wanted notices and enabling quicker acquisition of relevant clues about wanted suspects from the masses. Through literature review, we found that although research on microblog propagation patterns is abundant, few studies have examined wanted information as the object of investigation to explore the characteristics of different user subjects' microblog propagation patterns and user attributes during police microblog information dissemination. Existing research mostly consists of descriptive analyses from the perspective of police microblog operation and management [7], lacking empirical studies based on objective data of microblog user behavior.

2.2 Research on Microblog Propagation Prediction

Microblog propagation prediction research can generally be conducted from two perspectives: (1) Popularity prediction of microblog information, including predicting retweet scale, speed, and efficiency, typically using mathematical modeling methods based on epidemic models and classification/regression models based on machine learning [8]. Xu Yuemei et al. [9] employed convolutional neural networks and gradient boosting decision trees to predict the retweet scale of government microblogs based on user features, temporal features, and content features, identifying important features affecting this scale. (2) Prediction of user propagation behavior. Users participate in public life through various means, such as publishing opinions on social issues, commenting, or retweeting others' content [10]. Among these, user retweeting behavior is the most important method of microblog information diffusion [11-12]. Scholars have primarily used user features and microblog text topic and structure features, applying machine learning algorithms such as SVM and logistic regression models to achieve microblog user retweet prediction [13]. J. Zhu et al. [14] explored the effects of text content, influence, and time on retweeting, using a logistic regression classifier for prediction, finding that incorporating temporal factors better facilitates understanding of retweeting mechanisms.

In recent years, scholars have increasingly approached user retweet behavior prediction from the perspective of user social network relationships, combining user behavior log data with complex network methods and content analysis. B. Liang et al. [15] utilized one-class collaborative filtering methods based on fundamental factors affecting microblog retweeting. However, existing research has three main shortcomings: (1) It fails to distinguish between different user participation behaviors in microblog propagation; (2) Current propagation prediction analysis methods are mostly applicable to homogeneous information networks; (3) In prediction for police domain microblog propagation, event keyword features have not been considered. Given that the GATNE model [28] can handle heterogeneous information networks, is suitable for large-scale data modeling, and can incorporate rich node attributes and network structural features—aligning well with the multiple node types and diverse relationships present in social networks—this study addresses these limitations by proposing an attribute-specific heterogeneous information network embedding model to perform link prediction for police microblog propagation activities such as retweeting, commenting, and following. We construct heterogeneous information networks based on relationships between users and microblogs, as well as among users themselves, extract node attribute features including user features, text structure features, event features, and case keyword features, and finally combine propagation network structure with node attribute features, introducing the GATNE-I model for microblog propagation prediction.

3 Research Methods

This study conceptualizes police microblog propagation networks as heterogeneous information networks, where nodes represent users or microblog entries, and node attributes include user features, microblog text structure features, case features, temporal features, etc. For the specific domain of police microblog propagation, we introduce the heterogeneous information network embedding model GATNE-I to construct a microblog propagation prediction model based on attribute-specific heterogeneous network embedding, incorporating user, microblog, and temporal features as specific attributes into the GATNE-I model.

First, we preprocess wanted microblog data and retweet/comment user data, extracting user features and temporal features, and identifying case location, time, and name keywords in wanted microblogs. The XGBoost algorithm [29] is used to rank the importance of microblog features and user features. Simultaneously, we construct heterogeneous information networks based on retweet, comment, and follow relationships, with node categories divided into users and microblogs, and edge categories including retweet and comment relationships between users and microblogs, as well as follow relationships between users. Combining node attribute features, we construct an attribute-specific heterogeneous network embedding prediction model to forecast microblog propagation.

3.1 Construction of Microblog Attribute-Specific Heterogeneous Network

3.1.1 Microblog Heterogeneous Information Network Let a directed network graph be $G = (V, E, A)$, where $\phi : V \to O$ is the node type mapping function and $\Phi : E \to R$ is the edge type mapping function, with O and R representing the sets of all node types and edge types, respectively. Each node

$v \in V$ belongs to a specific node type, $A = \{x_i|v_i \in V\}$ is the set of all node features, where x_i is the feature associated with node v_i; $E = \cup_{r\in R} E_r$ indicates that E_r contains all edge types, $r \in R$ and $|R| > 1$. For each edge type $r \in R$, we partition the network into $G_r = (V, E_r, A)$, which we refer to as an attribute-specific heterogeneous information network.

This study abstracts the microblog heterogeneous information network as containing two types of nodes: microblogs (I) and users (U), with three types of relationships: user retweets, comments, and follows. We construct microblog heterogeneous information networks using these relationships as meta-paths: (1) Meta-path based on microblog retweet network. When a microblog is retweeted by two different users, the meta-path is represented as $U_1 - R_1 - I - R_2 - U_2$, where U_1 and U_2 represent different users who retweeted the same microblog I, and R_1 and R_2 represent the retweeted microblogs after users U_1 and U_2 retweeted microblog I. (2) Meta-path based on microblog comment network. When a microblog is commented on by two different users, the meta-path is represented as $U_1 - C_1 - I - C_2 - U_2$, where U_1 and U_2 represent different users who commented on the same microblog I, and C_1 and C_2 represent the microblog comments after users U_1 and U_2 commented on microblog I. (3) Meta-path based on user follow network. When a user is simultaneously followed by two different users, the meta-path is represented as $U_1 - U - U_2$, where U_1 and U_2 represent different users who follow the same user U.

3.1.2 Feature Extraction The wanted microblog propagation prediction model constructed in this study primarily explores factors affecting wanted microblog information propagation from three dimensions: user features, microblog features, and temporal features, as shown in Table 1 :

Table 1 Wanted Microblog Features and Feature Values

Feature Category	Feature Name	Feature Value
User Features	Authentication Type	No authentication / Personal authentication / Institutional authentication
	Number of Followers	0-999 / 1000-9999 / 10000-99999 / 100000-999999 / 1000000-10 million / Over 10 million
	Number of Followees	0-299 / 300-599 / 600-899 / 900-1999 / 2000-9999 / 10000-20000

Feature Category	Feature Name	Feature Value
	Total Microblogs Posted	0-999 / 1000-9999 / 10000-99999 / 100,000+
	Industry	Traditional media / Self-media / New media / Government agency / Public figure / Personal group organization / Enterprise / Public welfare organization / Industry - Other
	Public Security System	Yes / No
	VIP Status	Yes / No
	Location	Beijing / Shanghai / Guangdong / Hunan / Zhejiang / Hubei···
Microblog Features	Text Structure	Whether contains URL / Hashtag / Image / Video / Mention / Emoji / Case keywords
	Location Keywords	Whether contains location keywords
	Time Keywords	Whether contains time keywords
	Behavior Keywords	Whether contains behavior keywords
	Suspect Description Keywords	Whether contains suspect description keywords
	Wanted Notice Level	Whether contains wanted notice level
Temporal Features	Time Period	Late night (00:01-06:00), Early morning (06:01-08:30), Morning (08:31-12:00), Noon (12:01-14:00), Afternoon (14:01-18:00), Evening (18:01-24:00)
	Holiday	Holiday; Non-holiday

Feature Category	Feature Name	Feature Value
	Day of Week	Monday; Tuesday; Wednesday; Thursday; Friday; Saturday; Sunday
	Original Post	Original post; Non-original post

(1) **User Features.** We categorize user features into basic user attributes and user behavior characteristics. Basic attributes include authentication type, industry, whether a public security system user, gender, VIP status, and location (six dimensions). Authentication type feature values include institutional authentication users, personal authentication users, and non-authenticated users. Institutional authentication users inherently possess certain influence and authority, making their posts more likely to attract public attention and discussion. User industry is divided into traditional media, self-media, new media, government agencies, public figures, personal group organizations, enterprises, public welfare organizations, and industry-other. Personal group organizations refer to fan clubs, local associations, hometown associations, etc. Users whose authentication information or usernames contain words like "public security" or "police" are classified as public security system users. VIP status indicates whether the user has activated Weibo membership. Location information is obtained from user profile geographic data, with feature values including 34 Chinese provincial-level administrative regions plus "overseas" and "other" (36 feature values total). Since the number of followers and followees may affect microblog popularity [30], we also selected user follower count, followee count, and total posted microblogs as user behavior features.

(2) **Microblog Features.** Considering the particularity of wanted microblogs, we divide microblog features into text structure features and case keywords. In text structure features, URLs and hashtags show strong correlation with microblog retweeting [11], and users typically add images and videos to convey more information when posting. Observation of wanted microblog content reveals that users often include emojis to emphasize information when posting major wanted notices, such as [microphone] [shocked]. Therefore, we incorporate the presence of links, hashtags, images, videos, mentions (@), and emojis into microblog text structure features.

Article 266 of the "Regulations on the Procedures for Public Security Organs to Handle Criminal Cases" explicitly states: "Wanted notices should, as far as possible, include the wanted person' s name, alias, former name, nickname, gender, age, ethnicity, birthplace, household registration location, residence, occupation, ID number, clothing and physical characteristics, accent, behavioral habits, and recent photos of the wanted person. Fingerprints and photos of other physical evidence may be attached. Except for matters that must be kept confidential, the time, location, and brief case details of the crime should also be stated." Currently, wanted notices are generally disseminated through combined photo-text formats. Therefore, public security microblogs typically contain substantial suspect information. However, during wanted microblog propagation, other users do not need to follow these regulations when posting information, resulting in microblogs with less information, such as those lacking photos containing suspect text information. When a microblog contains more information, it tends to attract greater public attention and discussion. Consequently, we add case-related keyword features to microblog content characteristics, including crime time keywords, location keywords, behavior keywords, suspect description keywords, and wanted notice level. Among these, wanted notice level and reward availability can be directly determined through keyword screening, while time keywords, location keywords, and behavior keywords can be extracted through named entity recognition. Based on the crime categories in the "Criminal Law of the People' s Republic of China," we annotate the type of wanted case contained in each microblog. Since cases progress through different stages, suspect status may change from at-large to subsequently arrested, we also distinguish case progress based on the current suspect status shown in wanted microblogs, categorizing them as at-large, arrested, surrendered, etc.

(3) **Temporal Features.** Different microblog posting times affect the number of users who can receive the information. Therefore, based on microblog posting time periods, day of week, and legal holidays, and according to microblog user activity patterns, we divide posting time periods into late night (00:01-06:00), early morning (06:01-08:30), morning (08:31-12:00), noon (12:01-14:00), afternoon (14:01-18:00), and evening (18:01-24:00) [8]. Since our dataset spans a large time period and includes multiple wanted events, once the Ministry of Public Security releases a wanted notice, it may be retweeted by other users. Cases that easily cause social panic, with at-large suspects still posing threats to society, are more likely to be retweeted and reported by traditional and self-media. Therefore, we define original posts as the microblogs that first release wanted information for the same wanted case at different stages (e.g., arrested, surrendered) within the dataset.

3.2 Feature Importance Ranking

Feature importance is obtained by calculating and ranking each attribute in the dataset. The principle involves randomly extracting one feature from the dataset at a time and calculating the degree of performance metric degradation; greater change indicates higher feature importance. Methods primarily include ensemble learning algorithms such as random forest, decision trees, and XGBoost, broadly categorized into boosting and bagging approaches. The XGBoost algorithm improves upon gradient boosting decision trees (GBDT) by fitting data residuals and adding model complexity regularization terms to the loss function, effectively preventing overfitting. Its parallel and distributed design enables very fast training speeds. The XGBoost model treats missing values as sparse matrices, not considering their values during node splitting. Missing value data is allocated to left and right subtrees for separate loss calculation, with the better-performing subtree selected. If no data is missing during training but missing values appear during prediction, classification defaults to the right subtree.

This study uses the XGBoost algorithm for feature importance ranking, employing the `train_{{{test}}_{{split}}}()` function from Python' s sklearn package to randomly divide training and test sets. We treat the link prediction problem as a binary classification problem, setting the label to 1 if two nodes are connected and 0 otherwise. XGBoost' s input is a combination of feature vectors from two nodes—for retweet prediction, the input includes feature values of the retweeting user and the microblog. For multi-dimensional feature vector x_i, XGBoost' s output is shown in formula (1):

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \quad (1)$$

where k is the number of CART trees, F represents all possible CART trees, and $f_k(x_i)$ denotes the classification result of CART tree k. The XGBoost model' s objective function is shown in formula (2):

$$\mathrm{obj}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (2)$$

where the first term l in the objective function is the loss function measuring the difference between predicted and target values, and the second term Ω represents the regularization term, which is the sum of complexities of k CART trees, including the number of leaf nodes and leaf node scores.

3.3 GATNE-I Model

Y. Cen et al. [28] proposed the GATNE model (GATNE-I and GATNE-T) in 2019, which can handle real-world networks composed of large-scale nodes and multiple edge types, with each node associated with different attributes. This paper introduces the GATNE-I model to solve link prediction problems with different nodes and edges in heterogeneous information networks, performing link prediction for police microblog propagation activities such as retweeting, commenting, and following. In the model, each node' s embedding under each edge type consists of a general embedding b_i based on node features, an edge embedding u_i, and node attributes, corresponding to structural information, heterogeneous information, and attribute information, respectively. The general embedding is shared by each node under each edge type, while each node' s edge embedding is calculated through adjacent nodes' edge embeddings according to different edge types. The edge embedding calculation for node v_i under edge type r at layer k is shown in formula (3):

$$u_{i,r}^{(k)} = \text{aggregator}(\{u_{j,r}^{(k-1)}, \forall v_j \in N_{i,r}\}) \quad (3)$$

where i, j denote node numbers in the heterogeneous information network, r denotes the edge type, and the aggregator function can use mean aggregation or other types of pooling aggregation. In the GATNE-T model, the initial edge embedding $u_{i,r}^{(0)}$ for each node v_i under each edge type r is randomly initialized. The layer k edge embedding $u_{i,r}^{(k)}$ is defined as $u_{i,r}$, and all edge embeddings of node v_i are concatenated into a matrix U_i of size $s \times m$, where s denotes the edge embedding dimension and m denotes the number of edge types.

We calculate the coefficient $a_{i,r}$ for linear combination of vectors under each edge type r in U_i through a self-attention mechanism to obtain the weight of each edge type' s vector representation, as shown in formula (4):

$$a_{i,r} = \text{softmax}(w_r^T \tanh(W_r U_i))^T \quad (4)$$

where w_r and W_r are trainable parameters under edge type r with sizes d_a and $d_a \times s$ (d is the general embedding dimension), respectively, and T denotes vector or matrix transpose. We finally obtain the vector representation of each node v_i under edge type r in the GATNE-T model, as shown in formula (5), where b_i is the general embedding of node v_i, α_r is a hyperparameter representing the importance of edge embedding relative to the overall embedding, and $M_r^T \in \mathbb{R}^{s \times d}$ is a trainable transformation matrix.

$$v_{i,r} = b_i + \alpha_r M_r^T U_i a_{i,r} \quad (5)$$

The GATNE-T model cannot handle unseen nodes, while real-world network data is often incomplete. The GATNE-I model, which introduces node features, can solve this problem by defining the general embedding b_i as a parametric equation of node v_i' s attributes x_i. Different nodes v_i may have attributes x_i of different dimensions. In the GATNE-I model, $u_{i,r}^{(0)}$, which was randomly initialized in GATNE-T, is obtained through a node attribute function, as shown in formula (6), where $g_{z,r}$ is also a transformation function mapping features to node v_i' s edge embedding under edge type r.

$$u_{i,r}^{(0)} = g_{z,r}(x_i) \quad (6)$$

Additionally, the GATNE-I model adds an extra attribute term to the overall embedding of node v_i under edge type r. The expression for node v_i' s vector representation $v_{i,r}$ under edge type r is shown in formula (7), where β_r is a coefficient and D_z is a feature transformation matrix for node type z corresponding to node v_i.

$$v_{i,r} = h_z(x_i) + \alpha_r M_r^T U_i a_{i,r} + \beta_r D_z^T x_i \quad (7)$$

4 Experiments and Results Analysis

4.1 Dataset

This study' s dataset is from the domestic social media platform Sina Weibo. Using the search terms "wanted suspect," we collected original microblogs and their comments, retweets, and all user basic information. After manual verification to remove microblogs unrelated to wanted notices, we collected 14,905 original microblogs, 86,146 retweeted microblogs, and 62,548 microblog comments from January 1, 2016, to September 15, 2019. We collected follow and follower information for users who posted original microblogs. Due to the large data volume and overly sparse follow network, we removed users who appeared only once in the follow network, obtaining 168,059 follow relationships containing 143,370 Weibo users. In the microblog heterogeneous information network, nodes represent Weibo users, and edges represent retweet, comment, and follow relationships between users. Collected user fields include user ID, username, gender, province, VIP status, authentication, post count, followee count, and follower count. Microblog text information fields include microblog ID, user ID, posting time, login device, like count, retweet count, comment count, image link, and microblog text.

4.2 Analysis of Wanted Information Propagation Patterns

Taking the wanted microblog retweet network as an example, this study examines user attributes among different subjects posting microblogs during information propagation. Due to the large retweet network structure, we selected nodes appearing more than once in the dataset, obtaining a network structure with 697 nodes and 5,118 edges. Among these, 108 public security system users posted microblogs with 323 public security system users participating in retweeting. Visualization using Gephi software revealed that within the same province (municipality), public security users from different regions typically retweet microblogs posted by the highest-level public security user in that region. Ordinary users in the retweet network are also mostly from the same province (municipality). Taking "Peaceful Chongqing" as an example, as shown in Figure 1 [Figure 1: see original paper], we found that its retweeting users were mostly public security users from different districts in Chongqing municipality, such as "Peaceful Nan' an" and "Peaceful Qianjiang," while ordinary users like "Elegant Cat Forever" and "Fog Capital Old Cat" also belong to the Chongqing region. This indicates that wanted notice information propagation is generally limited to provincial/municipal boundaries, with relatively little propagation of wanted information between different provinces/municipalities among public security microblogs.

Taking "People's Daily"as an example, compared to public security system users, the retweet network contains more ordinary users and some public security system users, such as "Wuchang Baishazhou Police Station" and "Peaceful Jili," with public security system users distributed across various regions. This also demonstrates that news media, due to their inherent influence, enable wanted information propagation to break through user groups centered on public security microblogs.

4.3 Case Keyword Extraction

Wanted microblog texts contain substantial information related to wanted events, including crime time, location, persons (names), and criminal behavior. For some case keywords (such as wanted notice level), direct keyword screening of microblog texts can determine their presence. Since existing wanted notices are mostly disseminated through combined photo-text formats, we download images based on crawled microblog image links, then use Lightning OCR image text recognition software to obtain text information from images. Combined with microblog text content, we use the HanLP word segmentation tool in Python to segment original microblog texts and perform part-of-speech tagging to identify person names, locations, time information, etc., such as Qin Zhigang/n, Guangdong/ns, 23rd/t, etc. We supplement attribute fields for case keywords as shown in Table 2 . Case categories are classified according to crime categories and definitions in the "Criminal Law of the People' s Republic of China," combined with case suspect behavior keywords. Since cases progress through different stages and suspect status may change from at-large to subsequently arrested, we also distinguish case progress based on the current suspect status shown in wanted microblogs, categorizing them as at-large, arrested, surrendered, etc.

Table 2 Case Keyword Fields

Keyword Type	Examples
Location Keywords	Guangxi/ns, Huanjiang County/ns, Datang Town/ns, Zhaotun Village/ns, Beijing Station/ns, etc.
Time Keywords	2015/t, October/t, 23rd/t, recently/t, a few days ago/t, early morning/t, that night/t, etc.
Behavior Keywords	kill/v, rob/v, involve in organized crime/v, beat/v, lend/v, rape and kill/v, etc.

Keyword Type	Examples
Suspect Description Keywords	skin/n, round face/n, bald/n, skin color/n, body shape/n, jeans/n, etc.
Wanted Notice Level	Class A wanted notice, Class B wanted notice, International wanted notice
Reward	reward/v, bonus/v, prize/vn, reward 征集/n, etc.
Case Category	Organizing, forcing, enticing, harboring, or introducing prostitution; Smuggling; Smuggling, trafficking, transporting, or manufacturing drugs; Producing, selling, or disseminating obscene materials; Tax evasion; Public health hazards; Public safety hazards; Corruption and bribery; Producing or selling counterfeit goods; Disrupting market order; Disrupting public order; Crimes infringing upon citizens' personal rights and democratic rights (intentional homicide, intentional injury, rape, etc.); Property infringement crimes (robbery, theft, etc.); Disrupting financial management order; Disrupting environmental resource protection; Financial fraud; Cultural relics management obstruction; Judicial obstruction; Border management obstruction; Official business obstruction; Corporate management obstruction; Malfeasance
Case Progress	Surrender/nz, at-large/nz, capture/n, arrested/v, etc.

4.4 Microblog Feature Importance Ranking

To determine how feature values in different node network relationships affect link prediction, this study uses the XGBoost algorithm to screen important features, ranking feature importance for retweet and comment predictions separately, with results shown in Figure 2 [Figure 2: see original paper].

Experimental results demonstrate that in both retweet and comment predictions, microblog case keyword feature importance is significantly higher than other features, indicating that microblogs containing wanted case-related information are more likely to be retweeted or commented on. When public security microblogs or news media post wanted information, supplementing more case information more easily triggers user participation in information propagation. User follower count and holiday importance are significantly lower than other features, indicating that user follower numbers and whether it is a holiday have relatively small impact on predicting user retweeting or commenting on wanted microblogs. User location, day of week, user industry, and microblog posting time period all show relatively high feature importance (greater than 0.05). However, user authentication type importance is lower in retweet prediction than in comment prediction, while original post importance is much higher in retweet prediction than in comment prediction, reflecting that within the same wanted event, the earlier the microblog is posted, the more likely it is to be retweeted.

From the microblog perspective, conducting specific analysis on some features including microblog case keywords, temporal features, and text structure features, as shown in Figure 3 [Figure 3: see original paper], reveals that in the ranking of case keyword feature importance, Class A wanted notice (0.033), disrupting financial management order crime (0.029), and case category-other (0.055) have greater impact on comment prediction, while location keywords (0.027), suspect description keywords (0.022), property infringement crime (0.028), and arrested (0.028) have greater impact on retweet prediction. In microblog temporal features, Sunday (0.138) shows the highest importance in retweet prediction and relatively high importance (0.017) in comment prediction. Original post shows high importance in both retweet and comment predictions (greater than 0.02). Evening, Tuesday, and Saturday show lower importance in retweet prediction than in comment prediction. In microblog text structure features, video and hashtag feature importance are both relatively high, providing insights for public security system users posting wanted microblogs: posts should 尽可能 contain case-related information such as suspect characteristics and case progress, posting times can be selected on Saturdays and Sundays, and microblogs should include case-related videos, images, and hashtags to convey richer information and trigger more user attention. For example, during the Ministry of Public Security' s July 2019 nationwide operation to wanted 50 major at-large criminals, when local public security departments posted microblogs, they added the topic #MinistryofPublicSecurityWanted50MajorAtLargeCriminals#. Although local public security microblogs have relatively low influence, they could still attract considerable retweets.

From the perspective of users participating in retweeting or commenting, user location shows the highest importance in both retweet (0.180) and comment (1.87) predictions. Among these, North China regions such as Tianjin, Beijing, and Hebei show high importance in both retweet and comment predictions, while South China regions such as Guangxi, Hainan, Hong Kong, and Macau show low importance in both predictions. This also reflects that the user circle for wanted microblog propagation remains somewhat limited, with cooperative propagation among government microblogs being more evident. User features with followee counts of 300-599 and 2000-9999 show significantly higher importance in retweet prediction than in comment prediction, while user features with followee counts of 900-1999 have lower impact on retweet prediction than on comment prediction. This can help social media platforms like Sina perform user recommendations, pushing relevant microblogs to user groups likely to retweet or comment on wanted microblogs, thereby increasing the influence of public security and other government microblogs, promoting widespread dissemination of wanted information, and helping users contribute to assisting public security departments.

4.5 Experimental Results

This study conducts link prediction experiments based on the attribute-specific heterogeneous information network embedding (GATNE) model. To evaluate model performance, we select the same experimental data to construct baseline link prediction models including DeepWalk, Node2vec, Line, GAE, and SDNE, and compare their performance evaluation results with the GATNE-I model. Since DeepWalk, Node2vec, and other models cannot handle networks with different edge types, we separately apply baseline models to retweet, comment, and follow relationship networks for prediction. We first experiment on a partial dataset as shown in Table 3 . The ratio of positive to negative samples in the dataset affects prediction accuracy. We generally divide the original dataset into training sets (70%), test sets (approximately 20%), and validation sets (approximately 10%), with roughly equal numbers of positive and negative samples in each set. The experimental environment is Intel(R) Xeon(R) E5-2640v4 x86_{64}, 2.4GHz, 20 cores, Nvidia Tesla V100, 16GB memory. Model performance evaluation results are shown in Table 4 . AUC refers to the area under the ROC curve, which plots False Positive Rate (FPR) on the x-axis and True Positive Rate (TPR) on the y-axis. F1 is the harmonic mean of model precision and recall, as shown in formula (8). The PR curve plots precision and recall as variables, with recall on the x-axis and precision on the y-axis, where PR value refers to the area under the PR curve. Larger AUC, F1, and PR values indicate better model performance.

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (8)$$

Table 3 Partial Extracted Dataset

Relationship Type	Nodes	Edges
Retweet	697	5,118
Comment	697	5,118
Follow	697	5,118

Table 4 Model Performance Evaluation Results (Partial Dataset)

Model	Retweet (AUC)	Retweet (F1)	Retweet (PR)	Comment (AUC)	Comment (F1)	Comment (PR)	Follow (AUC)	Follow (F1)	Follow (PR)
DeepWalk	0.646	0.562	0.547	0.524	0.550	0.587	0.599	0.608	0.704
Node2vec	0.562	0.552	0.543	0.562	0.536	0.680	0.550	0.552	0.660
Line	0.559	0.542	0.566	0.687	0.739	0.712	0.766	0.736	0.719
GAE	0.770	0.733	0.744	0.603	0.569	0.668	0.675	0.680	0.652
SDNE	0.662	0.669	0.693	0.634	0.678	0.599	0.724	0.782	0.757
GraphFactorization	0.733	0.780	0.540	0.728	0.609	0.578	0.672	0.693	0.675
GATNE-T	0.601	0.631	0.622	0.760	0.711	0.732	0.574	0.543	0.756
GATNE-I	0.768	0.756	0.603	0.569	0.577	0.571	0.756	0.768	0.571

As shown in Table 4, compared with baseline models, the GATNE-T model demonstrates relatively high accuracy on retweet and comment relationship datasets. Moreover, due to the rich node features, GATNE-I' s performance is significantly better than GATNE-T. Although prediction performance on the follow network is relatively poor, overall, the GATNE-I(T) model shows good results. We subsequently conduct experiments on the full dataset, selecting DeepWalk and Node2vec as baseline models for comparison. Results are shown in Table 5 , with GATNE-I model word vector dimension set to 200, random walk sequence length of 10, 20 random walk sequences per node, window size of 5, and negative sample count of 5.

Table 5 Model Performance Evaluation Results (Full Dataset)

Model	Retweet (AUC)	Retweet (F1)	Retweet (PR)	Comment (AUC)	Comment (F1)	Comment (PR)	Follow (AUC)	Follow (F1)	Follow (PR)
DeepWalk	0.456	0.468	0.486	0.456	0.424	0.594	0.733	0.603	0.574
Node2vec	0.528	0.520	0.551	0.442	0.404	0.589	0.744	0.569	0.543
GATNE-T	0.685	0.639	0.662	0.716	0.644	0.737	0.756	0.668	0.675
GATNE-I	0.737	0.651	0.672	0.799	0.689	0.792	0.768	0.677	0.601

In the above experimental results, the GATNE-I model achieves AUC values of 0.799 and 0.734 on comment and retweet prediction datasets, respectively. Similarly, although the GATNE-I(T) model performs poorly on the follow dataset, overall, the GATNE-I(T) model still outperforms DeepWalk and Node2vec models. DeepWalk and Node2vec models are based on homogeneous information networks, while the GATNE-I(T) model can handle heterogeneous information networks and effectively process large-scale data, better aligning with the data scale of social networks in the real world.

Conclusion

This study addresses the specific domain problem of wanted microblog information propagation, evaluating the importance of features—including user attribute features, microblog case keyword features, microblog text structure features, and temporal features—for propagation prediction from user and microblog perspectives using the XGBoost algorithm. The research covers user retweeting and commenting behaviors, finding that microblog case keyword features demonstrate the highest importance in both retweet and comment predictions. Subsequently, we examine the basic attributes of users in retweet networks posted by public security users and media users, discovering that the propagation circle of wanted microblogs posted by public security users is relatively limited, with close cooperation among public security microblogs within provinces/municipalities. Furthermore, we construct heterogeneous information networks based on retweet, comment, and follow relationships, and propose an attribute-specific heterogeneous network embedding model based on extracted user and microblog attributes to predict retweeting, commenting, and user follow relationships for wanted microblogs.

Experimental results show that the model achieves accuracies of 0.734 and 0.799 in retweet and comment predictions, respectively, outperforming other baseline models. This helps public security organs better implement "online wanted" initiatives and promotes active citizen participation through personalized user recommendations, providing practical conditions for public security organs to maximize the utilization of social networks. The limitations of this study include: (1) We treat multi-level retweets and comments as single-level retweets and comments, and future research could more deeply explore relationships between different users in multi-level retweeting and commenting; (2) The proposed prediction model is based on heterogeneous information networks, but information propagation processes in the real world are dynamic and have not considered temporal factors; (3) When exploring attributes of propagation users in microblogs posted by different subjects, we have not considered that different levels of wanted cases may trigger different inter-provincial propagation patterns, which could be further investigated in future research.

Acknowledgments

We thank the National Experimental Teaching Demonstration Center for Library and Information Science for providing experimental support for this research.

References

[1] Zheng Jianguo, Zhu Junxuan, Cao Ruzhong. Research on link prediction of social network information dissemination based on context [J]. Information Studies: Theory & Application, 2018, 41(6): 94-99.

[2] Xiao Fei. Analysis of concepts and strategies for government microblog public opinion information work in public crisis events—Taking the Ya' an earthquake as an example [J]. Library and Information Service, 2014, 58(1): 44-47, 71.

[3] Liu Xiaoping, Tian Xiaoying. Comparative analysis of social network characteristics between traditional and new media microblogs [J]. Library and Information Service, 2018, 62(5): 106-114.

[4] Zhang Lianfeng, Zhou Honglei, Wang Dan, et al. Mining key nodes in microblog public opinion based on hypernetwork theory [J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(12): 1286-1296.

[5] Chen Ran, Liu Yang. Research on government microblog information propagation patterns based on retweeting behavior [J]. E-Government, 2017(7): 108-117.

[6] Li L, Zhang Q, Tian J, et al. Characterizing information propagation patterns in emergencies: a case study with Yiliang Earthquake [J]. International Journal of Information Management, 2018, 38(1): 34-41.

[7] Chen Guiwu. Research on the coupling of formal and informal politics in local government innovation processes—Taking police microblogs as an example [J]. Journal of Public Management, 2014, 11(2): 60-69, 141-142.

[8] An Lu, Yi Xingyue, Sun Ran. Prediction and evolution of microblog influence in terrorist event contexts [J]. Library and Information Knowledge, 2019(4): 52-61, 81.

[9] Xu Yuemei, Liu Yunwen, Cai Lianqiao. A prediction model for retweet scale of government microblogs based on deep fusion features [J]. Data Analysis and Knowledge Discovery, 2020, 4(2/3): 18-25.

[10] Boyd DM, Ellison NB. Social network sites: definition, history, and scholarship [J]. Journal of Computer-Mediated Communication, 2007, 13(1): 210-230.

[11] Huang Wei, Liu Yi, Xu Yejing, et al. Construction of a heat measurement model for online public opinion tweets [J]. Library and Information Service, 2019, 63(20): 17-25.

[12] Suh B, Hong L, Pirolli P, et al. Want to be retweeted? Large scale analytics on factors impacting retweeting in twitter network [C]//Proceedings of the 2010 IEEE second international conference on social computing. Washington, DC: IEEE Computer Society, 2010: 177-184.

[13] Tian Lei, Ren Guoheng, Wang Wei. Predicting microblog user retweeting behavior for reading promotion [J]. Journal of the China Society for Scientific and Technical Information, 2017, 36(11): 1175-1182.

[14] Zhu J, Xiong F, Piao D, et al. Statistically modeling the effectiveness of disaster information in social media [C]//Proceedings of the 2011 IEEE global humanitarian technology conference. Seattle: IEEE Computer Society, 2011: 431-436.

[15] Jiang B, Liang J, Sha Y, et al. Retweeting behavior prediction based on one-class collaborative filtering in social networks [C]//Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval. New York: Association for Computing Machinery, 2016: 977-980.

[16] Khan ML. Social media engagement: what motivates user participation and consumption on YouTube? [J]. Computers in Human Behavior, 2017, 66: 236-247.

[17] Jaccard P. The distribution of the flora in the alpine zone. 1 [J]. New Phytologist, 1912, 11(2): 37-50.

[18] Adamic L, Adar E. How to search a social network [J]. Social Networks, 2005, 27(3): 187-203.

[19] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering [C]//Proceedings of the 14th international conference on neural information processing systems: natural and synthetic. Cambridge: MIT Press, 2002: 585-591.

[20] Ahmed A, Shervashidze N, Narayanamurthy S, et al. Distributed large-scale natural graph factorization [C]//Proceedings of the 22nd international conference on World Wide Web. New York: Association for Computing Machinery, 2013: 37-48.

[21] Zhu D, Cui P, Zhang Z, et al. High-order proximity preserved embedding for dynamic networks [J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(11): 2134-2144.

[22] Cao S, Lu W, Xu Q. Grarep: learning graph representations with global structural information [C]//Proceedings of the 24th ACM international on conference on information and knowledge management. New York: Association for Computing Machinery, 2015: 891-900.

[23] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: online learning of social representations [C]//Proceedings of the 20th ACM SIGKDD international conference on

knowledge discovery and data mining. New York: Association for Computing Machinery, 2014: 701-710.

[24] Grover A, Leskovec J. Node2vec: scalable feature learning for networks [C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. New York: Association for Computing Machinery, 2016: 855-864.

[25] Ribeiro LFR, Saverese PHP, Figueiredo DR. struc2vec: learning node representations from structural identity [C]//Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. New York: Association for Computing Machinery, 2017: 385-394.

[26] Tang J, Qu M, Wang M, et al. Line: large-scale information network embedding [C]//Proceedings of the 24th international conference on World Wide Web. Geneva: International World Wide Web Conferences Steering Committee, 2015: 1067-1077.

[27] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks [C]//Proceedings of the 5th international conference on learning representations. 2017: 1-14.

[28] Cen Y, Zou X, Zhang J, et al. Representation learning for attributed multiplex heterogeneous network [C]//Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining. New York: Association for Computing Machinery, 2019: 1358-1368.

[29] Chen T, Guestrin C. Xgboost: a scalable tree boosting system [C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. New York: Association for Computing Machinery, 2016: 785-794.

[30] Suh B, Hong L, Pirolli P, et al. Want to be retweeted? Large scale analytics on factors impacting retweeting in twitter network [C]//Proceedings of the 2010 IEEE second international conference on social computing. Washington, DC: IEEE Computer Society, 2010: 177-184.

Author Contributions: Sun Ran: Literature research, experiments and data analysis, paper writing; An Lu: Research framework determination, paper writing and revision.

Abstract: [Purpose/significance] This study aimed to predict whether microblog users would retweet or comment on microblog entries containing wanted information. We also evaluated the important features that affected the spread of wanted microblog entries to help public security departments improve their operation performance and enhance communication and cooperation between police and the public. [Method/process] Based on the characteristics of the wanted microblogging, we combined user features, time features and structure features, and extracted event features in microblog entries, such as location keywords, time keywords, the wanted level and so on. The Xgboost algorithm was used to calculate the importance of different features in the retweet and comment prediction. In combination with the features of transmission network and node attributes, we trained and evaluated a prediction model based on heterogeneous information network embedding. [Result/conclusion] The values of the AUC in retweeting and commenting datasets are 0.737 and 0.799 respectively. As the model integrated network structure characteristics and different nodes' attributes, it was closer to the heterogeneous information network in reality and had higher accuracy than the traditional link prediction model. In addition, the result of features' importance showed that the keyword features of the proposed event features had the highest importance among all the features that affected the prediction of microblog entries retweeted and commented.

Keywords: information dissemination; public security microblog; link prediction; graph representation learning; heterogeneous information network