Knowledge-Based Systems 110 (2016) 30–48

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

freepaper.me FREE paper

SCIFNET: Stance community identification of topic persons using friendship network analysis

Zhong-Yong Chen[a], Chien Chin Chen[b,*]

[a] Department of Information Management, National Taiwan University, Address: Room 709, College of Management No. 1 Building, No.1, Sec. 4, Roosevelt Rd., Taipei City 106, Taiwan (R.O.C.)

[b] Department of Information Management, National Taiwan University, Address: Room 414, College of Management No. 2 Building, No.1, Sec. 4, Roosevelt Rd., Taipei City 106, Taiwan (R.O.C.)

ARTICLE INFO

Article history:
Received 26 December 2015
Revised 7 July 2016
Accepted 8 July 2016
Available online 11 July 2016

Keywords:
Text mining
Community detection
Clustering

ABSTRACT

A topic that involves communities with different competing viewpoints or stances is usually reported by a large number of documents. Knowing the association between the persons mentioned in the documents can help readers construct the background knowledge of the topic and comprehend the numerous topic documents more easily. In this paper, we investigate the stance community identification problem where the goal is to cluster important persons mentioned in a set of topic documents into stance-coherent communities. We propose a stance community identification method called SCIFNET, which constructs a friendship network of topic persons from topic documents automatically. Stance community expansion and stance community refinement techniques are designed to identify stance-coherent communities of topic persons in the friendship network and to detect persons who are stance-irrelevant about the topic. The results of experiments based on real-world datasets demonstrate the effectiveness of SCIFNET and show that it outperforms many well-known community detection approaches and clustering algorithms.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

With the prevalence of telecommunication technologies and the explosive growth in medium digitization, there are now enormous amounts of information on the Internet. As a result, people worldwide can easily obtain information about the latest topics, such as global economic trends, political events, and sports tournament results via the Internet. Usually, people are interested in topics that involve communities with different competing viewpoints or stances. However, they are often overwhelmed by the large number of topic documents that cover every detail of different stance communities. For example, in the topic about the selection of a new International Monetary Fund (IMF) president in 2011, Google News[1] collected hundreds of topic documents that reported the development of the campaign. Although the documents covered all perspectives on the topic (i.e., from the interactions between the candidates to the viewpoints of the general public), readers generally had difficulty assimilating the enormous amount of information in the documents. To ease the burden of reading so many topic documents, several topic mining techniques have been developed. For instance, Nallapati et al. [35] grouped topic documents into clusters, each of which presents a theme of a topic; Feng and Allan [22] extracted informative sentences from themes to summarize a topic; and Chen and Chen [5,6] further organized themes and summaries chronologically to depict the storyline of a topic. The techniques successfully condense the content of a topic. However, readers still need to invest a lot of time in digesting the generated summaries if they are not familiar with the topic.

A topic is basically associated with persons, times, and places [35]. Learning the associations between the persons mentioned in a set of topic documents (called topic persons hereafter) can help readers construct the background knowledge of the topic and digest the information quickly. For instance, in the above mentioned topic about the new IMF president selection, if readers had known that Angela Merkel supported Christine Lagarde (i.e., they are detected in the same community), they would have understood why she said "Christine Lagarde is an ideal embodiment of economics."

In this paper, we investigate the stance community identification problem, which involves clustering topic persons into stance-coherent communities. For instance, given the documents about the selection of the new IMF president in 2011, the stance community identification method discovers communities of persons, which represent the camps of the different candidates running for election, as shown in Fig. 1. Identifying stance communities of

* Corresponding author.
E-mail addresses: d98725003@ntu.edu.tw (Z.-Y. Chen), patonchen@ntu.edu.tw (C.C. Chen).

[1] https://news.google.com/

http://dx.doi.org/10.1016/j.knosys.2016.07.015
0950-7051/© 2016 Elsevier B.V. All rights reserved.

Fig. 1. The selection of the IMF president in 2011.

topic persons is a new research area, and to the best of our knowledge, only Chen et al. [7,8] have addressed the stance community identification problem. They proposed using Principal Component Analysis (PCA) [4]. Specifically, they examine the signs of the entries in the eigenvector associated with the largest eigenvalue to recognize stance communities of topic persons. The method can only handle two-stance topics; however, in practice, many topics involve more than two stances. Here, we present a novel stance community identification method called SCIFNET (Stance Community Identification based on Friendship NETwork), which analyzes a set of topic documents to identify stance communities and the corresponding persons in a topic. First, SCIFNET constructs a friendship network in which the nodes represent topic persons. The co-occurrence of the persons in the topic documents, the documents' stance orientation, and the co-neighboring level between nodes are leveraged to define the friendship strength between persons (i.e., the edge weights). We model stance community identification as a community detection task and design an objective function to evaluate the results. Stance community expansion and stance community refinement techniques, which are based on the objective function, are designed to iteratively cluster topic persons into stance-coherent communities and detect persons that are stance-irrelevant about the topic of interest. Their convergence proofs are presented such that the identification result converges to a local optimum. Evaluations based on real-world topics demonstrate the effectiveness of SCIFNET, and show that it outperforms well-known clustering and community detection approaches.

The proposed method has the following advantages over the current community detection research. First, most iterative clustering-based community detection methods, such as those in [20,31,48], would suffer the early merging problems of a node in a network tending to be merged (clustered) with a community simply because it is close to the community's seed. To get rid of this type of problem, we design the stance community refinement which iteratively refines the detected communities. Second, nodes in a social (friendship) network can play different roles. Differing from the overlapping node, bridge node, and hub node investigated in [13,14,21], the proposed method is able to identify stance-irrelevant nodes which stand for persons neutral to the stances of a topic. Finally, since topic persons may have opposing orientations, the constructed friendship network could have negative edges. While several community detection methods, such as [13,21,32] analyze network structures to infer communities, our method further examines edge signs to correctly detect stance communities of topic persons.

The remainder of this paper is organized as follows. In the next section, we review related works. Then, we describe SCIFNET in detail, and demonstrate its efficiency in experimental section. Final section contains our conclusions.

2. Related work

Our research is related to community detection [41]. Given a network of interests, the community detection task involves identifying sub-networks, each of which represents a coherent community [12,24,36,39]. For instance, given a social network, community detection methods identify groups of people with similar preferences [41]. The identified communities are useful to comprehend various social phenomena, such as epidemic spreading [43], and human interactions [14,15,40,42]. Basically, the methods partition a network into sub-networks based on the principle that maximizes the association between the nodes in each sub-network, while minimizing the association between the sub-networks [45]. In the following sub-sections, we review the existing community detection approaches, namely, the eigen-based community detection approach and the iterative clustering approach.

2.1. Eigen-based community detection approach

One of the techniques used in the eigen-based approach is spectral clustering, which exploits the eigenvectors of a Laplacian matrix [18] to find appropriate partitions of a network. The Laplacian matrix of a network is derived by subtracting the adjacency matrix A from the diagonal matrix D. The entry $a_{i,j}$ in A is 1 if node i and node j are connected, and 0 otherwise; and the entry $d_{i,i}$ in D is the degree of node i in the network. Shi and Malik [45] modeled image segmentation as a community detection problem. They represented an image as a network and employed the eigenvector associated with the second smallest eigenvalue (i.e., the fielder vector) of the Laplacian matrix to identify significant image segments. Ding et al. [16] used spectral clustering to cluster a set of documents and constructed a word-document matrix X in which the entries are the mutual information [33] between the words and documents. Then, a document network is constructed by considering each document as a node. The connection between nodes is represented by the weighted matrix $W = X^T X$; and the network is partitioned by using the fielder vector of the matrix W. The authors also introduced the Mcut metric to evaluate the partitioned network. The metric is integrated with a linkage-based refinement technique to improve the quality of the network partition.

One limitation of the above methods is that they usually make balanced cuts when partitioning a network; that is, the communities detected in the network need to be of a similar size. However, in practice, communities are of different sizes and magnitudes, so the balanced cut requirement is irrational [38,51]. To overcome this limitation, White and Smyth [51] developed a spectral clustering algorithm that maximizes the modularity [39] of a network partition. Specifically, given a network partition, the modularity measures the ratio of the edges within communities to all

freepaper.me FREE paper

the edges in the network and subtracts the expected number of connected nodes from the ratio in the same communities. The larger the value, the better will be the quality of the network partition. White and Smyth formulated the modularity maximization problem as a quadratic assignment problem and solved it analytically by using an eigen-decomposition method. The method constructs an eigenvector matrix U_K in which the columns are the eigenvectors of the matrix L_Q derived from the modularity maximization problem. Then, the row vectors of U_K are clustered by using the K-means algorithm [33] to find an appropriate network partition. Newman [38] developed an efficient algorithm to detect communities in a network. Initially, the algorithm treats a node as a community and constructs a modularity matrix B, where entry $B_{i,j}$ denotes the modularity between the community i and the community j. Next, it examines the signs of the entries in the principal eigenvector of B to identify the affiliation of the nodes. To refine the detected communities, i.e., the partitioned sub-networks, the algorithm then examines the modularity changed by moving nodes between communities and moves all the nodes that increase the modularity. Anchuri and Magdon-Ismai [2] investigated signed networks in which nodes are connected by positive or negative edges. They modified the modularity to incorporate negative edges into it and constructed a modularity matrix for a signed network. Communities are detected by examining the signs in the matrix's eigenvector associated with the largest eigenvalue. In addition, a refinement method based on the modified modularity is developed to calibrate the membership of the nodes.

2.2. Iterative clustering approach

The other community detection approach is iterative clustering. Girvan and Newman [24] devised an iterative clustering algorithm that measures the betweenness of edges to detect communities. The betweenness of an edge denotes the number of shortest paths between node pairs that run through the edge. The algorithm iteratively decomposes a network by removing the edge with the largest betweenness until a specific number of communities have been detected. Subsequently, Newman and Girvan [39] utilized the modularity they designed to enhance the betweenness-based community detection method. Meanwhile, Newman [37] developed a modularity-based community detection algorithm called FastModularity. Given a network, the algorithm first initializes each node as a community. Then, it iteratively merges communities until the modularity of the detected communities reaches a local optimum. The drawback with using the modularity for community detection is that the measure ignores missing edges in a community. In other words, it only measures how well the discovered community structure fits the existing edges [9]. In reality, it is difficult to obtain all the information about the analyzed network. Consequently, informative edges may be missing from the network and that would degrade the community detection performance. To resolve the problem, Chen et al. [9] proposed a measure called Max-Min modularity, which considers missing edges to improve the quality of community detection. Xu et al. [53] also designed an iterative clustering algorithm called SCAN (Structural Clustering Algorithm for Networks) for community detection. Initially, SCAN computes the ratio of co-neighbors between every node pair. A node is regarded as the core of a community if the number of high co-neighbor ratios between it and other nodes is also high. The algorithm expands communities from the identified core nodes by iteratively integrating the nodes' neighbors into the communities. It is noteworthy that the algorithm can identify hub nodes, which function as bridges to different communities. In social network analysis, hub nodes may play an important role in viral marketing. Yang et al. [55] developed an iterative bipartition method called FEC (Finding and Extracting a Community) for detecting communities in a signed network. The method first conducts a random walk on the network to measure the probability of reaching a node. Afterward, an adjacency matrix is constructed by sorting the nodes in accordance with their reaching probabilities. The algorithm then iteratively identifies a cutting point in the matrix to bipartition the network such that the positive edges within the partitioned sub-networks and the negative edges between the sub-networks are dense. Chen et al. [10] developed the L measure, which leverages the internal and external degrees of nodes in a community. A detected community is regarded as good if its L value is large. The authors also designed a two-phase algorithm that expands the communities in a network iteratively. The first phase identifies nodes whose degrees are higher than the average internal degree of a community. Then, in the second phase, the identified nodes are merged into a community if their inclusion increases the community's L value. The results of experiments show that the communities detected by using L are superior to those detected by using the modularity. Traag and Bruggeman [46] adjusted the modularity to include the negative edges of a network, and incorporated the modified modularity into Potts model [52] to detect communities. Yang et al. [54] integrated the link structure with content analysis for community detection. They introduced a popularity-based link model to measure the strength of the links between nodes and employed an iterative EM process to learn the membership of the nodes. Gao et al. [23] developed a generative model called CODA (Community Outlier Detection Algorithm) to detect communities and their outliers in a network. The model uses hidden Markov random fields [4] to compute the importance of a network structure. In addition, the nodes in the network are sorted by an objective function and low-ranked nodes are labeled as outliers. Eustace et al. [20] invented a two stage algorithm to detect local communities. In the first stage, the method randomly selects nodes as the seeds of local communities. Then it employs the alpha-close function to expand the communities with their close neighborhood. In the second stage, the local community begins to merge with the other local community which satisfies the threshold of using the beta-close function, and then merge together to construct the final communities.

Recently, a number of studies have started to detect communities with overlapping nodes [14,17,21,31,49,50]. For instance, Wang and Li [48] considered a node as a core vertex if the node has a high degree. They developed a community detection method which initializes core vertices as community seeds and then employed an intimate degree function to iteratively absorb new nodes into the communities. If an absorbed node has the same intimate degree for two (or more) communities, it will be labeled as an overlapping node of the communities. Li et al. [31] selected core vertices by using expert-defined rules and employed the absorbing degree to merge new nodes into existing communities. Similarly, a node is deemed as an overlapping node of communities if the absorbing degrees of the node to the communities are the same. Cui and Wang [13] extracted communities with overlapping nodes from a bi-partite network. The authors considered the node with the minimum degree as a key bi-community, and used the intimate degree to expand bi-communities with overlapping nodes. Notably, communities with overlapping nodes can also be detected by the matrix decomposition methods [19,21] and the maximal clique extraction techniques [14,15,32].

Our research differs from existing community detection because the networks analyzed by community detection approaches are usually pre-defined. In contrast, the friendship networks of topic persons in our method are derived automatically from topic documents. Our research further considers friendship orientations, and identifies friendly and opposing associations between topic persons in the friendship networks.

freepaper.me FREE paper

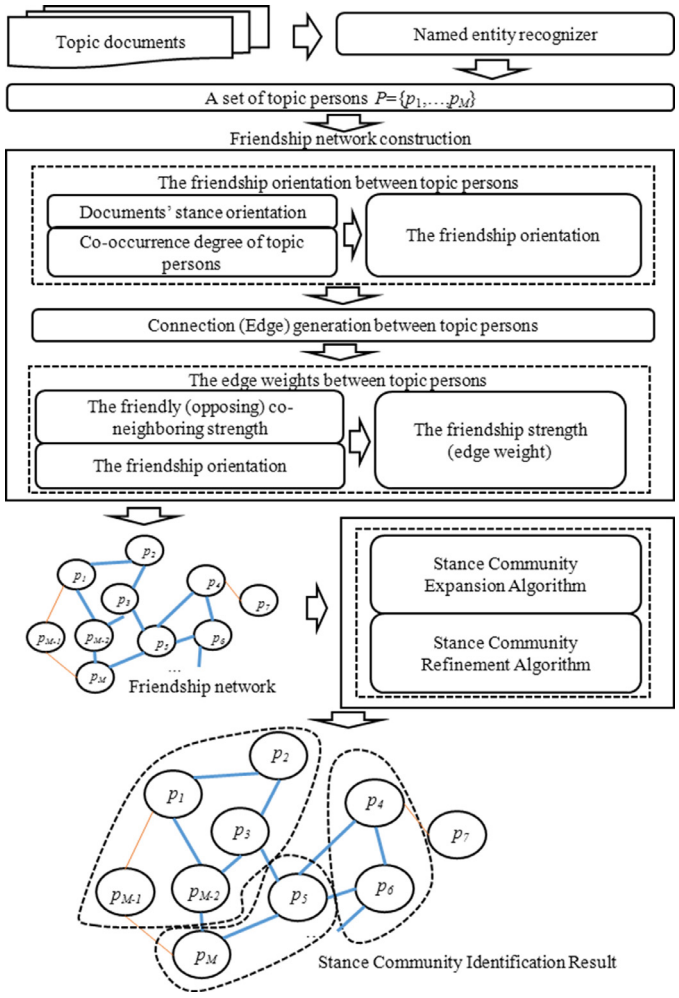

Fig. 2. The system architecture.

3. Methodology

We proposed a stance community identification method, SCIFNET, which clusters the persons mentioned in topic documents into stance-coherent communities. Fig. 2 shows SCIFNET's system architecture, which is comprised of three components: *friendship network construction*, *stance community expansion*, and *stance community refinement*. Specifically, given a set of documents reporting a topic with K stance communities, SCIFNET first extracts the topic persons mentioned in the documents. Then, it constructs a friendship network of the topic persons based on the co-occurrence of the persons in the documents and the stance orientation of the documents. Next, the stance community expansion process considers the stance community identification of topic persons as a community detection task and iteratively expands the K stance communities in the friendship network. In the last phase, the stance community refinement algorithm improves the identification result in accordance with an objective function, which measures the stance coherence of the detected communities. Note that a difficult issue in community detection is to determine the number of communities in a network and the issue is generally related to the optimization research regarding the cardinality of a clustering [33]. Like many community detection methods (e.g., [16,23,54]), we assume that the number of stance communities (i.e., K) is known in advance and concentrate on designing effective stance community expansion and refinement operations. Also, at the first attempt to model topic person stance identification as a community detection problem, we simply assume that each person belongs to a single stance. Later, the selection strategy of stance community cardinality will be discussed and a modified SCIFNET for detecting overlapping communities will be also provided. In the following subsections, we describe each system component in detail. We also show that using the components increases the value of the objective function such that the stance community identification result converges to a local optimum.

3.1. Friendship network construction

Let $D=\{d_1, d_2, \ldots, d_N\}$ be a set of topic documents, and let $P=\{p_1, p_2, \ldots, p_M\}$ be a set of topic persons mentioned in D. The friendship network construction generates a friendship network $G=\{P, E\}$, where the topic persons in P form the network's nodes; and $E=(p_i, p_j)$ is a set of edges that indicate the friendship orientation of the topic persons (i.e., whether the association between the persons is friendly or opposing). Generally, it is difficult to discover friendship orientations from text. However, Harris [26] observed that text units with opposing meanings seldom co-occur in the same context. In addition, Kanayama and Nasukawa [28] showed that text units with the same sentiment tend to occur (not occur) jointly to make the contexts coherent. Hence, the correlation coefficient [29], which measures the co-occurrence degree of topic persons in D, is probably a good measure for discovering the friendship orientation between topic persons. Nevertheless, we found that topic documents sometimes cover controversial issues. In the documents, people with different stances strongly criticize each other. Thus, only considering the co-occurrence degree of topic persons in D may overestimate the friendship of rivals and degrade the performance of stance community identification. Intuitively, topic persons who frequently co-occur in stance-friendly (stance-opposing) documents may have a friendly (opposing) association. To quantitate the stance orientation of a topic document, we adopt Turney and Littman [47]'s method and compute the stance weight of a document as follows:

$$sw_d = \sum_{word_i \in d} \log\left(\frac{\prod_{word_j \in Fwords} count(word_i, word_j) \cdot \prod_{word_k \in Owords} count(word_k)}{\prod_{word_j \in Fwords} count(word_j) \cdot \prod_{word_k \in Owords} count(word_i, word_k)}\right), \quad (1)$$

where sw_d represents the stance weight of document d; and *Fwords* and *Owords* are, respectively, sets of words with stance-friendly and stance-opposing semantics compiled by linguistic experts. The function $count(word_i, word_j)$ returns the number of documents in which $word_i$ and $word_j$ co-occur in our topic corpus. Basically, the equation utilizes pointwise mutual information (PMI) to compute the stance weight of a document. The stance weight sw_d is positive if d's content is strongly associated with *Fwords*, and negative if the content is strongly associated with *Owords*. We design the following *stance-oriented correlation coefficient* (SOCOR), which incorporates the stance weight into the correlation coefficient:

$$socor(p_i, p_j) = \left[\sum_{d \in D_{friendly}} sw_d * (p_{i,d} - \overline{p}_{i,friendly}) * (p_{j,d} - \overline{p}_{j,friendly}) + \sum_{d \in D_{opposing}} sw_d * (p_{i,d} - \overline{p}_{i,opposing}) * (p_{j,d} - \overline{p}_{j,opposing})\right] \Big/ \sqrt{\sum_{d \in D_{friendly}} \left[\sqrt{sw_d} * (p_{i,d} - \overline{p}_{i,friendly})\right]^2 + \sum_{d \in D_{opposing}} \left[\sqrt{|sw_d|} * (p_{i,d} - \overline{p}_{i,opposing})\right]^2} * \sqrt{\sum_{d \in D_{friendly}} \left[\sqrt{sw_d} * (p_{j,d} - \overline{p}_{j,friendly})\right]^2 + \sum_{d \in D_{opposing}} \left[\sqrt{|sw_d|} * (p_{j,d} - \overline{p}_{j,opposing})\right]^2}, \quad (2)$$

freepaper.me FREE paper دانلود کننده مقالات علمی

where $D_{friendly} \subseteq D$ is a set of topic documents whose stance weight is positive; $D_{opposing} \subseteq D$ is a set of topic documents whose stance weight is negative; and $\bar{p}_{i,friendly}$ and $\bar{p}_{i,opposing}$ are the average frequencies of p_i occurring in $D_{friendly}$ and $D_{opposing}$ respectively. Like the correlation coefficient, the range of $socor(p_i,p_j)$ is within [−1,1]. It is zero if the occurrences of p_i and p_j in D are independent of each other. However, if p_i and p_j tend to co-occur in stance-friendly (stance-opposing) documents, the $socor(p_i,p_j)$ is positive (resp. negative). Next, we define the friendship orientation in terms of the stance-oriented correlation coefficient.

Definition 1. The friendship orientation:

The friendship orientation between p_i and p_j is denoted as $socor(p_i,p_j)$ and $-1 \le socor(p_i,p_j) \le 1$.

We utilize SOCOR to construct the edge set E. In addition, to consolidate relationships between topic persons, we define a friendship orientation threshold θ. An edge (p_i,p_j) is established if $socor(p_i,p_j) > \theta$ or $socor(p_i,p_j) < -\theta$.

Jeh and Wisdom [27] and Antonellis et al. [3] demonstrated that the association between nodes in a network is proportional to their co-neighboring level. In other words, the greater the overlap between the neighbors of two nodes, the higher will be the likelihood that the nodes are associated with each other. In our research, however, edges indicate either friendly orientations or opposing orientations. To measure the co-neighboring strength, we define two types of neighbors, namely, friendly neighbors and opposing neighbors.

Definition 2. The Friendly Neighbors:

Let $p_i \in P$. The friendly neighbors of p_i, denoted by $\Gamma_{friendly}(p_i)$, form a set of nodes whose friendship orientations to p_i are larger than θ. Formally, $\Gamma_{friendly}(p_i) = \{p_j \in P |\ socor(p_i,p_j) > \theta\}$.

Definition 3. The Opposing Neighbors:

Let $p_i \in P$. The opposing neighbors of p_i, denoted by $\Gamma_{opposing}(p_i)$, form a set of nodes whose friendship orientations to p_i are smaller than $-\theta$. Formally, $\Gamma_{opposing}(p_i) = \{p_j \in P |\ socor(p_i,p_j) < -\theta\}$.

In Definitions 4 and 5, we employ the Jaccard coefficient to measure the friendly co-neighboring strength and the opposing co-neighboring strength respectively.

Definition 4. Friendly Co-neighboring Strength:

The friendly co-neighboring strength between p_i and p_j is denoted by $\gamma(p_i,p_j)$:

$$\gamma(p_i, p_j) = \frac{|\Gamma_{friendly}(p_i) \cap \Gamma_{friendly}(p_j)|}{|\Gamma_{friendly}(p_i) \cup \Gamma_{friendly}(p_j)|}.$$

Definition 5. Opposing Co-neighboring Strength:

The opposing co-neighboring strength between p_i and p_j is denoted by $\omega(p_i,p_j)$:

$$\omega(p_i, p_j) = \frac{\left|\Gamma_{opposing}(p_i) \cap \Gamma_{opposing}(p_j)\right|}{\left|\Gamma_{opposing}(p_i) \cup \Gamma_{opposing}(p_j)\right|}.$$

Clearly, if two nodes share several friendly (opposing) neighbors, their friendly (opposing) co-neighboring strength is strong. Finally, we combine the friendship orientation with the co-neighboring strengths, and define the friendship strength, i.e., the edge weight, as follows.

Definition 6. Friendship Strength:

The friendship strength, denoted by $\delta(p_i,p_j)$, represents the weight of edge (p_i,p_j).

$$\delta(p_i, p_j) = (socor(p_i, p_j) + 1)^{\frac{\gamma(p_i,p_j)+\omega(p_i,p_j)}{2}+\beta},\ if socor\left(p_i, p_j\right) > \theta.$$

$$\delta(p_i, p_j) = -(|socor(p_i, p_j)| + 1)^{(1-\frac{\gamma(p_i,p_j)+\omega(p_i,p_j)}{2})+\beta},\ if\ socor\left(p_i, p_j\right) < -\theta.$$

For friendly orientations (i.e., $socor(p_i,p_j) > \theta$), the friendly and opposing co-neighboring strengths function as an exponent to amplify the friendly relationships between nodes. We utilize a parameter $\beta \ge 1$ to ensure that the exponent is not less than 1; and we add 1 to a friendly orientation so that the base is greater than 1. As the enemies of foes may be friends, the friendship strength of p_i and p_j is strong and positive if they have a friendly orientation and share a lot of friendly and opposing neighbors. If p_i and p_j have an opposing orientation (i.e., $socor(p_i,p_j) < -\theta$), their friendship strength is negative. However, p_i and p_j may not fight against each other if they have many friends and adversaries in common. The negative friendship strength is thus diminished if the friendly and opposing neighbors of p_i and p_j overlap a great deal.

3.2. The objective function of SCIFNET

After constructing the friendship network of a topic, we identify stance communities in the network.

Definition 7. Stance Communities:

The stance communities $<c_1, c_2, \ldots, c_K>$ form a set of node clusters in the friendship network G such that $c_m \subseteq P$ and $c_m \cap c_n = null$ for $m \ne n$.

In general, community detection methods partition the nodes of a network into clusters (i.e., communities) in accordance with the principle that maximizes the association between the nodes in each cluster, while minimizing the association between the clusters [45]. We define the following objective function to identify a coherent stance community identification result.

$$C = \underset{<c_1,c_2,\ldots,c_K>}{\arg\max} \sum_{community:m}^{K} \left[\sum_{p_i,p_j \in c_m, i<j, (p_i,p_j)\in E} \delta(p_i, p_j) \right] - \sum_{community:m,n,m<n}^{K} \left[\sum_{p_i \in c_m, p_j \in c_n, (p_i,p_j)\in E} \delta(p_i, p_j) \right]. \quad (3)$$

To maximize the objective function, the identified stance communities need to maximize the first term of Eq. (3) and minimize the second term simultaneously. In other words, the stance community identification method seeks a set of stance communities that maximize the friendship strength within communities (the first term of the objective function) and minimize the friendship strength between communities (the objective function's second term).

3.3. Stance community expansion

Fig. 3 shows the proposed stance community expansion algorithm, and Fig. 4 provides an example of stance community expansion. In the algorithm, the symbol $P_{unlabeled}$ represents a set of unlabeled nodes (i.e., topic persons). Initially, $P_{unlabeled} = P$; that is, all nodes are unlabeled. The algorithm randomly selects K nodes as the seeds of stance communities and expands the communities iteratively by merging unlabeled nodes. In each iteration, a set of unlabeled nodes U that connect directly to a stance community are identified (i.e., $U = \{p_i \in P_{unlabeled} \mid (p_i,p_j) \in E, p_j \in c_k, 1 \le k \le K\}$). Each

دانلود کننده مقالات علمی freepaper.me FREE paper

freepaper.me FREE paper دانلود کننده مقالات علمی

The Stance Community Expansion Algorithm:

$P_{unlabeled} = P$

randomly select K nodes from $P_{unlabeled}$ to form the seeds of $\{c_1, c_2, \ldots, c_K\}$

havePositiveMergingScore = true

// for all unlabeled nodes, discover their communities

while ($P_{unlabeled} \neq \phi$ & *havePositiveMergingScore*) **do**

havePositiveMergingScore = false

// the set U contains the unlabeled node that has connection to the labeled node

$U = \{ p_i \in P_{unlabeled} \mid (p_i,p_j) \in E, p_j \in c_k, 1 \le k \le K \}$

for each p_i in U **do**

// the set Z_i contains the stance community that the unlabeled node p_i directly connected to

$Z_i = \{c_k \mid (p_i,p_j) \in E, p_j \in c_k, 1 \le k \le K\}$

$score_{max} = \max_{c_k \in Z_i} ms_{i,k}$

$community_{max} = \arg\max_{c_k \in Z_i} ms_{i,k}$

// the unlabeled node p_i will belong to the community with the maximum merging score

if $score_{max} > 0$ **then**

$c_{community_{max}} = c_{community_{max}} \cup \{p_i\}$

$P_{unlabeled} = P_{unlabeled} \setminus \{p_i\}$

havePositiveMergingScore = true

end if

end for

end while

return $C = \{c_1, c_2, \ldots, c_K\}$

Fig. 3. The stance community expansion algorithm.

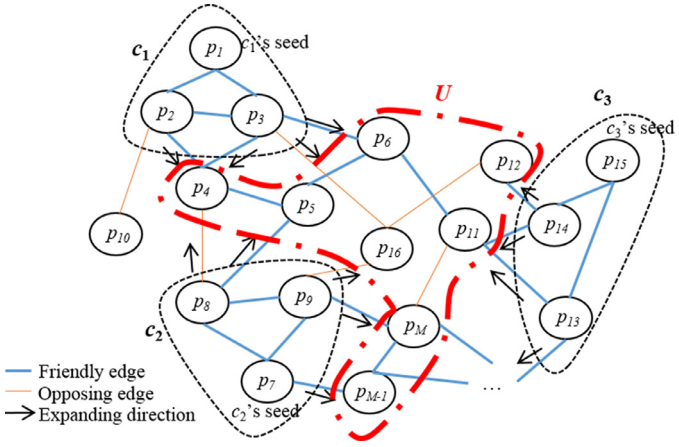

Fig. 4. An example of stance community expansion.

node p_i in U is then examined to determine an appropriate community label for it. Let Z_i denote the set of stance communities that the unlabeled node p_i is connected to directly; that is, $Z_i = \{c_k \mid (p_i,p_j) \in E, p_j \in c_k, 1 \le k \le K\}$. For instance, Z_4 shown in Fig. 4 comprises communities c_1 and c_2. We compute the merging score for each of the stance communities c_k in Z_i as follows:

$$ms_{i,k} = \sum_{p_j \in c_k, (p_i,p_j) \in E} \delta(p_i, p_j), \tag{4}$$

where $ms_{i,k}$ is the score of merging p_i with c_k. Basically, the merging score is the sum of the edge weights associated with p_i and stance community c_k. Intuitively, merging p_i with a community that has a positive merging score should produce a stance-coherent community. When more than one community has a positive merging score, the algorithm merges p_i with the stance community that has the maximum merging score. Below, we show that the step provides the most benefit for the objective function. Note that the merging score is negative if most of the nodes in c_k have an opposing friendship to p_i. Because merging p_i with a stance-opposing community is inappropriate, the algorithm revokes the merge operation if the maximum merging score is negative. The algorithm iteratively expands stance communities until all the unlabeled nodes in the friendship network are merged or no unlabeled node has a positive merging score with any stance community. Then, it returns a stance community identification result which will be polished by the stance community refinement algorithm.

The following cases show how the merge step of the algorithm benefits the objective function. In the first case, $|Z_i| = 1$ and the merging score of the connected stance community is positive[2]. Here, p_i is merged with the connected stance community. Because there is no other connected stance community, the merge operation will not change the second term of the objective function. Moreover, the operation increases the first term of the objective function by the positive merging score, so it benefits the objective function. In the second case, $|Z_i| > 1$ and the maximum merging score is positive[3]. Next, we show that merging p_i with the stance community that has the maximum merging score provides the most benefit for the objective function.

Proof. Let $|Z_i| = k$, and let $k > 1$. We have a sequence of merging scores $\langle ms_{i,1}, ms_{i,2}, \ldots, ms_{i,k} \rangle$ for the stance communities in Z_i. Let $ms_{i,1} \ge ms_{i,2} \ge \ldots \ge ms_{i,k}$ and let $ms_{i,1} > 0$. The stance community expansion algorithm merges p_i with c_1. The inequality $ms_{i,1} \ge ms_{i,n}$ holds for any stance community c_n in Z_i if $n \neq 1$. In other words,

$$\sum_{p_j \in c_1} \delta(p_i, p_j) \ge \sum_{p_j \in c_n} \delta(p_i, p_j). \tag{5}$$

Because Z_i has been determined, the summation of $\langle ms_{i,1}, ms_{i,2}, \ldots, ms_{i,k} \rangle$ (i.e., $\Sigma_{l = 1 \text{ to } k}\ ms_{i,l}$) is a fixed value. The inequality $ms_{i,1} \ge ms_{i,n}$ also implies that

$$\sum_{l \neq 1} ms_{i,l} \le \sum_{l \neq n} ms_{i,l} \tag{6}$$

[2] We exclude the case where $|Z_i| = 1$ and the merging score is negative. This is because the algorithm will not merge p_i with any stance community.

[3] We exclude the case where the maximum merging score is negative because the algorithm will not merge p_i with any stance community.

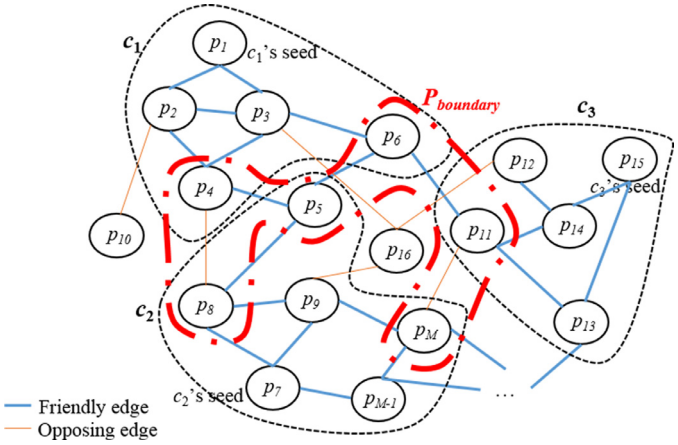

Fig. 5. An example of stance community refinement.

That is,

$$\sum_{l=2\ to\ k}\sum_{p_j\in c_l}\delta(p_i,p_j)\le\sum_{l=1\ to\ k,l\neq n}\sum_{p_j\in c_l}\delta(p_i,p_j) \tag{7}$$

or

$$-\sum_{l=2\ to\ k}\sum_{p_j\in c_l}\delta(p_i,p_j)\ge-\sum_{l=1\ to\ k,l\neq n}\sum_{p_j\in c_l}\delta(p_i,p_j). \tag{8}$$

By combining Eqs. (5) and (8), we have

$$\sum_{p_j\in c_1}\delta(p_i,p_j)-\sum_{l=2\ to\ k}\sum_{p_j\in c_l}\delta(p_i,p_j)\ge\sum_{p_j\in c_n}\delta(p_i,p_j)-\sum_{l=1\ to\ k,l\neq n}\sum_{p_j\in c_l}\delta(p_i,p_j). \tag{9}$$

□

The above inequality indicates that if the unlabeled node p_i is associated with more than one stance community, the stance community expansion algorithm will merge p_i with the community that benefits the objective function the most.

3.4. Stance community refinement

The stance community expansion algorithm iteratively expands stance communities from the seed nodes. In some cases, a node is merged with a stance community simply because it is close to the community's seed. However, it may be better to merge the node with some other community. For instance, node p_5 in Fig. 5 is merged with community c_2 even though it is strongly associated with community c_1. Also, the expansion result depends on the quality of the seeds. To minimize the effect of the above "*early merging*" problem and to lessen the influence of the seed initialization, we developed the following stance community refinement algorithm. The algorithm refines the communities iteratively. In each iteration, it identifies a set of boundary nodes $P_{boundary}\subseteq P$. Each node in $P_{boundary}$ belongs to a stance community and also connects to some other stance communities. In other words, $P_{boundary}=\{p_i \mid (p_i,p_j)\in E,\ p_i\in c_m,\ p_j\in c_n,\ m\neq n\}$. The algorithm re-clusters each boundary node to the stance community that produces the maximum merging score. It continues to identify and cluster boundary nodes until $P_{boundary}$ is empty or the identification result is stable; that is, no boundary node re-clustering benefits the objective function value and the value of the objective function converges to a local optimum.

Basically, our stance community refinement is a hill-climbing algorithm in that it iteratively improves the stance community identification result. However, to guarantee that a hill-climbing algorithm reaches a local optimum, we need to prove that each iteration of the algorithm monotonically increases (decreases) the objective function value [25,33,44]. Below, we prove that the value of the objective function increases monotonically in each boundary node re-clustering operation.

Proof. Let p_i be a boundary node. As a boundary node belongs to a stance community and also connects to some other stance communities, $|Z_i|$ must be greater than 1. That is, $|Z_i|=k>1$. Let $\langle ms_{i,1}, ms_{i,2}, \ldots, ms_{i,k}\rangle$ be the merging scores of the stance communities in Z_i, and let $ms_{i,1}\ge ms_{i,2}\ge\ldots\ge ms_{i,k}$. In addition, let $c_n\in Z_i$ be the stance community that p_i currently belongs to. The inequality $ms_{i,1}\ge ms_{i,n}$ holds. In other words,

$$\sum_{p_j\in c_1}\delta(p_i,p_j)\ge\sum_{p_j\in c_n}\delta(p_i,p_j). \tag{10}$$

Because Z_i has been determined, the summation of $\langle ms_{i,1}, ms_{i,2}, \ldots, ms_{i,k}\rangle$ (i.e., $\Sigma_{l=1\ to\ k}\ ms_{i,l}$) is a fixed value. The inequality $ms_{i,1}\ge ms_{i,n}$ also implies that

$$\sum_{l\neq 1}ms_{i,l}\le\sum_{l\neq n}ms_{i,l} \tag{11}$$

That is,

$$\sum_{l=2\ to\ k}\sum_{p_j\in c_l}\delta(p_i,p_j)\le\sum_{l=1\ to\ k,l\neq n}\sum_{p_j\in c_l}\delta(p_i,p_j) \tag{12}$$

or

$$-\sum_{l=2\ to\ k}\sum_{p_j\in c_l}\delta(p_i,p_j)\ge-\sum_{l=1\ to\ k,l\neq n}\sum_{p_j\in c_l}\delta(p_i,p_j). \tag{13}$$

By combining Eqs. (10) and (13), we have

$$\sum_{p_j\in c_1}\delta(p_i,p_j)-\sum_{l=2\ to\ k}\sum_{p_j\in c_l}\delta(p_i,p_j)\ge\sum_{p_j\in c_n}\delta(p_i,p_j)-\sum_{l=1\ to\ k,l\neq n}\sum_{p_j\in c_l}\delta(p_i,p_j) \tag{14}$$

Similar to the proof of stance community expansion, the above inequality indicates that the stance community refinement always re-clusters p_i into the community that benefits the objective function the most. The inequality also implies that

$$\sum_{p_j\in c_1}\delta(p_i,p_j)-\sum_{l=2\ to\ k}\sum_{p_j\in c_l}\delta(p_i,p_j)-[\sum_{p_j\in c_n}\delta(p_i,p_j)-\sum_{l=1\ to\ k,l\neq n}\sum_{p_j\in c_l}\delta(p_i,p_j)]\ge 0 \tag{15}$$

The left-hand side of the inequality is equivalent to the variation in the objective function when p_i is re-clustered. Note that the variation is always non-negative. In other words, re-clustering the boundary nodes in $P_{boundary}$ increases the value of the objective function monotonically. Because the set of possible stance community identification results is finite, the stance community refinement algorithm will eventually find a local optimum [25,33,44]. □

3.5. Stance-irrelevant topic person detection

A person mentioned frequently in topic documents may be irrelevant to the topic stances. For instance, in the topic about the 2012 French Presidential Election, U.S. President Barack Obama, one of the most influential people in the world, was frequently mentioned in the topic documents because journalists liked to analyze his attitude toward the candidates. However, President Obama showed no preference to any camp. SCIFNET can detect stance-irrelevant topic persons, which are defined as follows.

Definition 8. Stance-irrelevant Topic Persons:

freepaper.me FREE paper

The Stance Community Refinement Algorithm:

input: $C = \{c_1, c_2, \ldots, c_K\}$

$C_{old} = \{\}$

while $(C \neq C_{old})$ **do**

$C_{old} = C$

$P_{boundary} = \{p_i | (p_i, p_j) \in E,\ p_i \in c_m,\ p_j \in c_n,\ m \neq n\}$

if $P_{boundary} = \phi$ **then**

break

end if

// For each node which connects to more than one community, re-assign it to the community with the // largest merging score

for each p_i in $P_{boundary}$ **do**

$c_{original}$ = the community that p_i belongs to

$Z_i = \{c_k | (p_i, p_j) \in E,\ p_j \in c_k,\ 1 \leq k \leq K\}$

$score_{max} = \max_{c_k \in Z_i} ms_{i,k}$

$community_{max} = \arg\max_{c_k \in Z_i} ms_{i,k}$

if $c_{community_{max}} \neq c_{original}$ **then**

$c_{community_{max}} = c_{community_{max}} \cup \{p_i\}$

$c_{original} = c_{original} \backslash \{p_i\}$

$P_{boundary} = P_{boundary} \backslash \{p_i\}$

end if

end for

end while

return $C = \{c_1, c_2, \ldots, c_K\}$

Fig. 6. The stance community refinement algorithm.

Stance-irrelevant topic persons form a set $P_{irrelevant} = \{p_i \in P | p_i \notin c_k,\ 1 \leq k \leq K\}$.

In other words, a topic person is stance-irrelevant if he/she does not belong to any stance community. SCIFNET classifies two types of nodes as stance-irrelevant because they cannot be merged with a stance community. The first is the set of outliers which have no connections to other nodes in a network [53]. The nodes are stance-irrelevant because they do not show connections with any stance community. The second type comprises nodes that have connections with stance communities; however, most of the connections are with communities that have opposing associations with the nodes. Because the merging scores of the connected communities are negative, the nodes cannot merge with any stance community.

Technically, we can increase the value of the objective function by merging a node that belongs to the second type with a community that does not have any connections with the node. For instance, merging node p_{10} in Fig. 7 with c_2 increases the value of objective function by 1.5. Even if the node connects to every stance community, the value of the objective function can still be increased by merging the node with the community that has the minimum negative merging score. For example, merging node p_{16} in Fig. 7 with c_1 increases the objective function value by 2.1. The above strategies increase the value of the objective function because they reduce the friendship strength between stance communities, i.e., the second term of the objective function. However, although the two strategies are mathematically correct, merging a node with a community that does not have any connections or with the community that has the minimum negative merging score is irrational. Hence, in this study, we do not merge the second type of nodes.

In a future work, we will incorporate other information to handle the second type of nodes and refine the detection of stance-irrelevant topic persons.

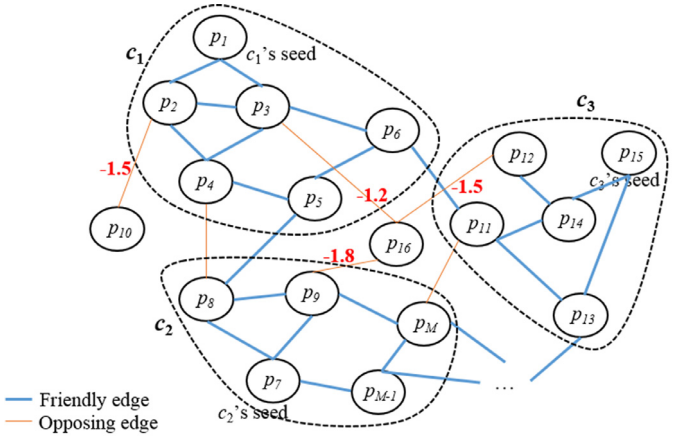

Fig. 7. An example of the associations of stance-irrelevant persons.

3.6. The computational complexity

In this section, we analyze the time complexity of friendship network construction, stance community expansion, stance community refinement, and stance-irrelevant topic person detection, which are the major components of SCIFNET. We also present the total time complexity of SCIFNET. The friendship network construction examines every person pair to measure their stance-oriented correlation coefficients and co-neighboring levels, whose time complexities are $O(N)$ and $O(M)$, respectively. As there are M^2 person pairs in a given topic, the overall cost of the friendship network construction is $O(M^3 + NM^2)$. Generally, the number of topic documents (i.e., N) is relatively larger than that of topic persons (i.e., M). Hence, the complexity of the construction process is $O(NM^2)$. The stance community expansion is based on the

The Modified Stance Community Expansion Algorithm for Detecting Overlapping Stance Communities

$P_{unlabeled} = P$
randomly select K nodes from $P_{unlabeled}$ to form the seeds of $\{c_1, c_2, \ldots, c_K\}$
$havePositiveMergingScore = true$
    // for all unlabeled nodes, discover their communities
    while ($P_{unlabeled} \neq \phi$ & $havePositiveMergingScore$) **do**
    $havePositiveMergingScore = false$
        //the set U contains the unlabeled node that has connection to the labeled node
        $U = \{ p_i \in P_{unlabeled} \mid (p_i, p_j) \in E, p_j \in c_k, 1 \leq k \leq K \}$
        for each p_i in U **do**
            // the set Z_i contains the stance community that the unlabeled node p_i directly connected to
            $Z_i = \{c_k \mid (p_i, p_j) \in E, p_j \in c_k, 1 \leq k \leq K\}$
            $pms_i = \{c_k \mid c_k \in Z_i \text{ and } ms_{i,k} > 0\}$
            $score_i = \sum_{c_k \in pms_i} ms_{i,k}$
            $community_{merge} = \emptyset$
            for each c_k in pms_i **do**
                if ($ms_{i,k} / score_i > mergedThreshold$) **then**
                    $community_{merge} = community_{merge} \cup c_k$
                end if
            end for
            if ($|community_{merge}| > 0$) **then**
                $havePositiveMergingScore = true$
            end if
            merged = false
            for each c_k in $community_{merge}$ **do**
                $c_k = c_k \cup \{p_i\}$
                merged = true
            end for
            if (merged) **then**
                $P_{unlabeled} = P_{unlabeded} \backslash \{p_i\}$
            end if
        end for
    end while
return $C = \{c_1, c_2, \ldots, c_K\}$

Fig. 8. The modified stance community expansion algorithm for detecting overlapping stance communities.

proposed merging score which computes the association of a node to a stance community. The complexity of the merging score calculation is O(M) as it needs to examine every node of the connected stance communities. Therefore, the overall complexity of the process is O(M^2). The stance community refinement is also based on the merging score, and iteratively refines each node's stance until the refinement converges to a local optimum. Letting T be the iteration number, the cost of the stance community refinement is O(TM^2). As mentioned in Section 3.5, the stance-irrelevant topic persons are the nodes that have no connection to other nodes in a friendship network. To detect those nodes, we have to examine all nodes in the constructed friendship network. The complexity of the operation is therefore $O(M)$. In sum, as mentioned earlier, N dominates the value of T and M, so the total time complexity of SCIFNET is O(NM^2).

3.7. SCIFNET for overlapping stance communities

In many topics, it is possible that a person belongs to more than one stance community. Hence, we present a variant of SCIFNET for detecting overlapping stance communities. Figs. 8 and 9 respectively show the modified stance community expansion and refinement algorithms. The main difference to the algorithms mentioned in the previous sections is in the way they merge a node. Rather than assigning a node to the stance community that has a maximum merging score, the modified algorithms merge the node with all the communities whose merging scores are above a predefined threshold. Specifically, let variable pms_i be the set of stance communities that have a positive merging score with node p_i. The sum of the positive merging scores $score_i$ is used to normalize the merging scores. For each stance community in pms_i, if the normalized merging score is larger than the merging threshold, we assign p_i to the stance community. In this way, p_i can belong to multiple stance communities.

3.8. The cardinality of stance communities

Selecting the appropriate number of communities (clusters) is a difficult and on-going research issue [33]. In practice, the value of K can be determined by experts who are familiar with the investigated topic. However, if the number of stance communities cannot be manually assigned, the following strategy is presented to determine the cardinality of stance communities. The cardinality strategy initiates the number of stance communities (i.e., K) with 2.

freepaper.me FREE paper

The Modified Stance Community Refinement Algorithm for Detecting Overlapping Stance Communities

input: $C = \{c_1, c_2, \ldots, c_K\}$

$C_{old} = \{\}$

while $(C \neq C_{old})$ **do**

$C_{old} = C$

$P_{boundary} = \{p_i | (p_i, p_j) \in E,\ p_i \in c_m,\ p_j \in c_n,\ m \neq n\}$

if $P_{boundary} = \phi$ **then**

break

end if

// For each node which connects to more than one community, re-assign it to the community with // the largest merging score

for each p_i in $P_{boundary}$ **do**

$Z_i = \{c_k | (p_i, p_j) \in E, p_j \in c_k,\ 1 \leq k \leq K\}$

$pms_i = \{c_k \mid c_k \in Z_i \text{ and } ms_{i,k} > 0\}$

$score_i = \sum_{c_k \in pms_i} ms_{i,k}$

$community_{merge} = \emptyset$

for each c_k in pms_i **do**

if $(ms_{i,k} / score_i > mergedThreshold)$ **then**

$community_{merge} = community_{merge} \cup c_k$

end if

end for

if $(|community_{merge}| > 0)$ **then**

$havePositiveMergingScore = true$

end if

merged = false

for each c_k in $community_{merge}$ **do**

$c_k = c_k \cup \{p_i\}$

merged = true

end for

if (merged) **then**

$P_{unlabeled} = P_{unlabeded} \setminus \{p_i\}$

end if

end for

end while

return $C = \{c_1, c_2, \ldots, c_K\}$

Fig. 9. The modified stance community refinement algorithm for detecting overlapping stance communities.

Next, it iteratively executes the stance community expansion and refinement algorithms such that each iteration increases the community number K by 1. The iteration stops when the remaining nodes (i.e., the nodes belonging to no community) have no positive merging score or when they are isolated nodes. Then, a stance identification result is returned and the corresponding K denotes the number of stance communities.

4. Experiment

In this section, we introduce the data corpus used in the experiments; demonstrate the effectiveness of each system component; and compare our method's performance with those of other well-known community detection methods and clustering algorithms. Then, we present a stance community identification result and discuss the stance-irrelevant persons detected by our method.

4.1. Dataset

Stance community identification is a relatively new research area. To the best of our knowledge, there is no official corpus for the subject; hence, we compiled a data corpus[4] for evaluations. The corpus comprises 30 topics and 4996 topic documents, all downloaded from the Google News. The collected topics cover three domains, namely sport, business issues, and political elections; and each topic involves about four competing stance communities. We also asked human experts to manually filter out irrelevant documents to ensure that the experiment documents are on-topic. To extract important topic persons mentioned in the topic documents, we used the well-known Stanford Named Entity Recognizer[5], which tags the person names in an input text. The recognizer extracted 6648 unique person names for all the topics. We found that a large number of the person names rarely appeared in the topic documents; and the frequency distribution followed Zipf's law [56]. In other words, there were very few frequent person names. Moreover, as there is no perfect named entity recognizer, several of the infrequent person names were incorrect or ambiguous (e.g., a string intermixed with the name of an

[4] Due to the length limitation, we established a web page at http://weal.im.ntu.edu.tw/SCIFNET.html which details titles, descriptions, number of documents, and stances of the evaluated topics.

[5] http://nlp.stanford.edu/software/CRF-NER.shtml

freepaper.me FREE paper

Table 1
The statistics of evaluated corpus.

# of topics	30		
# of topic documents	4996		
Avg. # of documents per topic	166.53		
Avg. # of stance communities per topic	3.97		
# of extracted topic persons	6648		
	$\lambda = 50\%$	$\lambda = 60\%$	$\lambda = 70\%$
# of evaluated topic persons	459	647	897
Avg. # of evaluated topic persons per topic	15.3	21.57	29.9

organization and the name of a person). To assess our method's performance accurately, for each of the evaluated topics, we removed the false person name entities and only evaluated the first frequent person names whose accumulated frequency reached $\lambda = 50$, 60, and 70 percent of the total frequency of all the extracted person names. The average number of evaluated person names under each setting of λ is shown in Table 1. All the names represent important topic persons.

We asked experts to group the evaluated topic persons into stance communities and establish a reliable ground truth for the performance evaluation. The kappa statistic which assesses the agreement between the experts is 74.73% and is good enough to conduct reliable evaluations. For the performance evaluation, we used the rand index [33], an important clustering evaluation metric, because the stance community identification method groups topic persons into clusters (i.e., communities). There are 1108,234 person pairs in the dataset. The rand index measures the percentage of all person pairs that are clustered correctly (i.e., if two persons with the same stance are placed in the same community or two persons with different stances are placed in different communities). The higher the score of the rand index, the better the stance community identification performance. In addition, we also represented the other information-theoretic metric named normalized mutual information (NMI) [33] which calculates the mutual information between the clusters and the classes, and was divided by the average of the entropy of the clusters and the classes. This metric can also reflect the quality of the clustering because it takes the quality of the clustering and the number of cluster into consideration [33]. Similar to the rand index, the higher the score of the NMI, the better the quality of the clustering. Because the stance community expansion algorithm depends on seed initialization, we randomly initialize our method twenty times. The rand index and NMI scores of all the evaluated topics over the initializations are averaged to obtain the overall stance community identification performance. For stance-irrelevant persons detected by the method, we measure their correctness in terms of the F1 score [33], which is the harmonic mean of the detection precision and the detection recall. The score is widely used to evaluate the overall effectiveness of a detection system.

4.2. System component analysis

4.2.1. Friendship orientation threshold

First, we consider the parameter θ, which is the threshold of friendship orientation used to establish the edges in a friendship network. In this experiment, θ is set between 0.1 and 0.9, and increased in increments of 0.1. Table 2 shows the lists of *Fwords* and *Owords* compiled by two linguistic experts. The stance word lists are used by the stance-oriented correlation coefficient (i.e., Eq. (2)) to compute the stance weight of a topic document. The parameter β, used by the friendship strength calculation (i.e., Definition 6), is set at 1. We discuss β and examine the effects of *Fwords* and *Owords* later. Figs. 10–15 show the rand index and the NMI scores under different settings of θ and λ. For each setting of θ, we

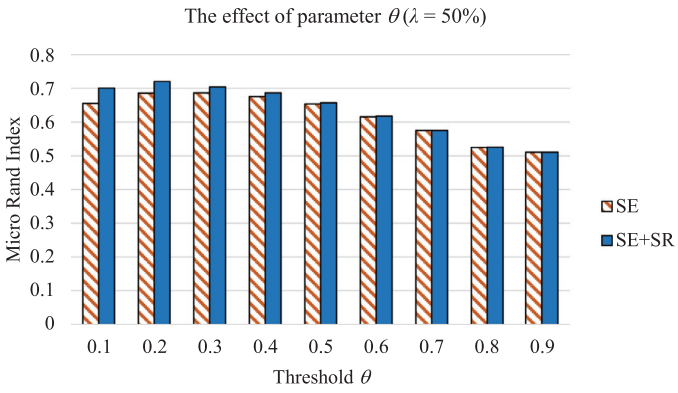

Fig. 10. The effect of parameter θ on rand index under $\lambda = 50\%$.

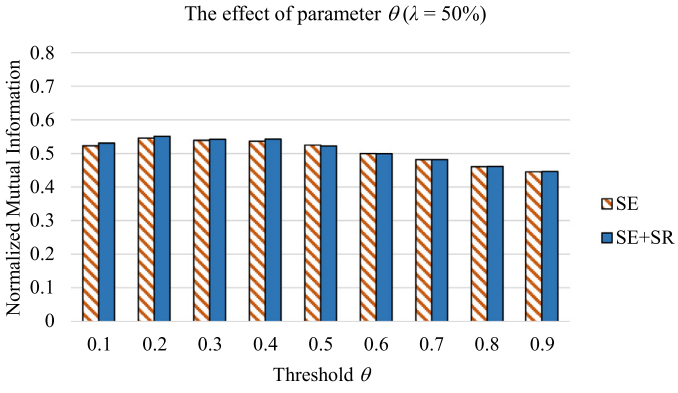

Fig. 11. The effect of parameter θ on NMI under $\lambda = 50\%$.

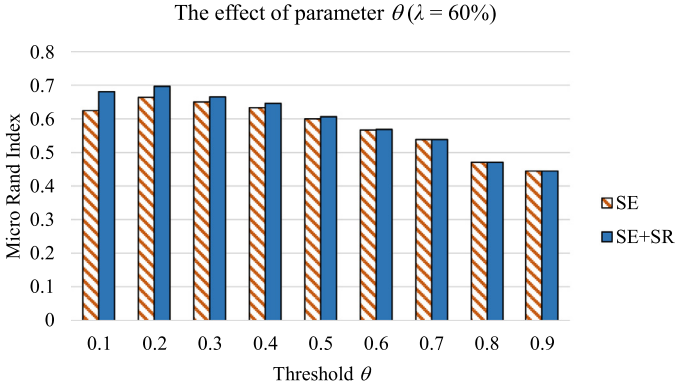

Fig. 12. The effect of parameter θ on rand index under $\lambda = 60\%$.

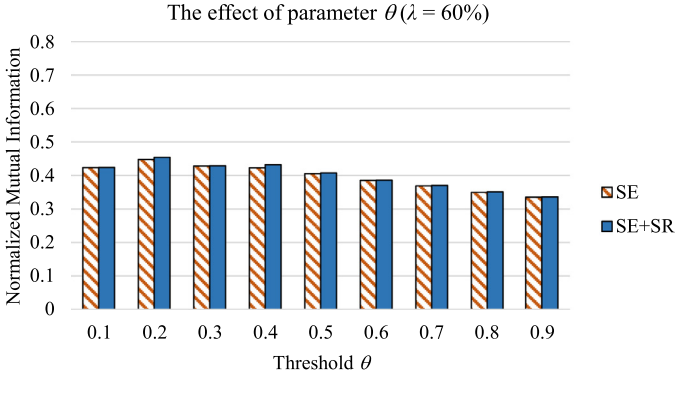

Fig. 13. The effect of parameter θ on NMI under $\lambda = 60\%$.

freepaper.me FREE paper

Table 2
The lists of *Fwords* and *Owords*.

Domain	Business issues	Political elections	Sports
Stance-friendly word - *Fwords*	support	cooperate	teammate
	member	support	like
	push	help	lead
	agreement	member	best
	help	good	good
	share	team	need
	approve	work	great
	benefit	partner	help
	partner	advocate	together
	consensus	friend	offend
Stance-opposing word - *Owords*	criticize	campaign	win
	rival	opposite	lose
	damage	rival	beat
	rape	fraud	defend
	fight	accusation	against
	campaign	contest	finish
	abuse	lost	end
	strike	beat	guard
	reject	debate	defense
	defend	defeat	hit

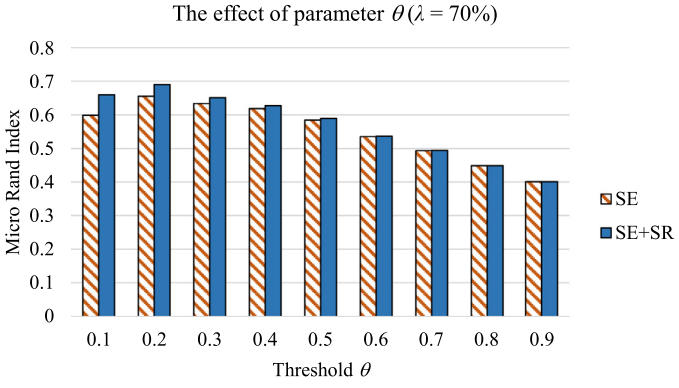

Fig. 14. The effect of parameter θ on rand index under $\lambda = 70\%$.

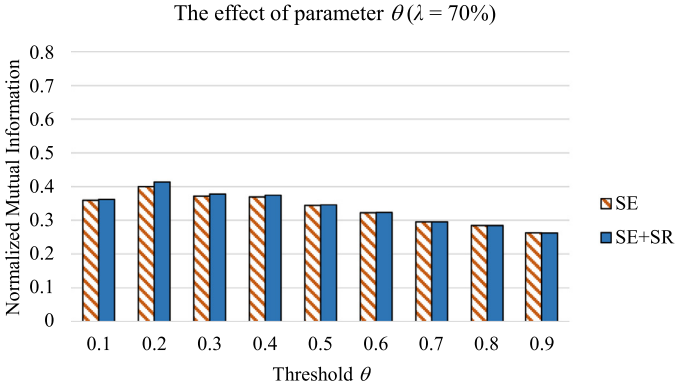

Fig. 15. The effect of parameter θ on NMI under $\lambda = 70\%$.

examine stance community expansion and stance community refinement techniques (denoted as SE+SR) in terms of the rand index. We also compare the performance based on stance community expansion only (denoted as SE), i.e., without stance community refinement.

As shown in the figures, the two metrics decreases as λ increases. A large λ implies that the stance community identification is difficult because the setting would include the infrequent topic persons in the stance community identification process. As

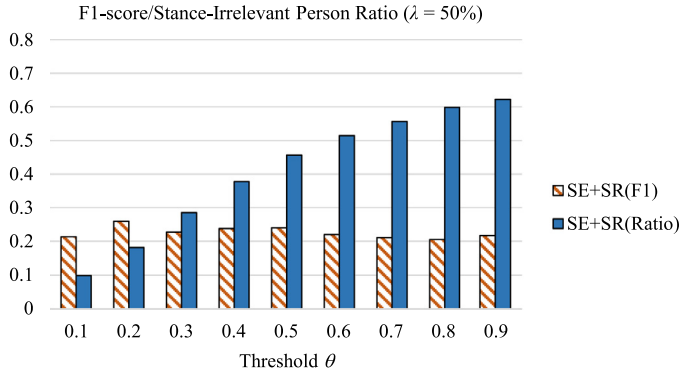

Fig. 16. The F1-score/ratio of the detected stance-irrelevant persons under $\lambda = 50\%$.

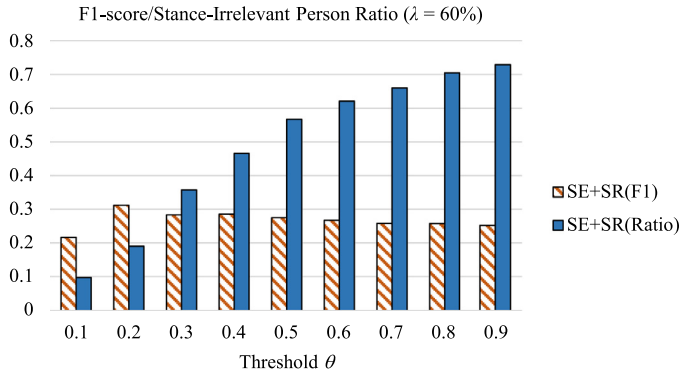

Fig. 17. The F1-score/ratio of the detected stance-irrelevant persons under $\lambda = 60\%$.

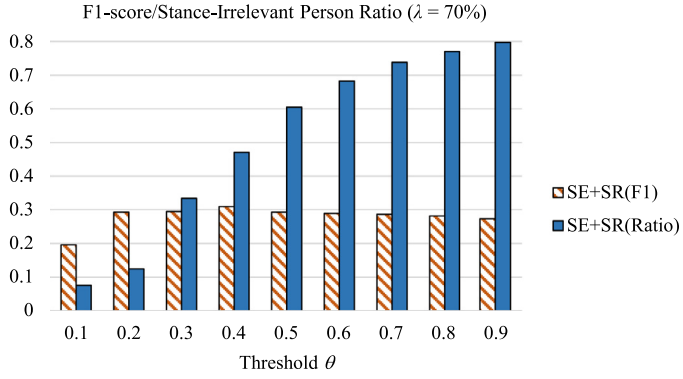

Fig. 18. The F1-score/ratio of the detected stance-irrelevant persons under $\lambda = 70\%$.

the construction of a friendship network is based on the occurrence of topic persons in the topic documents, including infrequent persons would reduce the quality of the network and therefore affect the stance community identification performance. Basically, the two metrics increases as the value of θ increases because a large θ filters out insignificant friendships between persons to improve the quality of the friendship network. When θ is greater than 0.4, the scores of the metrics drop gradually. Connections cannot be established between nodes when θ is large. As a result, the friendship network is too sparse to represent informative associations between persons and the stance community identification performance is inferior. It is noteworthy that SE+SR performs better than SE. The result demonstrates that stance community refinement resolves the "*early merging*" problem and the influence of the seed initialization in stance community expansion and therefore improves the stance community identification performance.

Figs. 16–18 show the F1 scores of stance-irrelevant topic person detection under different parameter settings. They also show

freepaper.me FREE paper

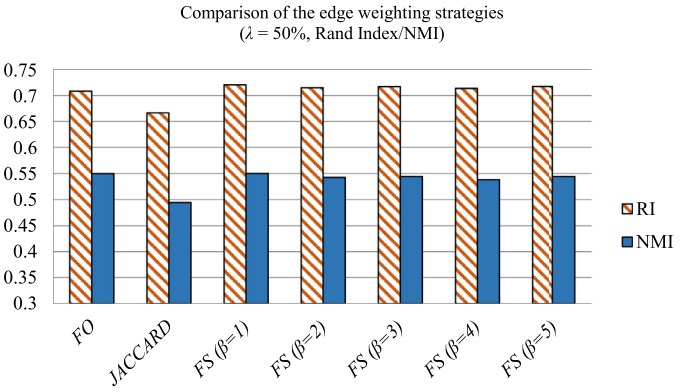

Fig. 19. Comparison of the edge weighting strategies under $\lambda = 50\%$.

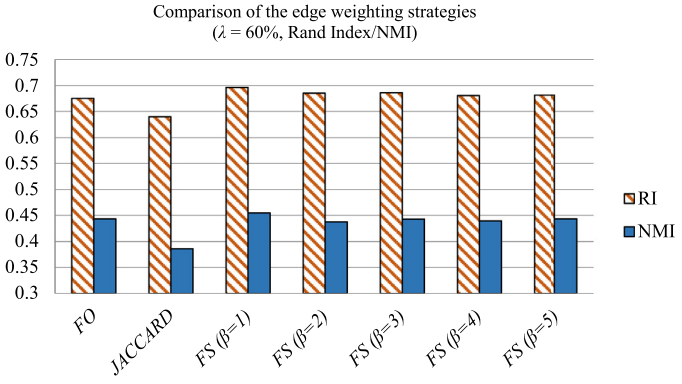

Fig. 20. Comparison of the edge weighting strategies under $\lambda = 60\%$.

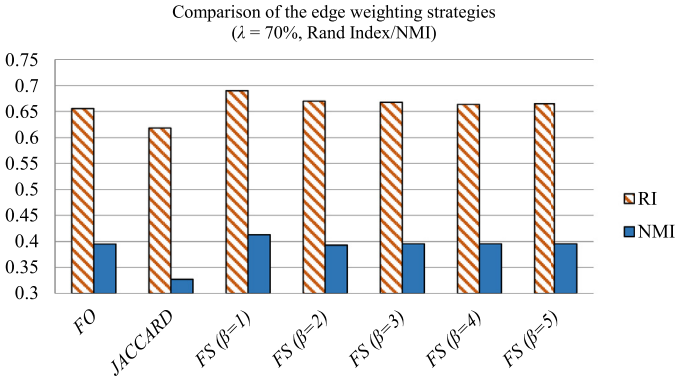

Fig. 21. Comparison of the edge weighting strategies under $\lambda = 70\%$.

the corresponding stance-irrelevant person ratio, which is the fraction of topic persons considered stance-irrelevant by our method. Note that the number of stance-irrelevant topic persons detected by SE+SR is the same as that detected by SE. This is because stance community refinement only re-clusters merged boundary nodes, so using it does not affect the stance-irrelevant topic person detection result. For ease of presentation, we only show SE+SR's F1 score and the stance-irrelevant topic person ratio. The F1 scores in the figures are inferior (around 0.2) because the number of stance-irrelevant topic persons in the evaluated topics is small. Hence, a misjudgment of the stance-irrelevant topic persons would reduce the F1 score significantly. The poor F1 scores also indicate that detecting stance-irrelevant topic persons is very difficult. Nevertheless, the scores are still superior to those of many of the community detection methods evaluated in the following experiments. As shown in the figures, a small θ value (e.g., $\theta = 0.1$) always produces a poor F1 score. The reason is that the friendship network constructed by a small θ contains many weak friendship edges that cause our method to merge a stance-irrelevant person with a stance community. Increasing the value of θ would improve the stance-irrelevant topic person detection performance, but setting it too high (i.e., higher than 0.5) would yield a sparse friendship network. Thus, many important topic persons are incorrectly classified as isolated nodes, which increase the stance-irrelevant topic person ratio. The corresponding F1 score is inferior because most of the detected stance-irrelevant persons are false alarms.

In summary, a large θ increases the ratio of stance-irrelevant topic persons and decreases the rand index and the NMI scores of stance community identification. Setting θ at 0.2 generally produces good scores for both metrics and F1 scores while maintaining a low stance-irrelevant person ratio. Therefore, we set θ at 0.2 in the following experiments.

4.2.2. Edge weight evaluation

Next, we discuss the friendship strength (FS), which combines the friendship orientation (FO) and the co-neighboring Jaccard coefficient (JACCARD) to compute the weight of a network edge. We evaluate the friendship strength by comparing it with its two constituents. In addition, we assess parameter β, which ensures that the friendship strength's exponent factor is not less than 1. As shown in Figs. 19–21, the metrics' scores under different settings of β are very similar. The results imply that the proposed friendship strength is insensitive to the setting of β. Nevertheless, setting β at 1 usually yields a superior performance, so we use the setting in the following experiments. Surprisingly, the identification performances based on the co-neighboring Jaccard coefficient are inferior. This is because the approach tends to underestimate the association of topic persons. For instance, if two persons do not have a common neighbor, the weight of the edge between them is zero even if they co-occur frequently in the topic documents. It is noteworthy that applying the two constituents together (i.e., the proposed friendship strength) achieves the best performance. As the constituents measure the association between nodes from different perspectives, applying them together identifies the friendship between topic persons accurately and therefore improves the system's performance. For example, in the sports topic "the 2011 NBA Conference Finals," if we simply employ the friendship orientation, the edge weight between Jason Terry and Shawn Marion, who are teammates of Dallas Maverick, would only be 0.280442. By combining the co-neighboring Jaccard coefficient with the friendship orientation, the edge weight increases to 1.448904. The improvement corresponds with the results reported by Jeh and Wisdom [27] and Antonellis et al. [3] who demonstrated that the association between nodes is proportional to their co-neighboring level.

4.2.3. Stance-oriented correlation coefficient evaluation

Next, we evaluate the stance-oriented correlation coefficient (i.e., SOCOR defined in Eq. (2)). The stance-oriented correlation coefficient enhances the traditional correlation coefficient (denoted as COR) by considering a document's stance weight, which is computed by using Turney and Littman's PMI method with the stance words listed in Table 2. Here, we compare our stance-oriented correlation coefficient with the traditional correlation coefficient. Turney and Littman also compiled a semantic orientation word list and used it to determine the semantic orientation of a text unit. To demonstrate the effect of our stance word list, we also compare the system's performance using the semantic orientation word list.

SOCOR outperforms COR, as shown in Figs. 22–24. The results demonstrate that the stance orientations of topic documents are informative for identifying the friendship orientation of topic persons. Notably, SOCOR with the semantic orientation word list is inferior. This is because the list is used to identify text units that

freepaper.me FREE paper

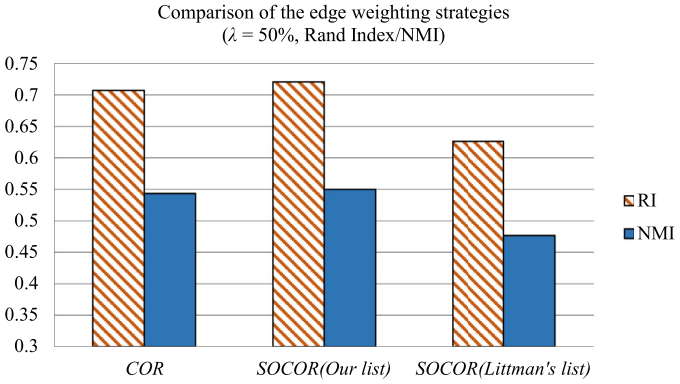

Fig. 22. Comparison of the correlation coefficient approaches under $\lambda = 50\%$.

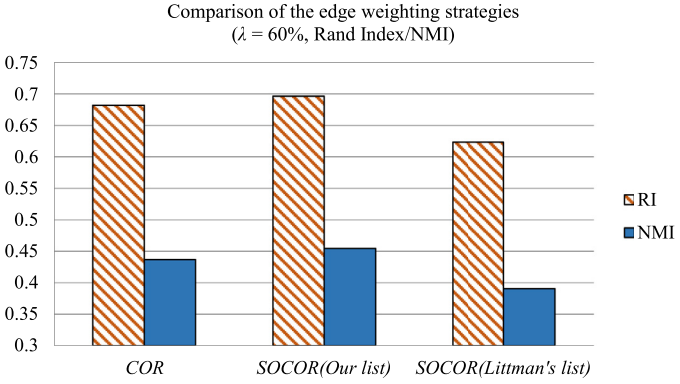

Fig. 23. Comparison of the correlation coefficient approaches under $\lambda = 60\%$.

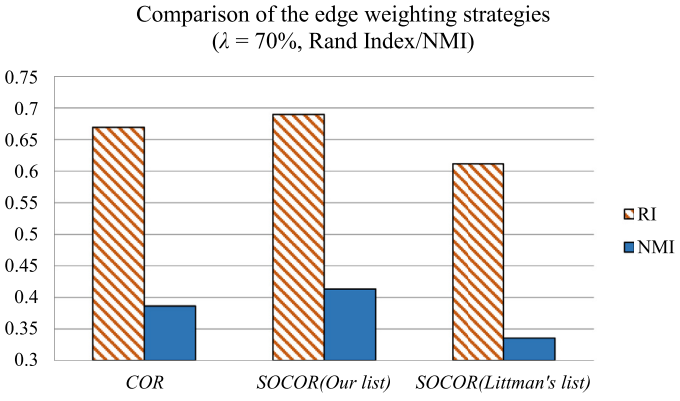

Fig. 24. Comparison of the correlation coefficient approaches under $\lambda = 70\%$.

convey positive or negative meanings, and the meanings may not reveal whether the associations between persons are friendly or opposing.

4.2.4. Overlapping stance community evaluation

In this section, we demonstrate the performance of our overlapping stance community identification. Figs. 25–27 show that the stance community identification performances are inferior when the merging threshold is high. This is because a high threshold will divide the members of the same stance community into different clusters, which deteriorates the system performance. Although relaxing the threshold generally improves system performance, a low threshold would result in noisy stance communities in which persons belonging to different communities would be merged. It is interesting to note that the performances presented here are inferior to those in the above experiments. This is because the topics we

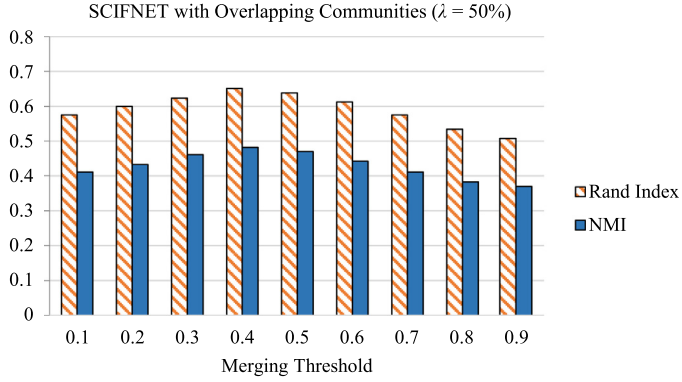

Fig. 25. SCIFNET with overlapping communities under $\lambda = 50\%$.

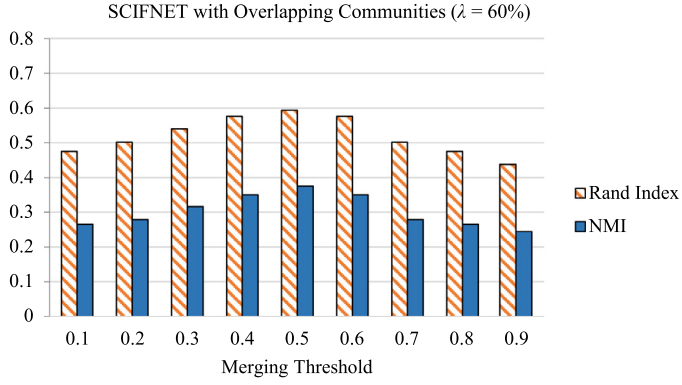

Fig. 26. SCIFNET with overlapping communities under $\lambda = 60\%$.

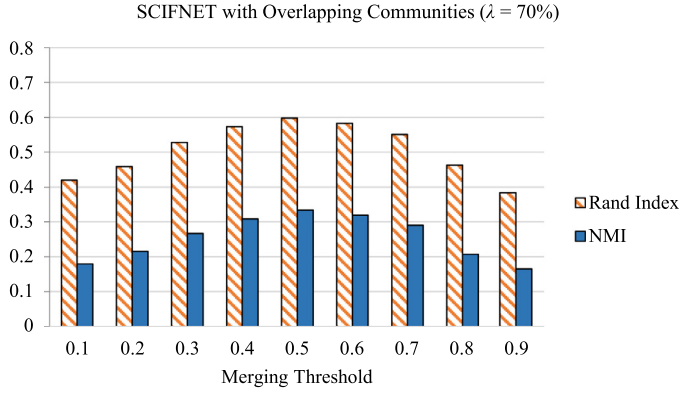

Fig. 27. SCIFNET with overlapping communities under $\lambda = 70\%$.

Table 3
The performance of the cardinality strategy.

λ	# of the returned stance communities	The rand index/NMI values of identified results
50%	2.467	0.6230/0.4286
60%	4.2	0.6702/0.4430
70%	7.2	0.5582/0.3316

evaluated do not have overlapping communities. The performances of the overlapping stance identification thus are poor when topic persons are assigned to more than one stance community.

4.2.5. Cardinality strategy evaluation

Finally, we evaluate the cardinality strategy presented in Section 3.8. Table 3 shows the performance of the cardinality strategy and the number of the returned stance communities. When

freepaper.me FREE paper

$\lambda = 50\%$, the number of the returned stance communities is less than the actual community number, i.e., 4. This is because the evaluated topic persons under $\lambda = 50\%$ are popular such that they tend to connect to each other in the constructed friendship network. The strategy thus incorrectly merges stance communities. When λ increases, less popular topic persons are included in the friendship network and the network becomes sparse. This sparsity in turn leads to a high cardinality that affects the stance identification performance. It is noteworthy that for under $\lambda = 60\%$, the strategy returns a good stance cardinality that the corresponding performances are comparable to those produced by fixing K at 4.

4.3. Comparison with other methods

4.3.1. Stance community identification evaluations

In this sub-section, we compare SCIFNET with 10 well-known community detection methods: FastModularity [37], SCAN [53], FEC [55], CODA [23], the signed modularity method (SM) [2], the seed-based intimate degree method (SID) [48], the seed-based absorbing degree method (SAD) [31], BASH [15], the maximal subgraph clustering coefficient method (MCC) [14], and the (α, β)-close method [20]. The methods require an input network. To ensure that the comparisons are fair, all the community detection methods run on the friendship networks generated by our method and partition each network into K communities. Note that the number of communities detected by SM sometimes is less than K. This is because the method detects communities according to the signs of the entries in the principal eigenvector. It stops community detection if the entry signs are all the same. Also note that SM and FEC are designed for signed networks. FastModularity, SCAN, CODA, SID, SAD, BASH, MCC, and (α, β)-close assume the analyzed networks are unsigned and examine the link structures to detect communities. Our friendship networks contain negative edges. To reduce the influence of negative edges on the methods, we also run the methods on the friendship networks without negative edges. We use the suffix "-neg" to indicate the result without negative edges. For instance, SCAN-neg stands for the result of SCAN on the friendship networks without negative edges. In SCAN, the clustering parameters ε and u are set at 0.5 and 2 respectively, as suggested by [53]; the link importance parameter of CODA is set at 0.2, as suggested by [23], and the parameter l of FEC is set at 10, as suggested by [55].

We also compare two popular clustering algorithms, namely, K-means [33] and HAC [34]. Both algorithms represent a topic person as an N-dimensional frequency vector in which an entry indicates the frequency that a topic person occurs in a topic document. To measure the association of topic persons, we utilize the cosine similarity [33] which is frequently used to determine the similarity of frequency vectors. For HAC, we consider four well-known cluster similarity strategies, namely, single-link, complete-link, average-link, and centroid-link strategies. In addition to the above methods, we compare a baseline method that clusters topic persons randomly. As the clustering results of CODA and K-means depend on their initializations, we randomize both methods twenty times and select the best, worst, and average results for comparison.

Table 4 shows the comparison results. All the compared methods perform better than the baseline, and our method achieves the best stance community identification performance. We observed that HAC and K-means tend to cluster popular topic persons together. This is because the cosine similarity is the inner product of two normalized frequency vectors [33], and it tends to yield a high similarity score if the calculated vectors contain many non-zero entries. As popular topic persons occur in many topic documents, the corresponding normalized frequency vectors contain a lot of non-zero entries. The clustering methods therefore overestimate the association of popular topic persons and group popular, but stance-different, persons together, which degrades the methods' performance. The inferior performance of HAC's single-link strategy is caused by the above defect because the strategy calculates the similarity of two clusters (i.e., communities) by examining the most similar person pair in the clusters. As a result, the strategy merges clusters containing popular persons even if the clusters represent different stances. In contrast, our method measures the association of topic persons in terms of the stance-oriented correlation coefficient and co-neighboring strength. Unlike the cosine similarity, the stance-oriented correlation coefficient considers how the occurrences of two topic persons vary jointly in a set of topic documents associated with stance orientations, thus being able to correctly measure the association of popular topic persons. For instance, in the political topic "the 2012 Korean presidential election," the friendship strength of Park Geun Hye and Park Jie-won, who represented different parties in the election, is -2.11474, but their cosine similarity is 0.984483. It is noteworthy that FastModularity, SCAN, and CODA perform better when the negative edges are removed from the friendship networks. As the methods are designed for unsigned networks, negative edges would distract their detection results. The FastModularity algorithm merges nodes into communities in terms of the modularity measure, which tends to merge communities that are connected by several edges. However, the measure ignores the edge weights of nodes. Many of the connected edges have small weights that impact the merged community's coherency and degrade the algorithm's performance. Our method merges communities in terms of the merging score (i.e., Eq. (4)). As the score is based on the edge weights (i.e., friendship strengths), the nodes in a community are highly associated. Consequently, the stance community identification result is better than that of the FastModularity algorithm. SCAN employs a Jaccard-like similarity to measure the co-neighboring strength between nodes and merges a node with a community if their co-neighboring strength is large. However, similar to FastModularity, SCAN ignores edge weights, which degrades its performance. In addition to the co-neighboring strength, our friendship strength considers the co-occurrence of nodes in topic documents associated with stance orientations. SCIFNET therefore outperforms SCAN significantly. The (α, β)-close method suffers the above problem too because its merge operation depends on the co-neighboring degree between a node and a community. As the merge process ignores edge weights, it may merge inappropriate nodes that deteriorate its stance identification result. SID employs a Jaccard-like function to merge a new node into an existing community. The Jaccard-like function is based on the number of common neighbors and it neglects the weights of the edges. Also, the method does not provide a refinement mechanism to avoid the early-merging problem of inappropriately merged nodes. These two defects degrade its community detection performance. SAD enhances the selection of community seeds by means of expert-defined rules. However, because the method still lacks a community refinement operation, its stance identification is sensitive to the seed initialization. BASH and MCC extract representative cliques from a network as the initial communities. We found from the experiments that popular topic persons of different stances normally have high degree and they tend to connect to each other. As a result of this, the fully connected cliques often group together the popular, but stance-different, persons that decrease the stance identification performance. While CODA integrates edge weights into its clustering objective function, the weights are based on the cosine similarity of the frequency vectors. Moreover, the objective function simply maximizes the sum of the edge weights in each community and ignores the association between the communities. For this reason, CODA groups together a lot of popular, but stance-different, topic persons. In addition to maximizing the association of nodes within communities, our objective function minimizes the association

freepaper.me FREE paper

Table 4
The rand index/NMI performance of the compared methods.

Method	$\lambda = 50\%$	$\lambda = 60\%$	$\lambda = 70\%$
SCIFNET (Best)	**0.7870/0.6654**	**0.7589/0.5721**	**0.7496/0.5271**
SCIFNET (Avg.)	0.7206/0.5502	0.6963/0.4545	0.6899/0.4131
SCIFNET (Worst)	0.6146***/0.4105^^^	0.5905***/0.3402^^^	0.6169***/0.3174^^^
FastModularity	0.6240***/0.4062^^^	0.6442***/0.3324^^^	0.6221***/0.2665^^^
FastModularity-neg	0.5934***/0.4103^^^	0.6200***/0.3300^^^	0.5971***/0.2659^^^
SCAN	0.6228***/0.4405^^	0.6523***/0.4022^	0.6798***/0.3653^^
SCAN-neg	0.6312***/0.3960^^^	0.6601***/0.3721^^	0.6869/0.3230^^^
CODA (Best)	0.7024***/0.5169	0.6887***/0.5049	0.6694***/0.4486
CODA (Avg.)	0.6606***/0.2990^^^	0.6487***/0.3052^^^	0.6378***/0.2756^^^
CODA (Worst)	0.6203***/0.0900^^^	0.6128***/0.0910^^^	0.6103***/0.0836^^^
CODA-neg (Best)	0.7240/0.6248	0.6977/0.5267	0.6667***/0.4551
CODA-neg (Avg.)	0.6592***/0.3369^^^	0.6499***/0.3065^^^	0.6374***/0.2706^^^
CODA-neg (Worst)	0.6009***/0.0575^^^	0.6124***/0.0877^^^	0.6102***/0.0767^^^
SM	0.6976***/0.5228	0.6919/0.4203	0.6881/0.3734^^
FEC	0.7125**/ 0.5420	0.6805***/ 0.4449	0.6469***/ 0.4055
SID-neg	0.6553***/ 0.5258	0.6061***/ 0.4047^	0.5223***/ 0.3429^^^
SAD-neg	0.5768***/0.4413^^	0.5622***/0.3947^	0.5211***/0.3274^^^
MCC-neg	0.5889***/0.4655^^	0.5522***/0.4146^	0.4863***/0.3339^^^
BASH-neg	0.5415***/ 0.3709^^^	0.5350***/ 0.3564^^^	0.5136***/ 0.3032^^^
(α, β)-close-neg	0.5688***/0.4364^^	0.5627***/4113^	0.5409***/0.2927^^^
HAC (Single-Link)	0.5968***/0.4865^	0.5323***/0.3628^^^	0.4545***/0.2935^^^
HAC (Complete-Link)	0.6916***/0.5386	0.6741***/0.4553	0.6136***/0.3691^^
HAC (Average-Link)	0.6979***/0.5175	0.6711***/0.4057^	0.6774***/0.3702^^
HAC (Centroid-Link)	0.6534***/0.5172	0.6148***/0.4202	0.5741***/0.3450^^^
K-means (Best)	0.7767/0.6623	0.7510/0.5654	0.7478/0.5145
K-means (Avg.)	0.7037***/0.5194	0.6881***/0.4283	0.6831***/0.3794^^
K-means (Worst)	0.5896***/0.3514^^^	0.5965***/0.2940^^^	0.5896***/0.2514^^^
Baseline (Avg.)	0.3983***/0.3487^^^	0.3479***/0.2547^^^	0.3085***/0.1933^^^

The results marked with *, **, and *** show, respectively, the improvements achieved by SCIFNET (Avg.) over the compared methods with 90%, 95% and 99% confidence levels based on the Z-statistic for two proportions, and the symbol ^, ^^, and ^^^ indicate the improvements based on the one-tailed paired t test [29].

between communities. Therefore, SCIFNET achieves a superior stance community identification performance. As shown in Table 4, FEC normally performed better than the other compared methods did; this is because the method is designed for signed networks and there are negative links in the produced friendship networks. Nevertheless, our method still outperformed FEC significantly. The SM method is also designed for signed networks. We found that SM sometimes cannot produce K communities for an evaluated topic because the signs of the entries in the principal eigenvectors are all positive. The method thereby groups persons with different stances together, but it is also based on the modularity which ignores the edge weights. Our method therefore outperforms the SM method.

Table 5
The F1 performance of stance-irrelevant topic person detection.

Method	$\lambda = 50\%$	$\lambda = 60\%$	$\lambda = 70\%$
SCIFNET (Best)	**0.358335**	**0.373005**	**0.363269**
SCIFNET (Avg.)	0.250637	0.292517	0.293951
SCIFNET (Worst)	0.037736	0.102941	0.178218
SCAN-neg	0.259259	0.287356	0.298182
CODA-neg (Best)	0.288889	0.325301	0.316667
CODA-neg (Avg.)	0.248889	0.247590*	0.247083***
CODA-neg (Worst)	0.177778*	0.168675**	0.183333***

The results marked with *, **, and *** show, respectively, the improvements achieved by SCIFNET (Avg.) over the compared methods with 90%, 95% and 99% confidence levels based on the Z-statistic for two proportions.

4.3.2. Stance-irrelevant topic person detection evaluation

One function of SCAN and CODA is to detect outliers (i.e., nodes that do not belong to any community). Here, we treat the outliers as stance-irrelevant topic persons and compare their stance-irrelevant topic person detection performance. Table 5 shows the comparison results. Note that CODA uses a clustering objective function to rank the nodes in a network and the last $\gamma\%$ nodes are denoted as outliers. To ensure a fair comparison, we adjusted $\gamma\%$ so that the number of stance-irrelevant topic persons detected by CODA is the same as that detected by our method.

As shown in Table 5, the F1 scores of the compared methods are all inferior because we select frequent topic persons for evaluation. All of them are important and influential in the evaluated topics, so very few of them are stance-irrelevant. Consequently, a misjudgment of the stance-irrelevant topic persons would reduce the F1 score dramatically. The inferior performance of the compared methods shows that the detection of stance-irrelevant topic persons is difficult and requires further investigation. Contrary to expectations, SCAN's F1 score is higher than our average F1 score. This is because of SCAN's high detection recall rate. As SCAN clusters nodes in terms of their co-neighboring strength, many weakly-connected nodes are treated as outliers. Consequently, its detection recall is high, which benefits its F1 performance. Nevertheless, our best F1 score is still the best stance-irrelevant topic person detection performance.

4.3.3. Discussion of seed initialization strategies

In this section, we evaluate SCIFNET incorporated with three well-known seed initialization strategies: Betweenness [24], Closeness [30], and Degree [11]. Given a network, the betweenness value of a node is the number of the shortest paths that the node is involved with. The betweenness strategy iteratively selects K nodes with the largest betweenness values as the seeds of communities. The closeness strategy sums the path lengths of a node to all other nodes in a network. The node with the minimum sum is regarded

freepaper.me FREE paper

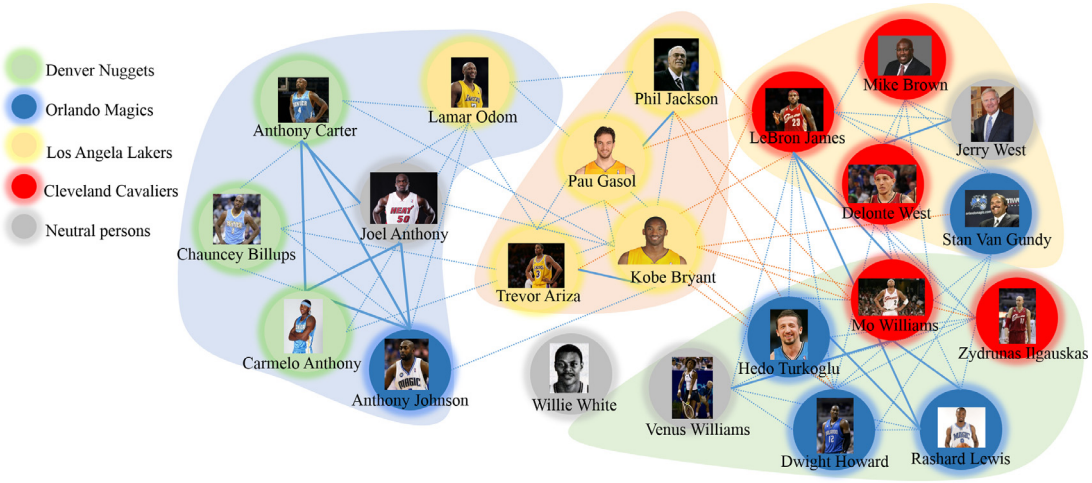

Fig. 28. The stance community identification result of the 2009 NBA Conference Final ($\lambda = 70\%$).

Table 6
The rand index/NMI performance of the seed initialization strategies.

Initialization strategy	$\lambda = 50\%$	$\lambda = 60\%$	$\lambda = 70\%$
Betweenness$_{\text{-refinement}}$	0.6872***/0.5128	0.6748***/0.4558	0.6719***/0.4224
Closeness$_{\text{-refinement}}$	0.6718***/0.5105	0.6331***/0.4235	0.5731***/0.3513^
Degree$_{\text{-refinement}}$	0.6053***/0.4689^	0.5805***/0.3983^	0.5110***/0.3429^^
Betweenness	0.7117/0.5212	0.6889/0.4578	0.7039/0.4322
Closeness	0.7020/0.5187	0.6719/0.4437	0.6556/0.3905
Degree	0.6905/0.5118	0.6692/0.4407	0.6689/0.4261

The results marked with *, **, and *** show, respectively, the improvements achieved by using stance community refinement with 90%, 95% and 99% confidence levels based on the Z-statistic for two proportions, and the symbol ^, ^^, and ^^^ indicate the improvements based on the one-tailed paired t test [29].

as the core of the network because it is close to all the other nodes. The strategy iteratively selects K nodes with the minimum sums as the seeds of communities. In the degree strategy, K nodes with the largest degrees in the network are selected as community seeds. As mentioned in Section 3.4, the proposed stance community refinement is capable of lessening the influence of inappropriate seed initializations. Here, in order to contrast the true effect of the seed initialization strategies, we evaluate the strategies with and without using the stance community refinement. Table 6 shows the performances of the initialization strategies and the suffix "-refinement" stands for the results without stance community refinement.

The comparison of results demonstrates that using stance community refinement considerably improves stance identification results. Furthermore, the performances differences between the initialization strategies become minor when the refinement is adopted. The results reveal that the proposed stance community refinement is able to reduce the influence of seed initializations. Notably, without stance community refinement, the performances of the strategies are inferior. This is because the strategies ignore the sign of edges that decrease the quality of the selected community seeds. For instance, we observed that the degree strategy tends to select popular nodes as the seeds of stance communities. A great portion of the edges connected to the selected seeds, however, have a negative weight that obscures the expanded stance communities. The result also reveals that discovering representative community seeds in a sign network is challenging and is worth investigation.

4.4. An example of stance community identification

The above experiments quantitatively evaluate the performance of SCIFNET. In this section, we consider a sports topic, namely the 2009 NBA Conference Finals, to assess our stance community identification result. The topic covers four basketball teams that competed for the title and we consider each team as a stance community. Fig. 28 shows the constructed friendship network. Stance-irrelevant topic persons are highlighted in gray; and teammates are highlighted in the same color. The blue edges and the orange edges depict friendly associations and opposing associations respectively. Their thickness indicates the friendship strength (i.e., edge weight). As shown in the figure, the friendship network accurately describes the associations of the topic persons. For instance, the orange edges always connect persons with different stances. While some stance-different persons are connected by blue edges, their friendship strength is very weak. It is noteworthy that many orange edges connect Los Angeles Lakers players and Orlando Magic players. This is because the two teams reached the finals. A large number of the topic documents report the teams' matchup and most of them contain stance-opposing words. As our method utilizes the stance weight of topic documents to measure the friendship strength of topic persons, the matchup-related documents help to capture the opposing orientations of the players.

The colored zones in the figure represent our stance community identification results. In this example, many topic persons are grouped into stance communities correctly. Moreover, one topic person (i.e., Willie White) is accurately classified as stance-irrelevant. Notably, our method prevents the teams' franchise players (i.e., Kobe Bryant, Carmelo Anthony, LeBron James, and Dwight Howard), who are also popular topic persons, from being merged. The outcome corresponds with the comparison result presented in the previous section, i.e., the proposed stance-oriented correlation coefficient is effective for measuring the friendship orientation of popular topic persons. We observed that incorrectly-grouped persons often appeared in a few topic documents. For instance, Cleveland player Zydrunas Ilgauskas, who only appeared in 12

freepaper.me FREE paper

topic documents, was grouped as a member of Orlando Magic. We analyzed the phenomenon and found that the stance-oriented correlation coefficient tends to overestimate the friendship of infrequent topic persons. This is because the coefficient is based on the occurrence pattern of topic persons. As infrequent persons are jointly absent from many topic documents, their friendships are overestimated. It is remarkable that Jerry West, an ex-Lakers player, is grouped as a member of Cavaliers. Jerry West was named "Mr. Clutch" because he made a lot of game-winning shots during his playing career. In Game 2 of the NBA conference finals, Cavaliers player LeBron James made an incredible game-winning shot. Many documents reported the event and tried to place him on a par with Jerry West. Their names thus co-occur frequently in the topic documents so they are grouped together. Interestingly, Venus Williams, a famous tennis player, is included in the experiment. During the matchup of Orlando Magic and Cleveland Cavaliers, Venus Williams was playing in the 2009 French Open. We observed that several topic documents collected from Google News were sports recaps that covered the NBA conference finals as well as the results of the tennis tournament. Consequently, Venus Williams was incorrectly classified as a member of Orlando Magic. The result suggests the analyzed topic documents need to be pure and on-topic. Diverse or noisy documents must be filtered out to enhance the result of stance community identification.

5. Concluding remarks

The Internet has become a crucial medium for disseminating and acquiring the latest information about topics. However, users are often overwhelmed by the enormous number of topic documents. Basically, times, places, and persons are the key elements of topics. Knowing the associations of topic persons can help readers construct the background knowledge of a topic and comprehend numerous topic documents quickly. In this paper, we defined the problem of stance community identification, which involves grouping important topic persons into stance-coherent communities. In addition, we presented a stance community identification method called SCIFNET that constructs a friendship network of topic persons from topic documents automatically. We developed the stance-oriented correlation coefficient to measure the friendship orientation of topic persons. The friendship orientation is then combined with the co-neighboring strength of the topic persons to measure their friendship strengths. Stance community expansion and stance community refinement techniques based on the designed objective function are used to identify stance communities of topic persons and identify stance-irrelevant topic persons. We also proved the techniques make the identified stance communities converge to a local optimum. The result of experiments on real-world topics demonstrate the effectiveness of SCIFNET and show that it outperforms many well-known community detection and clustering methods. Besides, the performance of our method is not sensitive to the parameter β and setting $\beta = 1$, and $\theta = 0.2$ generally achieves a superior stance community identification result.

The experiments suggest some interesting areas for future research. For instance, although the proposed stance-oriented correlation coefficient is effective in identifying the friendship orientation of popular topic persons, it is affected by the frequency sparseness problem of infrequent topic persons. Because infrequent topic persons are jointly absent from a lot of topic documents, the stance-oriented correlation coefficient may overestimate their friendship strength. Reducing the weight of documents when infrequent persons are jointly absent would resolve the overestimation problem. Moreover, considering off-topic documents may include irrelevant persons in the stance community identification process and degrade the system performance. Therefore, effective off-topic document elimination approaches should be developed to improve the stance community identification performance. Although we presented a simple strategy to select the appropriate cluster number, there is still room to improve the strategy. For instance, incorporating the number of stance communities into our objective function is one of our future research directions to automatically determine the appropriate value of K. In our experiment, the input topic documents were collected manually. To help Internet users comprehend emerging topics, we are integrating SCIFNET with techniques of topic detection and tracking [1], which automatically detect and track topic documents from different information sources (e.g., news agencies).

freepaper.me FREE paper

Acknowledgement

This research was supported in part by MOST 103-2221-E-002-106-MY2 from the Ministry of Science and Technology, Republic of China.

Reference

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, Y. Yang, Topic detection and tracking pilot study final report, in: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998, pp. 194–218.
[2] P. Anchuri, M. Magdon-Ismai, Communities and balance in signed networks: a spectral approach, in: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, 2012, pp. 235–242.
[3] I. Antonellis, H.G. Molina, C.C. Chang, Simrank++: query rewriting through link analysis of the click graph, in: Proceedings of the VLDB Endowment, 1, 2008, pp. 408–421.
[4] D. Barber, Bayesian Reasoning and Machine Learning, Cambridge University Press, New York, 2012.
[5] C.C. Chen, M.C. Chen, TSCAN: a novel method for topic summarization and content anatomy, in: Proceedings of the 31st annual international ACM SIGIR Conference on Research and Development in Information Retrieval, 2008, pp. 579–586.
[6] C.C. Chen, M.C Chen, TSCAN: a content anatomy approach to temporal topic summarization, IEEE Trans. Knowl. Data Eng. 24 (2012) 170–183.
[7] C.C. Chen, Z.-Y. Chen, C.-Y. Wu, An unsupervised approach for person name bipolarization using principal component analysis, IEEE Trans. Knowl. Data Eng. 24 (2012) 1963–1976.
[8] C.C. Chen, C.-Y. Wu, Bipolar person name identification of topic documents using principal component analysis, in: Proceeding of the 23rd International Conference on Computational Linguistics, 2010, pp. 170–178.
[9] J. Chen, O.R. Zaiane, R. Goebel, Detecting communities in social networks using max-min modularity, in: Proceedings of the SIAM International Conference on Data Mining, 2009, pp. 978–989.
[10] J. Chen, O.R. Zaiane, R. Goebel, Local Community Identification in Social Networks, in: Proceedings of the International Conference on Advances in Social Network Analysis and Mining, 2009, pp. 237–242.
[11] A. Chin, M. Chignell, A social hypertext model for finding community in blogs, in: Proceedings of the 7th Conference on Hypertext and Hypermedia, 2006, pp. 11–22.
[12] A. Clauset, M.E.J. Newman, C. Moore, Finding community structure in very large networks, Phys. Rev. E 70 (2004) 066111.
[13] Y. Cui, X. Wang, Uncovering overlapping community structures by the key bi–community and intimate degree in bipartite networks, Physica A 407 (2014) 7–14.
[14] Y. Cui, X. Wang, J. Li, Detecting overlapping communities in networks using the maximal sub-graph and the clustering coefficient, Physica A 405 (2014) 85–91.
[15] Y. Cui, X. Wang, J. Eustace, Detecting community structure via the maximal sub-graphs and belonging degrees in complex networks, Physica A 416 (2014) 198–207.
[16] C.H.Q. Ding, X. He, H. Zha, M. Gu, H.D. Simon, A min-max cut algorithm for graph partitioning and data clustering, in: Proceedings IEEE International Conference on Data Mining, 2001, pp. 107–114.
[17] Z. Ding, X. Zhang, D. Sun, B. Luo, Overlapping community detection based on network decomposition, Sci. Rep. 6 (2016) 24115.
[18] W.E. Donath, A.J. Hoffman, Lower bounds for the partitioning of graphs, IBM J. Res. Dev. 17 (1973) 420–425.
[19] J. Eustace, X. Wang, J. Li, Approximating web communities using sub-space decomposition, Knowl. Based Syst. 70 (2014) 118–127.
[20] J. Eustace, X. Wang, Y. Cui, Community detection using local neighborhood in complex networks, Physica A 436 (2015) 665–677.
[21] J. Eustace, X. Wang, Y. Cui, Overlapping community detection using neighborhood ratio matrix, Physica A 421 (2015) 510–521.
[22] A. Feng, J. Allan, Finding and linking incidents in news, in: Proceedings of the 16th Conference on Information and Knowledge Management, 2007, pp. 821–830.

[23] J. Gao, F. Liang, W. Fan, C. Wang, Y. Sun, J. Han, On community outliers and their efficient detection in information networks, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, pp. 813–822.
[24] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, in: Proceedings of the National Academy of Sciences, 99, 2002, pp. 7821–7826.
[25] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2006.
[26] Z. Harris, Distributional structure, Word 10 (1954) 146–162.
[27] G. Jeh, J. Widom, SimRank: a measure of structural-context similarity, in: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 538–543.
[28] H. Kanayama, T. Nasukawa, Fully automatic lexicon expansion for domain-oriented sentiment analysis, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2006, pp. 355–363.
[29] G. Keller, Statistics for Management and Economics, Cengage Learning, 2008.
[30] X. Liu, J. Bollen, M.L. Nelson, H.V. Sompel, Co-authorship networks in the digital library research community, Inf. Process. Manag. 41 (2005) 1462–1480.
[31] J. Li, X. Wang, J. Eustace, Detecting overlapping communities by seed community in weighted complex networks, Physica A 392 (2013) 6125–6134.
[32] J. Li, X. Wang, Y. Cui, Uncovering the overlapping community strcuture of complex networks by maximal cliques, Physica A 415 (2014) 398–406.
[33] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, New York, 2008.
[34] T. Mitchell, Machine Learning, McGraw-Hall, Maidenhead, 1997.
[35] R. Nallapati, A. Feng, F. Peng, J. Allan, Event threading within news topics, in: Proceedings of the 13th ACM International Conference on Information and Knowledge Management, 2004, pp. 446–453.
[36] M.E.J. Newman, Scientific collaboration networks: I. Network construction and fundamental results, Phys. Rev. E 64 (2001) 016131.
[37] M.E.J. Newman, Fast algorithm for detecting community structure in networks, Phys. Rev. E 69 (2004) 066133.
[38] M.E.J. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (2006) 036104.
[39] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2004) 026113.
[40] M. Oussalah, B. Escallier, D. Daher, An automated system for grammatical analysis of Twitter messages: A learning task application, Knowl. Based Syst. 101 (2016) 31–47.
[41] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, P. Spyridonos, Community detection in social media, Data Mining Knowl. Disc. 24 (2012) 515–554.
[42] G. Petrović, H. Fujita, SoNeR: social network ranker, Neurocomputing (2015).
[43] G. Ren, X. Wang, Epidemic spreading in time-varying community networks, Chaos 24 (2014) 023116.
[44] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, Prentice Hall, New Jersey, 2003.
[45] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 888–905.
[46] V.A. Traag, J. Bruggeman, Community detection in networks with positive and negative links, Phys. Rev. E 80 (2009) 036115.
[47] P.D. Turney, M.L. Littman, Measuring praise and criticism: inference of semantic orientation from association, ACM Trans. Inf. Syst. 21 (2003) 315–346.
[48] X. Wang, J. Li, Detecting communities by the core-vertex and initimate degree in complex networks, Physica A 392 (2013) 2555–2563.
[49] Z.-X. Wang, Z.-C. Li, X.-F. Ding, J.-H. Tang, Overlapping community detection based on node location analysis, Knowl. Based Syst. 105 (2016) 225–235.
[50] J.J. Whang, D.F. Gleich, I.,S. Dhillon, Overlapping community detection using neighborhood-Inflated seed expansion, IEEE Trans. Knowl. Data Eng. 28 (2016) 1272–1284.
[51] S. White, P. Smyth, A spectral clustering approach to finding communities in graphs, in: Proceedings of SIAM International Conference on Data Mining, 2005, pp. 76–84.
[52] F.-Y. Wu, The Potts model, Rev. Mod. Phys. 54 (1982) 235–268.
[53] X. Xu, N. Yuruk, Z. Feng, T.A.J. Schweiger, SCAN: a structural clustering algorithm for networks, in: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007, pp. 824–833.
[54] T. Yang, R. Jin, Y. Chi, S. Zhu, Combining link and content for community detection: a discriminative approach, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009, pp. 927–936.
[55] B. Yang, W.K. Cheung, J. Liu, Community mining from signed social networks, IEEE Trans. Knowl. Data Eng. 19 (2007) 1333–1348.
[56] G.K. Zipf, Human Behavior and the Principle of Least Effort, Addison-Wesley, Cambridge MA, 1949.

freepaper.me FREE paper