SCIPEDIA Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería

Revista Internacional Métodos numéricos para cálculo y diseño en ingeniería RIMNI

Key person analysis in persistent topic with online forum

Zehong LIN[1], Shuai WANG[2], xiaoxian Zhang[3]

1 College of Engineering, Harbin University, Harbin, China

2 Jilin Normal Univ, Coll Comp Sci & Technol, Siping 13600, Peoples R China

3 Computer Science and College,Changchun Institute of Technology, Changchun , Jilin

Abstract

The influence of users on online Forum should not be simply determined by the global network topology but rather in the corresponding local network with the user’s active range and semantic relation. Current analysis methods mostly focus on urgent topics while ignoring persistent topics, but persistent topics often have important implications for public opinion analysis. Therefore, this paper explores key person analysis in persistent topics on online Forum based on semantics. First, the interaction data are partitioned into subsets according to month, and the Latent Dirichlet Allocation (LDA) and filtering strategy are used to identify the topics from each partition. Then, we try to associate one topic with the adjacent time slice, which fulfills the criterion of having high similarity degree. On the basis of such topics, persistent topics are defined that exist for a sufficient number of periods. Following this, the paper provides the commitment function update criteria for the persistent topic social network (PTSN) based on the semantic and the sentiment weighted node position (SWNP) to identify the key person who has the most influence in the field. Finally, the emotional tendency analysis is applied to correct the results. The methods in real data sets validate the effectiveness of these methods.

OPEN ACCESS

Published: 21/12/2023

Accepted: 16/12/2023

DOI:
10.23967/j.rimni.2023.12.001

Keywords:
online Forum
persistent topic
key person
social network analysis

1. INTRODUCTION

As an electronic information service system on the Internet, an online Forum provides a public electronic forum on which each user can post messages and put forward views [1]. Online Forum gathers many users who are willing to share their experiences, information and ideas, and a user can browse others’ information and publish his/her own to form a thread through a unique registration ID [2].

Social network analysis (SNA) [3] can help us obtain the implicit characteristics of the users and information dissemination in a numerical manner. The forum topics are mainly divided into two categories: (1) emergency topics, which are characterized by a short duration with intense discussion; (2) persistent topics, characterized by long duration, typically closely related to one’s livelihood. Most studies have focused on the former, such as researching the discovery and prediction of online Forum hot topics and false information dissemination after emergencies [4]. There are two core issues that must be solved to identify key users in persistent livelihood topics: (1) extraction of persistent topics and (2) the identification of key users. To solve the first issue, we combine the time dimension and apply the latent Dirichlet allocation (LDA) topic model and the short text similarity assessment modelto discover the persistent topics [5]. To solve the second, SNA provides a series of node metrics (*e.g.,* central, prestige, trust and connectivity). The node position assessment, proposed by Przemysław Kazienko, is a very effective method for analysis, but it is more suitable for the global network while ignoring the semantic factors. Therefore, we provided the sentiment weighted node position algorithm (SWNP) and applied it to the persistent topic network to sort the users’ influence.

The algorithm must solve several problems. First, it must ensure that the extraction topic is related to the clustering results, so the algorithm uses the LDA model and the short text similarity assessment model for screening and gathering related posts while adopting adjacent time slice cross matching to ensure the topic sustainability on the timeline. After cataloging the posts, corresponding participants and replies relations, the persistent topic social network can be built and expressed as (PTSN= (V, E)), where V and E represent the nodes and their relationships [javascript:void(0);]in the local network, respectively. It then identifies the critical nodes in the local network, which have the greatest amount of influence on the specific topic and other users. After attempting different methods on real three-year online Forum data, the SWNP is provided and compared to the typical method.

The rest of the paper is organized as follows. We briefly review related work in section 2. We then present an overview of LDA and the short text similarity assessment model in section 3. In section 4, we propose persistent topic key person analysis in online Forum software, with detailed explanations. We discuss detailed experimental results on the research corpus in section 5, and we conclude this paper in section 6.

2. RELATED WORK

2.1 SNA in online Forum

Online Forum is an important platform for information dissemination. A user publishes a post to express his/her views on a given event, and others can browse the posts and create his/her own to form a thread through a unique registration ID [2]. A very important element of posting is the ability to add comments, which enables discussions. Accessibility to posts is

Correspondence: Zehong LIN (dloinng@163.com). This is an article distributed under the terms of the Creative Commons BY-NC-SA license

generally open, so anyone may read or comment. Online Forum is always busy with activity: every day, a large number of new users will register, and thousands of new posts and millions of new comments are written. The lifetime of posts is very short, and the relationships between users are very dynamic and temporal, providing a large amount of semantic information to explore intensely [6].

Research on online Forum is primarily rooted in public opinion guidance, sociology, linguistics and psychology, while data mining with technology is less frequently employed. However, nearly all online Forum websites record some basic statistics, which lend themselves well for data analysis and important findings. This network model, consisting of the board, posts and comments, can be analyzed by SNA to find the most important or influential users. Around such users, groups that share similar interests will form.

There are many types of online Forum: campus online Forum, commercial online Forum, professional online Forum, emotional online Forum and individual online Forum. We chose the comprehensive Tianya forum as the basis for our research because persistent livelihood topics are more likely to occur in this active social online Forum. There is some research on the Tianya forum datasets, such as the opinion leader algorithm based on users' interests, but the accuracy depends on the quality of the interest field [7].

2.2 Topic discovery

Some results have been achieved in the network topology and topic propagation models, but they are still new. Previous studies can be mainly divided into three categories: (1) the first type of research mainly focuses on the distribution of users to reveal their dynamic characteristics. (2) The second type of research mainly focuses on the topics of discovery and prediction. Wang [8] improved the information diffusion model based on topic influence, and proposed the topic diffusion trend prediction method based on the reply matrix. (3) The third type of research studies semantic communities for user characteristic analysis. Z.Bu [9] proposed a sock puppet detection algorithm that combines authorship-identification techniques with link analysis.

Compared to research around sudden hot issues, few studies consider persistent livelihood topic discovery, evolution and traceability. With the rapid dissemination of information, people's livelihood topics will continue to ferment, and they will inevitably have an impact on the management of the networked public opinion without necessary regulatory and counseling.

2.3 Key Person Extraction

There are two separate approaches to key person extraction in social networks: those based on context roles and those based on social network structure. The most common key person extraction methods rely on various centrality measures for each separate node. However, these algorithms lack a holistic view, and the node position in the social community is determined by its neighborhoods, such as in degree prestige and degree centrality. Other algorithms are more global, such as proximity prestige, rank prestige, node position, eccentricity and closeness centrality. Much of this research has been applied to different domains (e.g., influence spread, public opinion analysis, and terrorist group analysis) [10].

In fact, the user influence is not solely determined by the overall network topology but confirmed by the local network structure and semantic relationships among active users. No existing algorithm can meet this demand, and because the entire network is not the best choice, the influence field must be determined before the key person may be extracted. The PTSN is a semantic-based local network, so we propose a node position algorithm combined with semantic information to identify key persons.

3. PERSISTENT TOPIC EXTRACTION IN SOCIAL NETWORK

To obtain the persistent livelihood topic in online Forum, two basic methods are introduced here. The first is the LDA model for extracting topics, and the second is the short text similarity assessment model to distinguish persistent topics and emergency ones.

3.1 LDA

In statistical natural language processing, one common way of modeling the contributions of different topics to a document is to treat each topic as a probability distribution over words, viewing a document as a probabilistic mixture of these topics. Given documents *D* containing *K* topics and *N* unique words: W= {w_1,w_2,...,w_N}, where each w_i belongs to some document d_i, and $z_i$is a latent variable indicating the topic from which the *i*th word was drawn. The complete probability generative model is defined as follows:

$$\left\{\begin{array}{c} \theta^{(d)} \sim Dirichlet(\alpha) \\ z_i \mid \theta^{(d_i)} \sim Multinomial(\theta^{(d_i)}) \\ \varphi^{(z)} \sim Dirichlet(\beta) \\ w_i \mid z_i, \varphi^{(z_i)} \sim Multinomial(\varphi^{(z_i)}) \end{array}\right. \tag{1}$$

Here, the hyperparameters α and β are mainly used to control the sparsity of the distribution. According to this model, every word $w_i \in W$ will be assigned to a latent topic z_i.

In a corpus, the goal of LDA is to extract the latent topic z through evaluating the posterior distribution. The sum in the denominator involves T^nterms, where n is the total number of word instances in the corpus. However, it does not factorize, so Gionline Forum sampling is now widely adopted. Gionline Forum sampling estimates the probability of a word belonging to a topic, according to the topic distribution of the other words. At the beginning of the sampling, every word is randomly assigned to a topic as the initial state of a Markov chain. Each state of the chain is an assignment of values to the variables being sampled. After enough interations, the chain approaches the target distribution and the current values are recorded as the expected probability distribution. In the end, it obtains the topic T={t_1,t_2,...t_z} and t_i={(t_{i1},p_{i1}),...,($t_{ij,pij}$),...,(t_{iN},p_{iN})} where t_{ij} may appear in t_i with probability p_{ij}.

3.2 Short text similarity assessment model

Quan provides a short text on similarity computing methods based on probabilistic topics [11]. The algorithm uses a topic model on the short text feature vectors, then determines the semantic similarity by computing the cosine between the vectors. We improve the model for the online Forum title text using the minimum threshold and therefore require less computing cost.

The model analyzes topics in two adjacent time periods, so let the former topics T_{former}={t_1,...t_i,...t_n} and corresponding topic vector t_i={(t_{i1},p_{i1}),...($t_{ij,pij}$) ,...(t_{iN},p_{iN})}, the later ones T_{later}={t_1,...t_k,...t_m},and t_k={(t_{k1},p_{k1}),...($t_{kl,pkl}$),...(t_{kM},p_{kM})}. The existing similarity formula is not suitable, so equation (2) is used for this work. To obtain high similarity degree topics in adjacent time slots, *n×m* calculation time is need, i.e., each topic is required to match with all topics in another time period.

$$s_{i,k} = \sum_{word \in t_i \cap t_k} min(p(word)) \qquad (2)$$

where $s_{i,k}$ is the similarity degree of topic t_i and t_k, which equals the sum of the minimum probability of the words appearing in both topics. If $s_{i,k}$ is larger than threshold σ_1, it means the two topics are similar. If a topic continues over some periods, it can be considered a persistent topic.

Meanwhile, the size of topic t_i in a certain period can be measured by (3).

$$r_{i,d} = \sum_{word \in t_i \cap d} p(word) \qquad (3)$$

Here, the post title *d* is used to match the keywords of topic t_i, and then the sum of all the probabilities of success matching is the relevancy. If $r_{i,d}$ is greater than σ_2, then the post is related to the topic. The thresholds σ_1 and σ_2 will be confirmed in the experiment.

4. ANALYSIS OF KEY PERSON IN PERSISTENT TOPIC WITH ONLINE FORUM

Two important issues in social network analysis are individual role and social position. Analysis of key persons in persistent topics with online Forum is further considered.

Due to the time characteristics, the gathered data should be partitioned into subsequent *N* periods with the same length, which are always labeled from 0 to *N-1*, and these periods are separable or partly overlapped. In the experimental studies (see Sec. 5), we assumed that they have a length of 30 days.

The LDA was used to obtain the topics in each period and extract the persistent topic across multi-periods through the similarity assessment. Then, for each persistent topic, the social network was generated and the fundamental SNA measures were calculated to identify the key person.

In the first step, the interaction data are partitioned into the subsets by month, and the LDA and filtering strategy are used to identify the topics from each partition. Then, the algorithm attempts to associate one topic with another from the neighboring period while fulfilling the criterion of having a similarity degree larger than σ_1. On the basis of this comparison, the persistent topics that exist for the sufficient number of periods are defined. Following this, the algorithm uses sentiment weighted node positions in the interaction data to identify the key person who has the most influence in the field.

The algorithm consists of six subsequent steps:

Step 1. The gathered text stream should be partitioned into subsequent N periods with the same length.

Step 2. Extract topics, and then record the relevant posts, users, reply rates etc. To achieve this, the algorithm LDA described in Sec. 3.1 is used. The *z* topics will be obtained in every time slice.

Step 3. Simplify the topics using the filtering strategy. For a given period *ts*, after the attribute filter and topics set *N=Topic*(*ts*) are identified, each topic contains its keywords and the corresponding probability. The topic will be retained once it meets one of the following filtering strategies:

(1) The number of posts related to the topic ($r_{i,d}$ larger than σ_2) is greater than or equal to 10. σ_2 is 0.05 in Sec. 5, that is, the post is related if a post title contains a keyword of a certain topic.

(2) The total number of users involved in the topic is greater than or equal to 10% of the active users of the period;

(3) The topic's “hotness” (click times divided by the number of active users) is greater than or equal to 10%;

(4) The response rate (the total participation of users divided by the total number of clicks) is greater than or equal to 30%;

Step 4. The topics in adjacent time are crossed matching. To achieve this, the short text similarity assessment model described in Sec. 3.2 is used. The σ_1 is 0.09 in Sec. 5.

Step 5. Identify persistent topics that exist for a minimum period of time. Urgent topics have small time spans and simple network evolutions, which do not belong to the persistent topics that this article focuses on. Ephemeral topics do not last for more than two periods, but some may occur in the junction of two periods, so the ts_{req} is defined as the minimum period for topic longevity. In these experiments, it is assumed that ts_{req}=3.

A set of topics, which consists of similar topics during the periods *j,j+1,...,j+s,* the number of topic-related posts and users are respectively defined as follows:

$$\{ \begin{matrix} POST_i = \sum_{t=j}^{j+s} POST_i(ts) \\ USER =^i \cup_{t=j}^{j+s} USER_i(ts) \end{matrix} \qquad (4)$$

Step 6. The key persons in the persistent topic are identified using SWNP. First, *PTSN*= (V, E) is built. The traditional node position algorithm has an experimental basis for large-scale data [12] but does not consider interest, topics and sentiment factors, so this paper provides *SWNP*(*x*) to estimate the importance of the node *x* in a local network.

Every term/phrase is manually assigned a value between 0 and 1 according to its tone. Oppressive terms range between 0.5 and 1, and a higher value corresponds to a greater degree of oppression. Supportive terms range between 0 and 0.5, and a smaller value corresponds to a greater degree of support. If the phrase is neutral, it is assigned a value of 0.5.

For a given comment from one ID to the other, we can determine the implicit orientation by counting the number of positive or negative words in it (if there are several emotional words in one comment, we take the average).

$$sentiment_{i,j} = \frac{\sum_{k=1}^{n^j} O_{k,j}}{n^j} \qquad (5)$$

where $O_{k,j}$ is the emotional word weight in comments from *i* to *j*, n^j is the number of emotional words in all the comments from *i* to *j*. The $sentiment_{i,j}$ >0.5 indicates a negative emotional tendency with a negative commitment function, and $sentiment_{i,j}$ ≤0.5 indicates a positive commitment function. The *SWNP*(*x*) can be redefined as follows:

$$SWNP(x)=(1-\epsilon)+\epsilon\sum_{y\in Y_x}SWNP(y)\,|\,C(y\rightarrow x)\,| \qquad (6)$$

>where Y_x *x's* nearest neighbors, *i.e.,* nodes that are in the direct relation to *x*; $C(y\rightarrow x)$ is the commitment function; *ε is* the constant coefficient in the range [0,1], and its value denotes the openness of node position measurement on external influences: a smaller value indicates that *x's* node position is more static and independent while a larger value means that the node position is more influenced by others.

The value of the commitment function C(*y*→*x*) in *PTSN* must satisfy the following set of criteria:

(1) The value of commitment is from the range [-1;1]:$(x,y\in V)C(y\rightarrow x)\in[-1;1]$.

(2) The sum of all commitments' absolute values must be equal to 1 in the case of each node in the network: ?$(x\in V)\sum_{x\in V}|C(y\rightarrow x)|=1$.

(3) The commitment to oneself is 0:$(x\in V)$ $C(x\rightarrow x)=0$.

(4) If there is no relationship from *y* to *x*, then C(*y*→*x*) =0.

(5) If a member *y* is not active with respect to anybody and other n members x_i, $i=1,\ldots,n$ are active with respect to *y*, then instead of satisfying the above criterion 4, the commitment value is distributed equally among all of *y's* acquaintances x_i, *i.e.,*$(x_i\in V)$ $C(y\rightarrow x_i)=1/n$.

Some comments in online Forum are always presented without a clear view, so based on this consideration, we believe that comments labeled with strong emotions tend to communicate more information and therefore should attract greater attention. As such, if *x*→*y* shows the reply relationship from *x* to *y*, we assume that comments with strong emotions should transmit a greater commitment than just a passing glance. There are three specific cases:

(1) if *x*→*y* meets the strong negative (0.8, 1] or strong positive [0,0.2), $n_2=4n_1$;

(2) if *x*→*y* satisfies the general negative (0.6,0.8] or general positive [0.2,0.4), $n_2=2n_1$;

(3) if *x*→*y* belongs to relatively neutral [0.4, 0.6], $n_2=n_1$;

where n_1 is the total response number from *x* to *y*, and n_2 is the comments numbers after emotional weighted.

The value of the commitment function C(*x*→*y*) can be evaluated as the normalized sum of all activities from *x* to *y* in relation to all activities of *x*:

$$C(x\rightarrow y)=\frac{A(x\rightarrow y)}{\sum_{j=1}^{m}A(x\rightarrow y_j)} \qquad (7)$$

where *m* is the number of all nodes within the *PTSN*, A(*x*→*y*) is the function that denotes the activity of node *x* directed to node *y*, such as the number of comments from *x* to *y*. Using the emotional weighted $n_2$instead of $n_1$is more conducive to finding an important node in the semantic web.

5. EXPERIMENTS

5.1 Data Set

The dataset is from the Tianya forum (http://focus.tianya.cn), which is a popular bulletin-board service in China. It includes more than 300 boards, and the total number of registered user identifications (IDs) exceeds 32 million. Since its introduction in 1999, it has become the leading social-networking site in China due to its openness and freedom. We selected the Tianya By-talk board and collected data between January[javascript:void(0);] 2011 and December 2013 including 325288 users, 102756 posts and 4524756 replies. Among all the users, 12701 of them wrote at least 1 post in the period, 3724 wrote at least 2 posts and 573 at least 5 posts. Taking into consideration the users who wrote at least 1 post, the average number of posts for each user was 8.09. Most of the users' behavior consisted of replying to posts or even just browsing; the average number of comments for all users equaled 13.91, which was still greater than 8.09.

The largest hot topic post has 6571731 clicks and 66274 comments, and 71929 posts have more than 5 comments. In 2011 through 2013, 176346 users wrote at least one comment, 110261 wrote more than one comment, and the most active user posted 10276 comments. Considering only posts that have at least 5 comments, the average number of comments per post was 62.91. In 2011, users wrote 10324 posts and 400571 comments (38.8 comments/post); in 2012, they wrote 31146 posts and 1326819 comments (42.6 comments/post); and in 2013, they wrote 61286 posts and 2797366 comments (45.6 comments/post).

5.2 Identification of the topics in specific periods

We used the LDA to identify the topic in specific months, setting α=0.5, β=0.1, topic number *Z*=50 and Gionline Forum sampling iterations to 1000. Not all of each month's topics are related to the livelihood issues that this article focuses on, so these topics are omitted by the attribute filter described in Sec. 4.

After applying this attribute filter, there were a total of 978 topics with an average of 27 topics per month. The minimum number occurred in the 10th month with 9 topics and maximum was in the 6th month with 37 topics. To analyze the size of each topic, Fig. 1 shows the statistics on the number of topics related posts. Setting σ_2=0.05 retains more valid data for extracting persistent topics that is, if a title contains a keyword related to a certain topic, it will be retained. Eighty-two percent of retained topics ranged in size from 61 to 150 related posts.

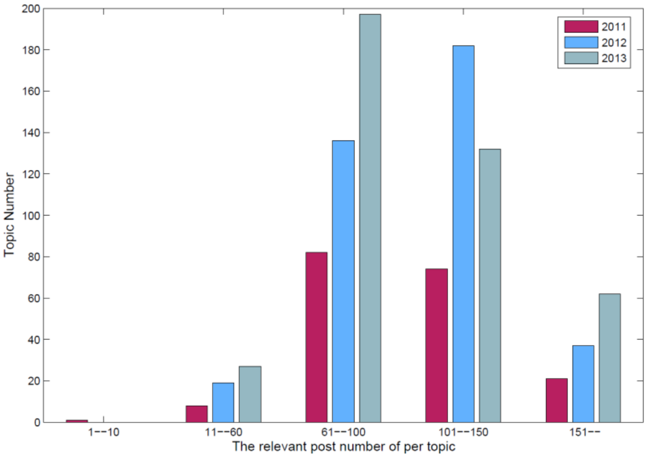

Fig. 1 The related posts number for each topic

5.3 Identification of the persistent topic

The next analysis concerned the identification of the persistent topics, which must exist over a given period. The persistent topic number is affected by σ$_1$. The keyword of a topic always has a frequency of approximately 0.05, while a similarity of 0.1 means the topics have at least two keywords, and then it can be certain that they are in fact the same. Experiments have proven that an important turning point occurs at σ$_1$=0.09, corresponding to the 18 relatively persistent topics. The persistent topics have high accuracy and quality by manual validation.

There are 18 persistent topics with 4637 related posts. A total of 91281 users (28% of total users) were present in the following analysis, which greatly reduces the data size for further analysis. There are 257 related posts per persistent topic, and according to the minimum period (three months), they have only 86 posts per topic per month. This number is less than the size of the general topics retained in Sec. 5.2, which also reflects the persistent topics that do not have a high post rate, click rate or response rate and instead have their own characteristics of long duration.

5.4 Analysis of duration time of the persistent topic

Thirteen persistent topics (72%) lasted for 3 months, which is the minimum duration necessary to consider the topic as a persistent one in our analysis. Four persistent topics lasted exactly 4 months, and the longest lasted 5 months. The distribution of persistent topics is relatively uniform; only in May 2013 (the 29th month) and June 2013 (the 30th month) was there four co-existing persistent topics. Data analysis found that this was during the time of graduation season and the university entrance exam. Additionally, youth films such as “So Young” and singing reality shows such as “X Factor” and “Chinese Idol” caused such topics to remain hot and evolve continuously around this time, although topic evolution is beyond our research.

At the same time, the obtained persistent topics have high diversity, for there is little overlap within the same period. Though two topics with interval time may be similar, they are apparently two different events. Issues concerning graduation, college entrance examinations and employment will repeat themselves every year in different fashions, although this type of topic evolution analysis is not within the scope of this study. Therefore, this algorithm ensures diversity among the persistent topics.

5.5 Persistent topic social network（PTSN）

The goal on the next analysis is to count the posts and the users in the persistent topic. Table 1 shows the basic information of 18 persistent topics, and there are 257 posts and 5071 users per persistent topic. Social network *PTSN*=(*V*, *E*) can be built for each persistent topic, where *V* is a finite set of registered users who take part in the topic (*i.e.,* the *IDs*). *E* is a finite set of social relationships (*i.e.,* posts and replies).

Table 1. The basic information of the persistent topic

No.	*periods*	*posts*	*users*	*No.*	*periods*	*posts*	*users*
1	3	246	4835	10	3	268	4981
2	3	202	5124	11	3	227	5671
3	3	316	6147	12	4	249	5019
4	4	340	5410	13	4	342	3957
5	3	198	4105	14	3	179	6105
6	3	279	4716	15	4	283	4281
7	3	248	6124	16	3	305	6289
8	5	336	4398	17	3	269	5042
9	3	187	5627	18	3	163	3450

5.6 Node position iterative data processing

The experiments revealed that the number of iterations necessary to calculate the node positions for all users in each *PTSN* depends on the value of the parameter *ε*, see Eq.(6): the greater the value of *ε*, the greater the number of iterations (Fig. 2). Each node was initialized in *PTSN* with *SWNP*=1 and the stop condition *τ*=0.00001. The iterative processing of *SWNP* uses six different *ε* (0.01, 0.1, 0.3, 0.5, 0.7, and 0.9) for comparative analysis. Because the given 18 *PTSN*s have similar sizes, their tendencies are similar.

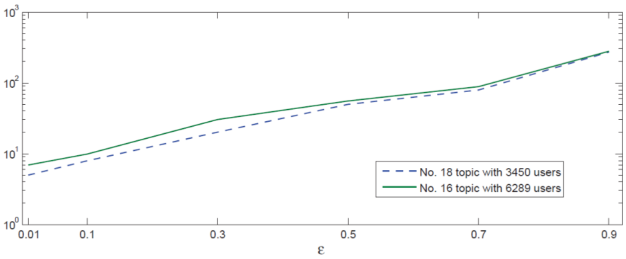

Fig. 2 The number of iterations in relation to *ε*

The experiments revealed that the *SWNP* does not increase the number of iterations and processing time compared with *NP*. Because the sentiment analysis only gives every comment a one-off score to determine its emotional inclination (positive or negative), linearly enhancing the corresponding comments without a change in the iteration processing simply adds a linear time complexity to the iterative process. For a clearer demonstration, the *SWNP* value generally refers to the absolute value except in particular emphasis. Next, the distribution characteristics of the *SWNP* are analyzed to discover the important nodes.

5.7 Distribution characteristics of SWNP

Experiments analyze the distribution characteristics of SWNP in 18 PTSN, and Fig. 3 gives the average SWNP and their standard deviation in *No.16* and *No.18 PTSN* with different *ε*. The average SWNP does not depend on *ε*, and it can be formally demonstrated that the SWNP equals approximately 1 in all cases. On the other hand, the standard deviation differs substantially depending on *ε*: the greater the *ε*, the greater the standard deviation. Namely, the *SWNP* value has increased disproportionately with bigger *ε*, which has been proven by the experimental data.

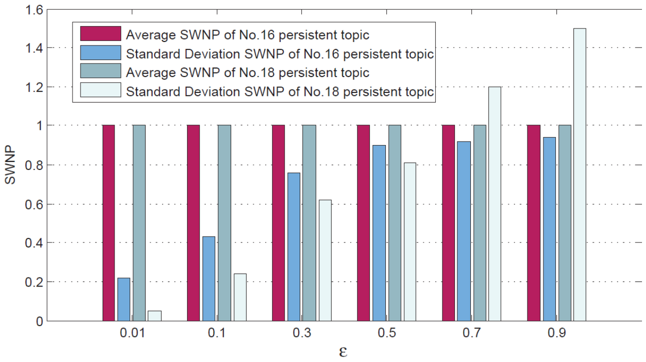

Fig. 3 Average *SWNP* and their standard deviations in relation to *ε*

The distribution characteristics of *SWNP* are determined by its network topology structure; for example, the standard deviation variation tendency of *No.18* is more noticeable than *No.16*. This result indicates the greater difference of *SWNP* in *No.18 PTSN*, as there are a few nodes with ultra value. It can also be noted that the average *SWNP* over 81% of users is less than 1. This result means that only a few members exceed the average value that equals 1. This result also shows that the members' *SWNP* difference increased for greater *ε*, and it is valid for all the 18 *PTSN*. The *No.18 PTSN* has the standard deviations of the most obvious change: while *ε*=0.9, fewer than 1% of users have a *SWNP*>1, and these users are clearly important. Fig. 4 shows the percentage of users with *SWNP*≥1 and *SWNP*≥2 within *No.18* and *No.16 PTSN* in relation to *ε*.

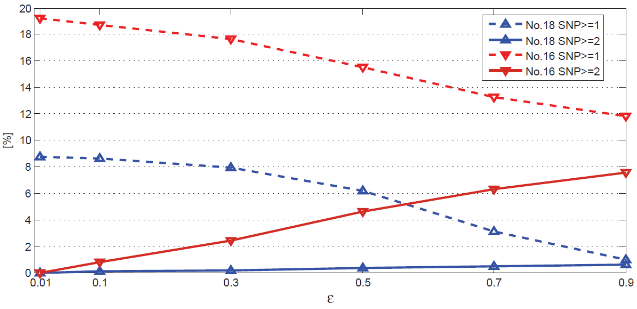

Fig. 4 The percentage of users with *SWNP*≥1 and *SWNP*≥2 within No.18 and No.16 *PTSN* in relation to *ε*

It can be seen that the different *PTSN*s have the same *SWNP* distribution trend, with the *SWNP*≥1 nodes decreasing and *SWNP*≥2 nodes increasing. The average percentage of nodes with *SWNP*≥2 is 4.7% in all the 18 *PTSN* (*No.16* with 7.54% and *No.18* with 0.57%). This conclusion can help us identify the important nodes in persistent topic social networks. The percent of *SWNP*≥1 and *SWNP*≥2 are 3.12% and 0.49% in *No.18 PTSN* while *ε*=0.7, so it can be assured that 3.12% users are active users and the 0.49% users are key person in this topic. In fact, the greater the *ε*, the more distinguishable the results, but the larger number of iterations directly influences the processing time. Generally, the parameter is determined by the different network scales, but the nodes with high *SWNP* values do not necessarily represent key persons, as the adjacent nodes may pass a lot of negative energy (if the commitment function is less than 0). Therefore, sentiment analysis is needed to actually identify the key persons.

5.8 The Top N key persons in PTSN

Extracting the Top *N* key persons in *PTSN* is achieved through a ranking nodes process based on the importance degree. The algorithm sorts the nodes according to the *SWNP*, and then modifies the list using the emotional attributes. The comparing algorithms mainly used are *IDC* (Indegree Prestige Centrality), *ODC* (Outdegree Prestige Centrality) and *PR* (PageRank). *IDC* is based on the indegree number, so it takes into account the number of members that are adjacent to a particular member of the community, as follows: $IDC(x) = i(x)/(m-1)$, where *m* is the number of nodes in the network, and *i*(*x*) is the number of members from the first level neighborhood that are adjacent to *x*. In other words, more prominent people receive more nominations from members of the community. *ODC* takes into account the outdegree number of the member *x* for edges that are directed to the given node, as follows: $ODC(x) = o(x)/(m-1)$, where *o*(*x*) is the number of the first level neighbors to *x*. On the other hand, users who have low outdegree centrality are not very open to the external world and do not communicate with many members. *ODC* and *IDC* are the simplest and most intuitive measures that can be used in network analysis. Google uses *PR* to rank the pages in its search engine to measure the importance of a particular page to the others. Table 2 gives the top 10 important nodes using different methods in the *No.18 PTSN* with 3450 nodes.

Table 2 Top 10 users in *No.18 PTSN*

Pos.		*ε=0.01*	*ε=0.1*	*ε=0.3*	*ε=0.5*	*ε=0.7*	*ε=0.9*	*IDC*	*ODC*	*PR*
1	ID	122756	**22614**	307146	307146	8961	8961	14864	7996	22614
	Val.	1.834	**5.634**	12.458	15.762	20.546	25.874	0.214	0.130	0.0133
2	ID	235523	307146	8961	8961	307146	307146	248153	200416	70064
	Val.	1.627	5.301	12.041	15.240	20.121	21.371	0.197	0.124	0.0105
3	ID	57681	8961	20547	20547	20547	196349	84134	14267	89712
	Val.	1.526	4.982	11.878	15.046	16.824	18.627	0.182	0.120	0.0092
4	ID	**22614**	20547	**22614**	**22614**	**22614**	276482	33224	14864	6401
	Val.	**1.475**	4.870	**11.526**	**14.872**	**16.345**	18.064	0.176	0.106	0.0088
5	ID	307146	57681	276482	57681	57681	20547	313375	9246	85216
	Val.	1.404	4.633	10.954	14.534	15.015	17.349	0.172	0.095	0.0080
6	ID	8961	276482	235523	122756	122756	235523	51229	81820	578
	Val.	1.377	4.315	10.467	13.801	15.246	16.202	0.154	0.087	0.0078
7	ID	276482	122756	122756	235523	235523	70064	52166	241357	3601
	Val.	1.306	4.157	10.348	13.008	15.205	15.977	0.143	0.084	0.0076
8	ID	20547	235523	57681	276482	196349	57681	7996	120608	14027
	Val.	1.288	3.946	8.002	11.328	14.548	15.279	0.132	0.079	0.0073
9	ID	196349	196349	70064	70064	314627	122756	921712	122412	39240
	Val.	1.270	3.415	7.856	11.340	13.851	14.675	0.121	0.070	0.0070
10	ID	70064	70064	196349	196349	70064	314627	810204	15246	317540
	Val.	1.256	3.097	6.912	9.282	11.067	12.544	0.117	0.059	0.0067

The important node ranking is relatively stable when used with different values of *ε*. As the simplest and most intuitive measures that can be used in network analysis, the *ODC* and *IDC* have low accuracy. The node sort result of *PR* is a good one, but there are two main shortcomings: (1) without the commitment function in *PR*, all links have the same weight and importance. The PR is distributed by its outdegree and gives no considerations to the strength of the interaction. (2) No sentiment analysis to identify the effective opinion leaders. After ranking, we analyzed the ratio of negative emotions ($C(y \rightarrow x) \leq 0$) for the selected node (*e.g.,* ID22614). Due to 73% of the commitment functions being less than 0, the node is an active user but not a positive advocate, which helps to control the spread of false information as well as in public opinion analysis and other follow-up work.

The *SWNP* can identify key persons in the specific topic, so it cannot be evaluated by the typical methods, such as Google's search engine or the users ranking list by computing click rate. To further confirm the stability of the algorithm, the top 10 users in different *PTSN* are used to analyze their community duties and real occupational information. By checking and calculating though artificial verification, a high level of accuracy is maintained.

6. CONCLUSIONS

Two main independent approaches are provided in the paper for identifying key persons in online Forum: (i) discovery of the persistent topics and (ii) extraction of the key person using SWNP. Identifying persistent topics mainly combines the *LDA* model and similarity model on the timeline. *SWNP* is a new method of node position analysis, which takes into account both the node position of the neighbors and the strength and emotional tendency of connections between network nodes. The data are from Tianya forum, as indicated in Sec 5. The experiment shows that the number of persistent topics is far less than urgent topics, and most of them exist for approximately 3 months with uniform distribution on the timeline. In the established *PTSN*, the high influence persons are extracted through the *SWNP* iterative calculation and have been analyzed by contrast experiment and artificial verification. The weighted sentiment in *SWNP* mainly reflects that the emotional intensity can be converted to the number of comments, which changes the value of the commitment function and the iterative results. In addition, negative emotions can be used to alter the notion of the key persons to a certain extent, such as discovery of the different ideas of factions, online water armies and false advertisement publishers.

REFERENCES

1. Kulunk A , Kalkan S C , Bakirci A , et al. Session-Based Recommender System for Social Networks' Forum Platform[C]// 2020 28th Signal Processing and Communications Applications Conference (SIU). 2020.

2. Lu D, Lixin D. "Sentiment Analysis in Chinese BBS," *Intelligence Computation and Evolutionary Computation*, 2013, pp.869-873.

3. Wasserman, S. and Faust, K. "Social Network Analysis: Methods and Applications," *Cambridge University Press*, New York, 1994.

4. Liu, Hong, and Bi Wei Li. "Hot Topic Detection Research of Internet Public Opinion Based on Affinity Propagation Clustering," *Computer Informatics Cybernetics and Applications*, 2012, pp. 261-269.

5. Zhang, Fang, G. Y. Si, and Pi Luo. "Study on rumor spreading model based on evolution game," *Journal of System Simulation*, 2011, pp. 1772-1775.

6. Gregory A L , Piff P K . Finding uncommon ground: Extremist online forum engagement predicts integrative complexity [J]. PLOS ONE, 2021, 16.

7. Liu J, Cao Z, Cui K, "Identifying Important Users in Sina Microblog," Multimedia *2012 Fourth International Conference on. IEEE*, 2012, pp. 839-842.

8. Wang Wei, "The Study of Topic Diffusion State Presentation and Trend Prediction within BBS," *Procedia Engineering*, vol.29, 2012, pp.2995-3001.

9. Z. Bu, Z. Xia, J. Wang, "A sock puppet detection algorithm on virtual spaces," Knowledge Based Syst. 37.2013 :366-377.

10. Carrington P., Scott J., Wasserman S., "Models and methods in Social Network Analysis," *Cabrige University Press, Cambrige*, 2005.

11. Quan XJ, Liu G, Lu Z, Ni XL, Liu WY. "Short text similarity based on probabilistic topics," *Knowledge and Information Systems*, vol.25, 2010,pp.473-491.

12. Kazienko, P., Musiał, K. and Zgrzywa, A, "Evaluation of Node Position Based on Email Communication," *Control and Cybernetics*, 38 (1), 2009, pp.67-86.

Funding: The work was supported by the Self-funded project of Harbin Science and Technology Bureau(2022ZCZJCG023,2023ZCZJCG006),the Harbin Social Science Federation special topic(2023HSKZ10),the Research Project of Education Department of Jilin Province (No:JJKH20210674KJ and No:JJKH20220445KJ) and 2022 Science and Technology Department of Jilin Province(20230101243JC).