PageRank-based Multidimensional Measurement of Weibo User Influence (Postprint)

Authors: Luo Fang, Xu Yang, Pu Qiumei, Qiu Qizhi

Date: 2019-04-01T00:00:00+00:00

Abstract

In recent years, the development of social networks has propelled research in multiple fields, such as public opinion monitoring, advertising recommendation, opinion leader identification, etc., while the influence measurement of social network users serves as the foundation for the aforementioned research. Taking Sina Weibo as the research object, this study aims to propose a Weibo user influence measurement algorithm with broader applicability and more comprehensive consideration of factors, integrating three-dimensional factors of user basic attributes, user interaction behaviors, and user post content into the traditional PageRank algorithm, and presents a multi-dimensional Weibo user influence measurement algorithm named MDIR (multi-dimension influence rank). Experimental results demonstrate that, compared with five other commonly used influence measurement algorithms, the MDIR algorithm can more comprehensively and accurately reflect the actual influence of Weibo users.

Full Text

Preamble

Vol. 37 No. 5
Application Research of Computers
ChinaXiv Cooperative Journal

Multi-dimensional Measure of Microblog User Influence Based on PageRank

Luo Fang[1], Xu Yang[1], Pu Qiumei[2], Qiu Qizhi[1]
(1. School of Computer Science & Technology, Wuhan University of Technology, Wuhan 430063, China;
2. School of Information Engineering, Minzu University of China, Beijing 100081, China)

Abstract: In recent years, the development of social networks has promoted research in many fields, such as public opinion monitoring, advertising recommendation, and opinion leader identification. The influence measurement of social network users forms the foundation for these research areas. Taking Sina Weibo as the research object, this paper aims to propose a microblog user influence measurement algorithm with broader applicability and more comprehensive considerations. By integrating three-dimensional factors—user basic attributes, user interaction behavior, and user blog content—into the traditional PageRank algorithm, we propose a multi-dimensional microblog user influence measurement algorithm called MDIR (multi-dimension influence rank). Experimental results demonstrate that the MDIR algorithm can more comprehensively and realistically reflect the actual influence of microblog users compared to five other commonly used influence measurement algorithms.

Keywords: microblog; user influence; PageRank; user behavior
Classification: TP391
doi: 10.19734/j.issn.1001-3695.2018.10.0798

0 Introduction

With the rapid development of Internet technology, application technologies centered on user interaction and personalized experience—represented by blog technology—have further propelled the transition from the Web 2.0 era, characterized by openness and sharing, to the Web 3.0 era featuring information fusion. Microblog, abbreviated from micro-blog, is a broadcast-style social networking platform based on a follow mechanism for sharing brief real-time information. According to data from the 41st "Statistical Report on China's Internet Development" released by CNNIC [1], by December 2017, China's microblog user base reached 376 million, driving continuous growth in user adoption to 40.9%, an increase of 3.8 percentage points from December 2016. On the microblog platform, each user can not only publish original microblog content but also freely forward, comment on, and like other users' posts. These mutual forwarding, commenting, and liking behaviors among different users facilitate the formation of microblog information dissemination networks. Additionally, microblog platforms feature low usage thresholds, concise and highly readable content, which enables faster information dissemination and broader influence.

Microblog user influence can be understood as a user's ability to cause behavioral changes in other users after posting a microblog. During microblog information propagation, different influential users' operations (such as forwarding and commenting) and attitudes (such as support or opposition) produce varying effects on the scope and depth of information dissemination. User influence measurement has important applications in network public opinion monitoring, advertising placement, and user recommendation. For public opinion monitoring, special measures must be taken for certain high-influence users to prevent the proliferation of public opinion. For advertising placement, selecting high-influence users as initial centers for ad dissemination can maximize propagation effects. For user recommendation, "opinion leaders" [2] in users' areas of interest are often default recommendation targets. In summary, user influence measurement plays an indispensable role in current hot research topics.

To reasonably measure microblog user influence, this paper improves upon the PageRank algorithm and proposes the MDIR (multi-dimension influence rank) algorithm. Compared with other influence measurement algorithms, the MDIR algorithm considers more comprehensive and reasonable influence factors, yielding more objective user influence rankings.

1 Related Work

Current research on social network user influence measurement by scholars both domestically and internationally generally falls into three categories based on user basic attributes, interaction behavior, and published blog content.

a) Influence measurement methods based on user basic attributes. User basic attributes represent the most primitive manifestation of user influence. Common basic attributes include follower count and post count, which are feature factors considered by most currently popular influence measurement algorithms. Cha et al. [3] selected three attributes from Twitter users—follower count, retweet count, and comment frequency—and calculated user influence based on node degree, comparing the correlation of the results. Their experiments showed that user follower count is not directly proportional to the number of retweets and comments on their posts. Mao et al. [4] analyzed microblog user activity, but their analysis relied solely on comment count without considering other factors or eliminating interference from "zombie fans" on the platform.

b) Influence measurement methods based on user interaction behavior. Interaction behavior between users is a direct manifestation of user influence. Common interaction behaviors include forwarding, commenting, mentioning, and liking. Zhang Hao et al. [5] proposed the UI algorithm based on user basic attributes and interaction behavior. They introduced the concepts of "user influence" and "user being-influenced." "User influence" was calculated based on attributes such as follower count, post count, and microblog retweet count, while "user being-influenced" was based on the interaction between fans and followed users, such as the percentage of a fan' s retweets and comments on a particular followed user' s posts relative to their total interactions with all followed users. However, the authors' quantification method for various interaction behaviors was normalization, which does not reflect actual microblog propagation patterns. Wang Ding et al. [6] built upon Zhang Hao et al.' s research by considering weight values for different behaviors. Their experiments provided detailed verification of ranking results from the data perspective, but they only constructed network topology based on user follow relationships without considering the large number of "zombie fans" on microblogs—users who follow others but do not interact with them, thus playing no role in information propagation. Sun Hong et al. [7] comprehensively considered users' actual microblog activity behaviors and the topological structure of the microblog network, proposing the MBUI-Rank algorithm. Their experimental results showed that this algorithm produced more accurate and objective user influence calculations. Qi Chao et al. [8] improved upon the PageRank algorithm and proposed the BWPR algorithm, constructing topological networks separately for forwarding, commenting, and mentioning behaviors. Although this algorithm achieved good results in experiments, it only considered the explicit feature of user interaction behavior without accounting for implicit user interest preferences.

c) Influence measurement methods based on blog content. Microblogs posted by users carry substantial information. Analyzing blog content can reveal topics of interest or emotional attributes of users, features widely applied in influence measurement. Weng et al. [9] proposed the TwitterRank algorithm based on Twitter data, which considered not only network structure but also analyzed topic similarity of each user' s tweets based on content. The algorithm summed users' influence values across each topic to obtain their overall network influence. Although this method achieved good experimental results, it only considered tweet count and topic similarity while ignoring user interaction behavior features. Shi Yakai et al. [10] measured inter-user influence by introducing user behavior in network topology structure and user interest similarity based on blog content, but they did not consider user basic attributes such as follower count and verification status.

Given the shortcomings of current research, this paper improves upon the PageRank algorithm and proposes a multi-dimensional microblog user influence measurement algorithm—MDIR (multi-dimension influence rank)—that integrates user basic attributes, user interaction behavior, and user blog content.

2.1 PageRank Algorithm

The PageRank algorithm is a classic algorithm widely used for webpage ranking on the Internet [11]. Its core idea is to study the network topology structure and calculate the number of inbound links (i.e., page link counts) to determine the page' s ranking order. The PageRank algorithm formula is shown in Equation (1):

$$PR(p_i) = (1-d) + d \sum_{p_j \in I(p_i)} \frac{PR(p_j)}{|O(p_j)|}$$

where $PR(p_i)$ represents the PageRank value of page p_i; $I(p_i)$ is the set of inbound links to page p_i; $|O(p_j)|$ is the number of outbound links from page p_j; and d is the damping coefficient, typically set to $d = 0.85$ [12].

When applying the traditional PageRank algorithm to microblog user influence measurement, users on the microblog platform are analogized to webpages in the Web network, considering only the follow and followed relationships between users. Direct application of the PageRank algorithm for user influence measurement has the following problems:

a) Subjective determination of initial PR values. The PageRank algorithm calculates initial PR values for webpages by averaging, which is not suitable for microblog user influence measurement. Due to differences between users (specifically in follower count, post count, verification status, etc.), using the averaging method to obtain initial PR values ignores the impact of users' own attributes on microblog propagation, resulting in less objective final user influence rankings.

b) Unreasonable PR value distribution method. Webpages evenly distribute their PageRank values to all pages they link to, which is clearly unreasonable for microblog users. Not all users treat all their followed users equally; most users only show interest in a portion of them and are willing to invest more attention.

c) Incomplete consideration of network structure. The PageRank algorithm' s application to microblogs is based on user follow relationships, making it heavily dependent on follower count. However, since followers may include many invalid zombie fans and silent fans, follower count cannot truly reflect user influence. Moreover, microblog users are dynamic and engage in various behaviors such as forwarding and commenting on posts. These behaviors significantly promote microblog propagation and occur not only between users with follow relationships but also through features like Sina Weibo' s "Weiba" and "Weitopic," where users can access and interact with other users' posts without following them.

2.2 MDIR Influence Measurement Factors

To address the problems identified above in applying the PageRank algorithm to user influence measurement, the MDIR algorithm proposed in this paper comprehensively considers three dimensions:

a) Factors based on user basic attributes. Users' basic attributes include follower count, post count, and verification status. Follower count and post count are relatively intuitive influence factors. Generally, more followers mean more users can see the posted microblog information, thus influencing more people. If follower count determines the scope of a user' s influence, then post count determines its depth. For follower groups of the same size, more posts mean each follower is exposed to the user' s microblog information more frequently, resulting in deeper influence. Verification status is a potential influence factor; verified Sina Weibo accounts have higher credibility, making their posts more likely to be commented on and forwarded, thus increasing user influence.

b) Factors based on user interaction behavior. On the microblog platform, posts published and forwarded by users subtly influence their fans, while fans' commenting, forwarding, and mentioning behaviors in turn promote information dissemination. More retweets mean wider information propagation; more comments indicate higher attention to the microblog; more mentions suggest higher prestige established by the user among their fan group.

c) Factors based on user blog content. Microblog information posted and forwarded by users reflects their personal interests. By mining blog content, we can obtain a topic distribution vector for each microblog. Combining all topic distribution vectors of a user yields a topic distribution vector representing the user' s interests, referred to as the interest distribution vector. Weng et al. [9] noted that on Twitter, users with more similar topics of interest are more likely to follow each other. Although their experiments were conducted on the Twitter platform, the results can be applied to microblog platforms due to structural similarities. Li Zhihong et al. [13] pointed out that microblog users have "homophily"—users are more inclined to establish relationships (forwarding, commenting, etc.) with others who share similar interests.

In summary, the MDIR algorithm proposed in this paper calculates users' initial influence values based on their basic attributes, addressing the subjectivity drawback of traditional PageRank. It calculates inter-user propagation willingness based on interaction behavior and blog content, improving the uniform influence distribution problem of traditional PageRank. By constructing a microblog information propagation network based on user interaction behavior, it effectively eliminates interference from numerous "zombie fans" and "silent fans" on user influence.

3.1 Microblog Propagation Network Construction

As described in Section 2.2, this paper constructs a microblog information propagation network (referred to as the microblog propagation network) through users' forwarding, commenting, and mentioning behaviors.

Definition 1: Microblog Propagation Network. Let the microblog propagation network be $G(V, E, B)$, where: - $V = \{v_i | i = 1, 2, 3, ..., n\}$ represents the set of nodes in the microblog propagation network, with v_i representing microblog users involved in one or more of the three behaviors: forwarding, commenting, or mentioning. - $E = \{(v_i, v_j) | v_i, v_j \in V, i \neq j \wedge v_i R v_j \vee v_i C v_j \vee v_i M v_j\}$ represents the set of edges in the microblog propagation network, where $v_i R v_j$ indicates user v_i forwarded user v_j' s microblog, $v_i C v_j$ indicates user v_i commented on user v_j' s microblog, and $v_i M v_j$ indicates user v_i mentioned user v_j in a microblog. - $B = \{B_{ij} | (v_i, v_j) \in E\}$ represents the set of edge weights in the microblog propagation network, where B_{ij} indicates the weighted sum of user v_i' s forwarding, commenting, and mentioning of user v_j.

3.2 Initial Influence Calculation

This paper calculates users' initial influence values based on three basic attributes: follower count, post count, and verification status. Since follower counts and post counts vary dramatically between users, this paper employs a logarithmic normalization method to process these values and reduce data span. The calculation formula is shown in Equation (6):

$$Init\_Infl(v_i) = \frac{\lg NF(v_i)}{\lg NF_{max}} + \frac{\lg NW(v_i)}{\lg NW_{max}} + Verify(v_i)$$

where $NF(v_i)$ represents user v_i' s post count; $NW(v_i)$ represents user v_i' s real follower count, i.e., user v_i' s in-degree in the microblog propagation network; NF_{max} represents the post count of the user with the most posts; NW_{max} represents the real follower count of the user with the most followers; $Verify(v_i)$ represents user v_i' s verification status. Lappas T et al. [15] noted that when user v_i is verified on Weibo, $Verify(v_i) = 0.5$ is most appropriate; when not verified, $Verify(v_i) = 0$.

3.3 MDIR Algorithm

The three behaviors have different contribution ratios to microblog propagation, so we should comprehensively consider the contribution proportion of each behavior. This paper uses α, β, and γ to represent the contribution ratios of forwarding, commenting, and mentioning behaviors, respectively, as shown in Equation (2):

$$B_{ij} = \alpha R_{ij} + \beta C_{ij} + \gamma M_{ij}$$

where R_{ij}, C_{ij}, and M_{ij} represent the counts of user v_i forwarding, commenting, and mentioning user v_j, respectively. For the values of α, β, and γ, this paper uses the order relation method, which involves pairwise comparison of each variable' s importance, aggregating the comparison results in a certain way, and finally calculating the values through computation.

First, construct a judgment matrix A, as shown in Equation (3):

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

The element a_{ij} in matrix A represents the relative importance of the i-th variable to the j-th variable. The properties of elements in matrix A are $a_{ij} = 1/a_{ji}$ and $a_{ii} = 1$. Based on the relative importance level table for variables summarized by Saaty et al. [14] (Table 1) and the relative importance relationships among the three behaviors, we can assign values to matrix A elements. Specifically, forwarding behavior is between slightly important and equally important compared to commenting behavior, so we take $a_{12} = 2$; forwarding behavior is between obviously important and absolutely important compared to mentioning behavior, so we take $a_{13} = 8$; commenting behavior is between equally important and slightly important compared to mentioning behavior, so we take $a_{23} = 2$. Substituting these into Equation (3) yields the final matrix shown in Equation (4):

$$A = \begin{bmatrix} 1 & 2 & 8 \\ 1/2 & 1 & 2 \\ 1/8 & 1/2 & 1 \end{bmatrix}$$

Solving the system of equations, the final values are shown in Equation (5):

$$\alpha = 0.727, \quad \beta = 0.182, \quad \gamma = 0.091$$

The MDIR algorithm recalculates the target user' s influence through a reasonable strategy, with its calculation formula shown in Equation (7):

$$MDIR(v_i) = (1-d) + d \sum_{v_j \in in(v_i)} ratio(v_j, v_i) \times MDIR(v_j)$$

where $in(v_i)$ represents the set of in-degrees of user v_i in the microblog propagation network; d is the damping coefficient, generally valued at 0.85; $ratio(v_j, v_i)$ represents the influence contribution proportion from user v_j to user v_i, calculated as shown in Equation (8):

$$ratio(v_j, v_i) = \frac{W(v_j, v_i)}{\sum_{v_k \in out(v_j)} W(v_j, v_k)}$$

where $out(v_j)$ represents the set of out-degrees of user v_j in the microblog propagation network; $W(v_j, v_i)$ represents user v_j' s propagation willingness to user v_i, expressed as the product of the interaction frequency $BF(v_j, v_i)$ and interest similarity $SIM(v_j, v_i)$ between v_j and v_i.

For inter-user interaction frequency, it can be calculated from the edge weight set B of the microblog propagation network G, as shown in Equation (9):

$$BF(v_j, v_i) = \frac{B_{ij}}{\sum_k B_{jk}}$$

For inter-user interest similarity, this paper uses the LDA model to extract users' topic distribution vectors and then calculates similarity between adjacent users based on the microblog propagation network. First, historical microblog information posted, forwarded, and commented on by users is aggregated into a document. The "document-user" collection is then used as model input, and the LDA model outputs a topic distribution vector corresponding to each user. The collection of all users' topic distribution vectors is denoted as matrix DT, where D and T correspond to the number of users and topics, respectively, with matrix element DT_{ij} representing the probability of user v_i on topic t_j.

Based on the "user-topic" matrix DT, this paper uses KL distance to calculate inter-user similarity. KL distance is a method for describing differences between two probability distributions P and Q, with its calculation formula shown in Equation (10):

$$KL(P||Q) = \sum_x P(x) \ln \frac{P(x)}{Q(x)}$$

As shown in Equation (10), the more similar two probability distributions are, the smaller their KL divergence. However, KL divergence is not symmetric, i.e., $KL(P||Q) \neq KL(Q||P)$. Therefore, to more appropriately represent similarity between microblog users, this paper first averages the KL distances and then takes the reciprocal, as shown in Equation (11):

$$SIM(v_i, v_j) = \frac{2}{KL(DT_i||DT_j) + KL(DT_j||DT_i)}$$

where DT_i and DT_j are the i-th and j-th rows of matrix DT, respectively, representing the topic distribution vectors of user v_i and user v_j.

The main processing steps of the MDIR algorithm are as follows:

Input: Microblog propagation network $G(V, E, B)$, damping coefficient d, threshold ε.

Output: Influence set S for all users.

1. Initialize microblog propagation network edge weights
2. for each edge $(v_i, v_j) \in E$ do
3. Calculate $B_{ij} = \alpha R_{ij} + \beta C_{ij} + \gamma M_{ij}$
4. end for
5. Initialize user influence values
6. for each user $v_i \in V$ do
7. $P\_MDIR(v_i) = 0$
8. end for
9. $S.Init()$ // Initialize set S
10. // User influence calculation, iteration ends when all user influence values converge
11. while $|S| < |V|$ do

for each user $v_i \in V$ do
// Influence value from previous round
$P = MDIR(v_i)$
// Calculate current round influence value using Equation (7)
$MDIR(v_i) = (1 - d) + d \sum_{v_j \in in(v_i)} ratio(v_j, v_i) \times MDIR(v_j)$
end for
if $|P\_MDIR(v_i) - MDIR(v_i)| \leq \varepsilon$ then
$S.add(v_i)$ // Save user v_i' s influence
end if
end while
return S

According to Equation (8), we can construct the probability transition matrix M, thereby transforming the MDIR algorithm' s solution process into a Markov process. The convergence conditions for a Markov process are: (a) matrix M is a stochastic matrix; (b) matrix M is an irreducible matrix; (c) matrix M is an aperiodic matrix. Based on Equation (8), matrix M directly satisfies convergence conditions (a) and (b). For convergence condition (c), this paper filters users before experiments to ensure the constructed microblog propagation network is a strongly connected graph, guaranteeing the irreducibility of matrix M. In summary, the MDIR algorithm proposed in this paper is convergent.

4.1.1 Dataset Selection

This paper uses Sina Weibo as the data source, crawling partial user microblog information under 12 hot topics on Sina Weibo during May 2014 as research data. Due to the complexity and redundancy of the crawled user information, this paper filters out users with fewer than 10 posts and fewer than 10 followings. The relevant data of the microblog propagation network constructed from the filtered data is shown in Table 2 .

4.1.2 Experimental Environment

Since constructing the microblog propagation network and iterative calculation of the MDIR algorithm consume significant time, this paper designs parallelization based on MapReduce for these two processes. The experiment uses four PCs to build a Hadoop cluster, each running 64-bit CentOS-7. The specific cluster configuration is shown in Table 3 .

4.2.1 Comparative Experiments

To make the experimental results more convincing, this paper selects currently popular or classic user influence measurement algorithms for comparison. First is the PageRank algorithm. Since the MDIR algorithm is an improved version based on PageRank, comparing it with the original PageRank algorithm better highlights the proposed algorithm' s advantages. Second is the BWPR algorithm. Qi Chao et al. [8] proposed the BWPR algorithm for calculating user influence based on PageRank, with its main improvement being the determination of influence distribution factors for fan users based on inter-user interaction behavior. Third is the TwitterRank algorithm. Weng et al. [9] proposed the TwitterRank algorithm for the Twitter platform, which mainly integrates user interest similarity on top of PageRank. Although this algorithm is based on Twitter, Twitter and Sina Weibo share similar structures, making it meaningful to generalize TwitterRank to the microblog platform. The final two are ranking algorithms based on user follower count and post count.

4.2.2 Evaluation Criteria

In real environments, microblog user influence has numerous measurement standards, making it difficult to establish a unified criterion. This paper adopts the M-fold cross-validation method proposed by Ding Zhaoyun et al. [16] to verify algorithm precision, recall, and F-value. First, we calculate the Top-K influence user sets I_x from each of the five comparison algorithms and the proposed MDIR algorithm. Then we construct the standard ranking set I_M as results voted correct by any M (where $1 < M \leq 6$) algorithms. The mathematical description of set I_M is shown in Equation (12):

$$I_M = \bigcup_{x \in \text{Combine}(6,M)} \bigcap_{i \in x} I_i$$

where $\text{Combine}(6, M)$ is the number of combinations selecting M algorithms from six algorithms. For example, given four algorithms A, B, C, D with Top-K influence user sets I_A, I_B, I_C, I_D, assuming $M = 2$, the standard set I_M is constructed as shown in Equation (13):

$$I_M = (I_A \cap I_B) \cup (I_A \cap I_C) \cup (I_A \cap I_D) \cup (I_B \cap I_C) \cup (I_B \cap I_D) \cup (I_C \cap I_D)$$

The precision P_A of algorithm A is calculated as shown in Equation (14):

$$P_A = \frac{|I_A \cap I_M|}{|I_M|}$$

The recall R_A of algorithm A is calculated as shown in Equation (15):

$$R_A = \frac{|I_A \cap I_M|}{|I_A|}$$

The F_A value of algorithm A is calculated as shown in Equation (16):

$$F_A = \frac{2 \times P_A \times R_A}{P_A + R_A}$$

4.3.1 Accuracy Verification

This paper simultaneously conducts cross-validation on six algorithms for $M = \{2, 3, 4\}$. When $M \geq 5$, the standard set contains fewer elements, and the precision and recall of each algorithm are similar, so these cases are ignored. For $M = 2, 3, 4$, we compare the precision, recall, and F-value of Top-K (where K takes values 100, 200, ···, 1000) influence users obtained by the six algorithms. Experimental results show that the MDIR algorithm achieves good performance under all three metrics.

Under the Top-K user scale, precision represents the ratio of correctly calculated Top-K users to the total number of users K. The three groups of experimental results shown in Figure 1 [Figure 1: see original paper] indicate that the MDIR algorithm' s precision under different user scales K and cross-validation folds M outperforms other comparison algorithms. When $M = 3$ and $M = 4$, the standard set contains fewer users, resulting in fewer intersecting elements between any algorithm' s result set and the standard set, so overall precision is lower compared to when $M = 2$.

4.3.2 Recall Verification

In the Top-K influence user scale, recall is the ratio of "correctly" calculated Top-K users to the number of influence users in the standard set, reflecting the degree to which influential users in microblogs are discovered. The recall distributions of the six algorithms under $M = 2, 3, 4$ are shown in Figure 2 [Figure 2: see original paper]. Experiments show that the MDIR algorithm performs well under all three M values, with particularly obvious distinction when $M = 3$. Since recall is jointly determined by $|I_A \cap I_M|$ and $|I_A|$, both increase as M increases, making recall changes less obvious.

4.3.3 F-value Verification

The F-value comprehensively considers both recall and precision, reflecting the overall degree of recall and precision of an algorithm. The F-values of each algorithm in cross-validation are shown in Figure 3 [Figure 3: see original paper]. Figure 3 shows that the MDIR algorithm proposed in this paper has obvious advantages in all three experiments. Meanwhile, the BWPR algorithm based solely on user interaction behavior shows decreasing efficiency as the statistical scale increases because user influence is also related to user basic attributes and blog content. The TwitterRank algorithm based on user blog similarity shows an upward trend in efficiency as the statistical scale increases, but its efficiency remains suboptimal because it does not incorporate user interaction behavior. The PageRank algorithm shows a similar trend to user follower count ranking, sufficiently demonstrating that PageRank heavily depends on user follower count, which is also a limitation of the original PageRank algorithm. User post count ranking and follower count ranking perform poorly overall due to their single-factor consideration. In contrast, the MDIR algorithm proposed in this paper comprehensively considers three-dimensional factors—user basic attributes, inter-user interaction behavior, and blog content—and specifically integrates them into the original PageRank algorithm, making its final results superior to other comparison algorithms in precision, recall, and F-value.

4.3.4 Convergence Verification

The SF-UIR algorithm is a microblog user influence measurement algorithm proposed by Wang Ding et al. [6] in 2018. This algorithm adds features of users' own behavior and fan user features in the network topology structure to PageRank, addressing the shortcomings of traditional PageRank' s poor objectivity and uniform influence transfer ratio distribution. Experiments verified the comprehensiveness and authenticity of the influence ranking results calculated by this algorithm. To further illustrate the applicability of the MDIR algorithm proposed in this paper, we compare the convergence speeds of the two algorithms.

Using the dataset described in Table 2, we parallelize both algorithms and calculate user influence values under the same convergence threshold, tracking and recording the iteration counts during the calculation process. Results show that the MDIR algorithm converges after 58 iterations, taking 34 minutes. For the SF-UIR algorithm, it converges after 65 iterations, taking 39 minutes. These experimental results demonstrate that the MDIR algorithm proposed in this paper has good convergence performance.

4.4 Algorithm Time Efficiency Verification

Based on the MDIR algorithm introduction in Section 3.3, since we need to calculate the inter-user influence contribution ratio $ratio(v_j, v_i)$, the time complexity of the MDIR algorithm proposed in this paper increases compared to the PageRank algorithm. However, the calculation of $ratio(v_j, v_i)$ is simple, involving only microblog text content and user interaction relationships, which can be extracted during preprocessing steps before the experiment. Additionally, the MDIR algorithm enables faster accumulation of influence values for high-influence users and faster convergence for low-influence users. To verify the time efficiency of the parallelized MDIR algorithm, this paper compares the time consumed from data reading to final convergence between the single-machine serial MDIR algorithm and the MapReduce-based parallel MDIR algorithm when processing the same data scale. Eight comparative tests were designed based on different user scales, with results shown in Figure 4 [Figure 4: see original paper].

The figure clearly shows that when processing 10,000 user data with smaller scale, the parallelized MDIR algorithm takes 10 minutes, slightly higher than the 7 minutes consumed by the serial MDIR algorithm. However, starting from 30,000 user data, the execution time of the parallelized MDIR algorithm changes relatively smoothly, while the serial MDIR algorithm grows almost exponentially.

When reaching a data scale of 80,000 users, the parallel algorithm requires almost half the time of the serial algorithm. If the data volume continues to increase, the serial algorithm may cause program abnormal termination due to excessive memory consumption, while the parallelized MDIR algorithm will maintain good performance.

5 Conclusion

By analyzing problems in traditional PageRank algorithms for user influence measurement, this paper conducted targeted feature selection from three dimensions and proposed the MDIR algorithm. Compared with current related research, the MDIR algorithm incorporates more influence factors. Considering users' own basic attributes makes calculation results more objective; considering inter-user interaction behavior and blog content makes results more reasonable; and considering network topology structure avoids unnecessary calculations and obtains more effective results. Experiments also demonstrate that the MDIR algorithm proposed in this paper achieves good performance across multiple influence evaluation metrics.

In future research, we will first consider user influence in different topics, mining opinion leaders who have high influence across multiple topics. Second, a large number of "water army" accounts in microblogs can interfere with calculation results, requiring the incorporation of microblog "water army" identification technology into the measurement algorithm to make final calculation results more objective.

References

[1] CNNIC. The 41st Statistical Report on the Development of China' s Internet Network [R]. Beijing: China Internet Network Information Center, 2018.

[2] Wang Chenxu, Guan Xiaohong, Qin Tao, et al. Modeling on Opinion Leader' s Influence in Microblog Message Propagation and Its Application [J]. Journal of Software, 2015, 26(6): 1473-1485.

[3] Cha M, Haddadi H, Benevenuto F, et al. Measuring User Influence in Twitter: The Million Follower Fallacy [C]// Proc of International Conference on Weblogs and Social Media. California: AAAI Press, 2010: 14.

[4] Mao Guojun, Zhang Jie. A PageRank-based Mining Algorithm for User Influences on Micro-blogs [C]// Proc of Pacific Asia Conference on Information Systems. Berlin: Springer, 2016.

[5] Zhang Hao, Liu Gongshen, Su Bo. A Computer Method for Microblogging Users Influence [J]. Computer Applications and Software, 2015, 32(3): 41-44.

[6] Wang Ding, Xu Jun, Duan Cunyu, et al. Improved User Influence Evaluation Algorithm Based on PageRank [J]. Journal of Harbin Institute of Technology: Social Sciences Edition, 2018, 50(5): 60-67.

[7] Sun Hong, Zuo Teng. Research on Microblog User Influence Algorithm Based on PageRank [J]. Application Research of Computers, 2018, 35(4): 1028-1032.

[8] Qi Chao, Chen Hongxu, Yu Hongtao. Method of Evaluating Microblog Users' Influence Based on Comprehensive Analysis of User Behavior [J]. Application Research of Computers, 2014, 31(7): 2004-2007.

[9] Weng Jianshu, Lim E P, Jiang Jing, et al. TwitterRank: Finding Topic-sensitive Influential Twitterers [C]// Proc of ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2010: 261-270.

[10] Shi Yakai, Ma Huifang, Zhang Di, et al. Microblog User Influence Algorithm Based on User Behavior and Content [J]. Application Research of Computers, 2016, 33(10): 2906-2909.

[11] Brin S, Page L. Reprint of: The Anatomy of a Large-scale Hypertextual Web Search Engine [J]. Computer Networks, 2012, 56(18): 3825-3833.

[12] Wang Chi, Chen Wei, Wang Yajun, et al. Scalable Influence Maximization for Independent Cascade Model in Large-scale Social Networks [J]. Data Mining and Knowledge Discovery, 2012, 25(3): 545-576.

[13] Li Zhihong, Zhuang Yunbei. Real-time Measurement Model of the Influence of Microblog Users with Double Dimensions Based on PageRank Algorithm [J]. System Engineering, 2016, 34(2): 128-137.

[14] Saaty T. Fundamentals of the Analytic Network Process-Multiple Networks with Benefits [J]. Journal of System Science and System Engineering, 2004, 13(3): 348-379.

[15] Lappas T, Terzi E, Gunopulos D, et al. Finding Effectors in Social Networks [C]// Proc of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2010: 1059-1068.

[16] Ding Zhaoyun, Zhou Bin, Jia Yan, et al. Topical Influence Analysis Based on the Multi-relational Network in Microblogs [J]. Journal of Computer Research and Development, 2013, 50(10): 2155-2175.

Note: Figure translations are in progress. See original paper for figures.