Building Character Graphs and Dividing Communities in Chinese Novels Based on Graph Data Extraction: Community Division for Character Emotional Polarity Networks

XIAOLEI ZHAO1, XUN LIANG1, (Senior Member, IEEE), FENG TANG1, XUAN ZHANG1, XIAOPING ZHOU1,2, (Member, IEEE), AND XINYU GUO3

1School of Information, Renmin University of China, Beijing 100872, China
2Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
3Department of Relational Database Service, Alibaba Corporation, Beijing 100123, China

Corresponding author: Xun Liang (xliang@ruc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 71531012 and Grant 71271211, in part by the Natural Science Foundation of Beijing under Grant 4172032, and in part by the National Social Foundation of China under Grant 18ZDA309.

ABSTRACT Chinese authors indirectly portray real social relationships in society by incorporating their own experiences and feelings about politics, economic life, and cultural habits. Social and natural knowledge are mixed and hidden in natural literary language. Therefore, by well defining the type of original text, the pipeline for processing the text and the rules to build the graph, we can extract enough valid and useful information. Based on higher-level information, we can build a literary character graph to promote computers’ comprehension of literature. The identification of opposite characters is a challenging topic because the character relationships are intricate and complex. The cryptic expression in literary works can enhance the readability of the plot, but it is obvious that it increases the cost of understanding. Even for human readers different types of characters are difficult to classify and identify, and it usually needs careful reading for several times to get the subtle differences among the characters. Though extracting knowledge from Chinese literature is a complicated and difficult topic, this paper models the graph of polar literary characters and divides polar communities by extracting polar vertices and polar edges based on the concept of literary sentiment polarity. The results show that using a long-sentence window is a good trade-off. The experiments of modern Chinese polar literature show that the accuracy of the community division method for the integrated graph polarity is obviously better than the method based on the co-occurrence network, and it can automatically match the positive and negative communities as well as build the complete graph structure. It clearly reflects the meaning of literary works from the perspective of polarity. This systematic model describes general standard operation steps for the machine to understand complex Chinese literature. The experiments conducted on seven benchmark Chinese novel datasets demonstrate that the method based on emotional polarity shows a significant improvement compared to baseline performance. Though we use Chinese datasets in this paper, the model and methods are important references for literature analysis and graph extraction in other languages.

INDEX TERMS Community division, character graphs, emotional polarity, natural language processing, social network analysis.

I. INTRODUCTION

Literature is an important form for people to spread culture and exchange ideas by shaping images and reflecting social life through language [1]. The rich text mining value of long texts carries and records much information related to human society in the history of Chinese culture. With literary long texts, many political, economic and cultural content descriptions in human life are preserved for posterity and can be studied. Digital technology has promoted the
The exponential increase in the amount of electronic text information makes it fast and effective to obtain target information in massive texts, which has become an urgent need [3]. Text mining techniques can be used on-demand to acquire the knowledge and potential useful value for users who are curious or interested [4] and to extract previously unknown, understandable, and potentially useful value patterns and knowledge from a large number of long-text corpora [5]. It can solve the problems of management, utilization, and application of many unorganized electronic long texts. Analyzing the network of persons in literary works [6] helps publishers and readers find the information they need in long texts, assists publishing decision support, mines long-text knowledge bases, and effectively improves the accuracy of text retrieval and plot understanding.

A typical research area in social network analysis is community division, also known as community discovery [7]. Through user behavior analysis and social network sentiment analysis [8], we determine the closely connected node groups in social networks. Modern Chinese literature is a very good dataset. It reflects real life and reveals the contradictory aspects of modern Chinese society [9]. This type of literature has rich emotions and obvious community differentiation. This article uses “polarity” to describe the sharp opposition between this emotion and the community.

Community detection is an important tool to analyze hidden information such as functional module and topology structure in complex networks [10]. Traditional social network analysis of literary works is based on co-occurrences to determine the relationship between characters in a novel and to construct and analyze the character network [1], which usually discards the wonderful emotions, plots, and characters in the literary works. If one wants to mine the rich and internal textual information, the intrinsic linguistic connection with the literary language of the plot must be explored first.

Although novels are fictional, authors fully combine the understanding of life and social experience from real society in the creative process. Therefore, his or her creative works are a reflection of society, and the analysis based on it is reasonable and can be indirect; it reflects the social value of the author’s cognitive perspective.

Previous studies [11]–[14] mostly focused on selecting influential nodes in social network data independently, while they still need identifying other classes of nodes for pivot selection. Reference [15] simply utilized the node centrality to generate the structure-based network and did not consider contexts. [16] provided an overview of the characters and the relationships between them from the narrative texts but no further analysis of the network structure was presented. In [17], the authors conducted a community detection and character categorization which identify the protagonists and antagonists. This dimension gives limited information from the aspect of frequency about the original literature. Overall, few of previous works have completed graph structure extraction [18] and community division from exacted polar nodes. Extracting knowledge from Chinese literature is a complex and difficult topic for the following reasons:

First, to improve the attractiveness and readability of the works in the creative process, the author usually deliberately conceals, covers, reverses, reveals, and darkens the relationship between the protagonist and the antagonist, which makes it difficult to divide the characters into subgroups [19].

Second, the novel’s natural language has complex semantics and twists and turns. Although there are many techniques that can be used for reference, more fully theoretical experimental verification is needed in named entity recognition [20] and emotional polarity confirmation.

Third, although the amount of digital text is sufficient, the proportion of the labeled text is low, manual labeling confirmation lacks a uniformly standard definition, and the cost is high, which brings difficulties to our work.

In addition, the characters in novels usually have dynamic personalities, which are hidden in various aspects of expression. Even for human readers, it is difficult to classify and identify different types of characters, and it usually needs careful reading for several times to get the subtle differences among those characters.

Finally, there are limited number of datasets available to apply machine learning algorithm [21], [22] and it is quite time-consuming to produce denotations on large scale [23]. It is laborious to label the prior knowledge manually in some scenarios where prior knowledge is hard to obtain [24].

Our work combines the theory on natural language processing, knowledge graph and social network. We systematically propose a integrated framework, conduct the experiments on typical datasets, present the outcomes and solve this unsolved topic by solving the problems synthetically from different fields. We establish the systematic method and start from relatively polar literature that consists of characters with a steady and clear personality. It is a novel trial for community division with character graph building and polarity information extraction. This paper uses natural language processing based on the concept of literary sentiment polarity to select a representative polar literary dataset, extract characters and sentiment indication information in literary space, and mine the node polarity and link polarity contained in the text. After modeling the network of polar literary characters, we divide
polar communities in the network. This study is a comprehensive application in the field of text classification and information extraction that integrates natural language processing and social network theory, combined with extracting graph data and visualizing for these data and knowledge structure.

The results show that the use of long-sentence windows can ensure the accuracy of literary interactions in the window during information extraction, and the information contained in the window is not very sparse, which can better ensure the network modeling accuracy. As there is no related benchmark for this work, we select the basic community detection method with general co-occurrence parameter to work as baseline. The accuracy of the community division results of the integrated network polarity method is significantly improved compared to the method based on the co-occurrence network results. It can automatically match the positive and negative communities and includes the complete link structure of the network that clearly reflects the literature from the perspective of polarity, which is the intention of this work.

Our main contributions in this paper are as follows: First, a systematic model is well defined to process long natural language text. It provides general standard operation steps for the machine to understand complex Chinese literature and extract graph data. Second, we propose four methods including evaluation standard of parameter to achieve building a character graph and dividing communities. They combine emotional information with social network computing to form a character network with emotional knowledge. Third, experiments conducted on seven benchmark Chinese novel datasets give out the accuracy and reliability of the methods as well as visualized graph data structure to make them more applicable and demonstrate that the method based on emotional polarity shows a significant improvement compared to baseline performance. Instead of costing manpower to analyse graph knowledge in texts as before, the automatic methods provide more change for application on data mining. With texts in literature industries, the methods are designed to benefit writers, readers and publishers. With other types of compatible texts, the methods are favourable to politicians, governments, media industry and knowledge digitalization.

II. GRAPH DATA EXTRACTION AND CHARACTER GRAPH BUILDING
A. POLARITY DATA AND THE POLAR LITERARY CHARACTER NETWORK
Literate emotional polarity (LEP) refers to the obvious literary emotional tendencies of the characters and the relationships between the them in literary works under a specific classification criteria based on community relation factions, such as the antagonist, when the authors develop the plots and describe the characters’ behaviors.

The discrimination of literary polarity provides a sense of community division within the literature. As characters in a literary work can be determined by emotional tendencies, character relationships can also be determined by emotional tendencies and values. Literary works of polarity are referred to as polar literature (PL).

The content of PL shows polarity. Therefore, the nodes, relationships, and communities extracted according to the content also show polarity. Together, they form a network model containing polarity information and therefore we can build the knowledge graph [25]. We will define and explain these concepts as below.

Definition 1 (Polarity of Vertices, PoV): In PL, all the characters described by the author are called literary character sets $V = \{v_1, v_2, \ldots, v_N\}$. If the literary content has a clear positive or negative tendency to a certain node $v$ in $V$, $v$ has the vertex polarity, and $v$ is called a polar vertex (PV). Each vertex with polarity can determine the degree to which the literary content praises or denounces its emotions and the vertex polarity value $p(v) \in R$ is recorded, it is also called the character polarity.

When the author’s description and the reader’s understanding reach a consensus that emotions are positive, $p(v)$ is larger than a polar threshold value $thresh$, and it is known as a positive node, denoted by $v^+$; a negative emotion corresponds to a negative node, where $p(v)$ is less than $thresh$, referred to as $v^-$. If we confirm that a node has polarity but has not completed the community division work and has not specifically determined its polarity, then we will consider the polarity node as an unclassified node $v^0$. If the emotions of node $v$ tend to be neutral, multifaceted, changeable, or the authors and readers cannot reach a consensus in order to determine $p(v)$ or obtain $p(v)$ close $thresh$, then $v$ does not have polarity, which is called a nonpolar vertex, denoted by $v^0$.

In this paper, it is necessary to remove $v^0$ from the network. In PL, the number of $v^0$ nodes is small, and the frequency is low, so it has little effect on community division [26].

Definition 2 (Polarity of Edges, PoE): In PL, the relationship between all the characters described by the author is called a link set $E = \{e_1, e_2, \ldots, e_M\}$, where any link $e$ corresponds to a node tuple $(v_i, v_j)$. If the link $e$ of a certain pair of polar nodes $(v'_i, v'_j)$ in the literary content has a clear tendency of emotional attraction or repulsion, then $e$ has a link polarity, and $e$ is called a polar edge (PE). For each link with polarity, the degree of emotional attraction or repulsion can be determined and recorded as the edge polarity value $p(e(v'_i, v'_j)) \in R$. PoE is also known as the polarity of relations.

If $v'_i$ and $v'_j$ of $e$ corresponds to the same polarity, then $p(e(v'_i, v'_j)) > 0$, it is called a positive link; otherwise, $p(e(v'_i, v'_j)) < 0$, it is called a negative link. Because the edge polarity value can indicate the edge polarity, in this paper, the link polarity value is simply referred to as the edge polarity. If the emotional tendency of the link $e$ is neutral, or there are variations among the readers in understanding the description or the emotion is undetermined to get $p(e(v'_i, v'_j))$ or $p(e(v'_i, v'_j)) = 0$, then the vertex does not have polarity, and the edge $e(v'_i, v'_j)$ is called a nonpolar link. Because the nonpolar node’s emotional tendency is fuzzy and the emotional relationship with other nodes (polar or...
nonpolar) is also fuzzy, the nodes in the corresponding two-tuple of the polar link are all polar nodes.

Definition 3 (Literary Character Network \(G(V, E)\)): It refers to the simple undirected positive-weighed network composed of the literary collections \(V\) and link collections \(E\) in literature.

Definition 4 (Polarized Literary Character Network \(G'(V', E')\)): It refers to the undirected network composed of PV set and PE set in PL. In this paper, \(v^0\) will be excluded from \(V\) to get \(V'\). For literature with rich polarity, each node has one and only one polarity \(v' \in \{v^+, v^-\}\), \(v' \neq v^0\), and each PoE has a nonzero value \(p(e(v'_i, v'_j)) \neq 0\), and all \(v'\) and \(p(e(v'_i, v'_j))\) sufficiently characterize all the information in \(G'\). Not all literary character networks can construct polar literary character networks.

Definition 5 (Characters Community, CC): CC is also called the literary vertices sub-network and refers to sub-network corresponding to the characters that are closely connected in \(G(V, E)\).

The phenomenon of communities in the network diagram is called the community structure, and the community structure is a common feature of the network. The general literary character network comprises two types of communities: communities that do not intersect with each other are called nonoverlapping (disjoint) communities, and communities containing intersections are called overlapping communities [27]. Given a network, the process of determining the community structure is called community detection. The detected communities are denoted by \(G_1, G_2, \ldots, G_C\), \(C \geq 2\), \(G_1 \cup G_2 \cup \ldots \cup G_C = G\). For nonoverlapping communities, \(G_i \cap G_j = \emptyset\).

Definition 6 (Polarity of Community, PoC): In \(G'(V', E')\), two nonoverlapping communities consist of positive vertices and negative vertices having a tendency toward emotional praise, these communities are referred to as a polar community (PC).

A polar literary character network \(G'\) has only two positive and negative polar communities. The positive community \(G^+\) is composed of positive vertices, and the negative community \(G^-\) is composed of negative vertices. In the polarized literary character network, the vertices must belong to either a positive or negative community. In this study, the network has at least one positive vertex and one negative vertex.

The goal of community division (CD) is to extract the information inherent in the literature by identifying PV, PE, and classifying the vertices into the corresponding PC. A key issue of literary polarity is the classification criteria used to determine the polarity. The construction of a literary character network is based on the reading experience and plot development. In the novels from the new era, literature has developed in a pluralistic trend, and various plot modes have been blended with each other, thereby comprehensively and profoundly showing the complex and multifaceted spiritual world of human beings [28]. There are usually two different frameworks for character classification in novels: one framework is the nature of the character itself, and the other framework is the role of the character in the work [29].

The criterion of actionability is not suitable for the classification standard of literary polarity. The judgment of social morality may be different for different readers, so the polarity should be judged by the author. From the author’s specific perspective, readers and authors can reach a consensus and obtain reliable emotional orientation results, which is a good polarity division standard, and a literary work applicable to this standard is a polarized literary work. This paper’s work is different from analyzing general social network communities; the two communities have obvious polar distinctions.

B. Text Space and the Text Space Entropy

Definition 7 (Morpheme c): In morphology, morphemes are the smallest grammatical units with a combination of phonetics and semantics.

Morphemes can be divided into free morphemes, bound morphemes, and semi-free morphemes, according to whether they can form words alone [30]. According to the different information planes used in this text, we define three types of text spaces: a natural morpheme space, a character-indicating text space, and an emotion-indicating text space [31]. According to the extraction order of text information, we first introduce the natural morpheme sequence:

Definition 8 (Natural Morpheme Sequence, NMS): \(S_N = \{c_i\}, i = 1, 2, 3, \ldots\) NMS is a finite sequence of all morphemes \(c\) in a natural language text. It is the embodiment of the unprocessed natural language of a long-text novel. It contains complete morphemes and does not include high-level information such as the section, segment, emotion, and social network level. NMS is characterized by large morpheme capacity but sparse information content.

Definition 9 (Character-Indicating Sequence, CIS): \(S_V = \{c_{V_i}\}, i = 1, 2, 3, \ldots\), in which \(c_{V_i} \in V\) and \(c_{V_i} \in S_N\). CIS indicates the specific morphemes \(c_{V_i}\) that can indicate the appearance of the character extracted from the \(S_N\), and the original morpheme sequence \(S_N\) is retained to form a \(S_V\).

Definition 10 (Emotion-Indicating Sequence, EIS): \(S' = \{c'_{V_i}\}, i = 1, 2, 3, \ldots\), in which \(c'_{V_i} \in S_N\), \(S' \subset S_N\). PIS contains morphemes \(c'_{V_i}\) that can express emotional polarity extracted from \(S_N\). NMS, CIS, and EIS are independent finite-length sequences of morphemes. To characterize the interaction and correlation between morphemes, it is also necessary to indicate the scope of morphemes, as defined below.

Definition 11 (Literacy Field Window, LFW): Under different segmentation modes (chapter identifiers, paragraph delimiters, and periods and other punctuation), the sequence is segmented to represent the literary unit with the largest distance of morpheme interactions.

The segmentation mode refers to the methods of semiotics. This paper does not consider the influence of inner punctuation on semantics [32]. The size of the LFW is called granularity, denoted as \(Q\). When a different \(Q\) is selected, the amounts of information contained in the text space generated by the text sequence are different. The chapter window
is divided according to the chapter identifier, the paragraph window is divided according to the paragraph separator, the long-sentence window is divided according to the end of the sentence, and the short-sentence window is divided according to other punctuation [33].

Definition 12 (Polarity of Window, PoW): PoW is the tendency of the LEP expressed by the entire text in the LFW. When \( Q \) is a short sentence, it is called the short-sentence polarity; when \( Q \) is a long sentence, it is called the long-sentence polarity; when \( Q \) is a paragraph, it is called the paragraph polarity. When \( Q \) is a chapter, it is called the chapter polarity, PoW is calculated as below,

\[
p(LFW) = p(S_j \subseteq S) = \sum p(c_j \in S_j).
\]  

(1)

Definition 13 (Literacy Filter, LF): Under an LFW of a certain granularity \( Q \), the morpheme sequence can be divided into a series of LFW units \( t = \{c_i, c_{i+1}, \ldots, c_l\} \). The morpheme sequence is uniquely mapped to obtain a text space \( T = \{t_i\}, i = 1, 2, 3, \ldots \), composed of literary interaction units, and the extractor transformation process is called \( T = F(S, Q) \), the process of LF is denoted as below,

\[
T = F(S, Q) = LFW_{q}(S) = (S_j \subseteq S)_{Q}.
\]  

(2)

Under a set of \( F \) with different \( Q, S_N, S_V \), and \( S' \) can be used to obtain the natural morpheme text space \( T_N \), character-indicating text space \( T_V \), and emotion-indicating text space \( T' \) respectively. The unit \( t_N \) of the natural morpheme text space \( T_N \) is called the natural morpheme window, the unit \( t_V \) of the character-indicating space \( T_V \) is called the character-indicating window, and the unit \( t' \) is the emotion-indicating space \( T' \) is called the emotion-indicating window. \( t_N, t_V \) and \( t' \) can be composed of morphemes \( c \), character-indicating morphemes \( c_V \), and emotion-indicating morphemes \( c' \). The granularity \( Q \) of \( F \) determines the information content of the filtered text space \( T \). Therefore, when \( Q \) is selected, the information contained in the text space \( T \) needs to be evaluated, as defined below.

Definition 14 (Text Space Entropy, TSE): The information content of the text space \( T \) divided by the granularity \( Q \) is referred to as the TSE, denoted as \( H \).

The text space entropy is used to evaluate the information content of the text space when extracting subspaces. Information theory states that the amount of information is related to changes in the uncertainty, and the changes in uncertainty are related to the number and probability of possible outcomes [34],

\[
H(T) = E[I(T)] = E[-\ln(P(X))].
\]  

(3)

Among them, \( X \) is the variable of interest when extracting information from the literary interaction unit \( t \), \( P \) is the probability quality function of \( X \) [35], and \( E \) is the expectation function, \( I(T) \) is the amount of spatial information in \( T \) and called the spatial self-information. In particular, since the \( X \) enumeration type is limited, when \( X \) represents the number of \( n^+ \) and \( n^- \) in a character-indicating window \( t_V \), the entropy of a text space can be represented as below,

\[
H(T) = \sum_i P(x_i)I(x_i) = -\sum_i P(x_i)\ln P(x_i).
\]  

(4)

Using chapter identifiers, paragraph separators, periods and other punctuation as segmentation identifiers, the corresponding information content is also different when the smallest unit of the natural morpheme space is divided. When \( Q \) is selected, the abovementioned text space entropy is used to evaluate the granularity.

In the following section, we divide the community based on the natural morpheme space \( T_N \), character-indicating space \( T_V \), and emotion-indicating space \( T' \) to obtain the polar community division result.

III. BUILDING CHARACTER GRAPH AND DIVIDING COMMUNITIES USING EMOTIONAL DATA OF CHARACTER NETWORK

A. COMMUNITY DIVISION BASED ON A CO-OCCURRENCE NETWORK STRUCTURE

In \( T_V \), the structure of the network model used in the community division method according to the co-occurrence of character nodes is the GN algorithm proposed by Girvan and Newman [36]. It gives out the sub-communities using weights of edges that are positive integers.

Representing the entire network and then using algorithms to complete community discovery based on the co-occurrence network structure of the community division is the method called community division based on co-occurrences (CDC).

We divide our methods into 4 parts to fully demonstrate that CDEP performs much better than CDC qualitatively and quantitively, which is GN community discovery based on co-occurrence frequency. CDC is a general but sketchy method to transform natural text into knowledge graph.

The community division of the co-occurrence network structure is completed based on the general literary character network \( G(V, E) \). Regardless of the emotional polarity information, the literary character set \( V \) and the link set \( E \) are extracted from \( T_V \), and the weights of the links \( w[e(v^i, v^j)] \) in \( E \) indicate the interaction strength between nodes.

In this way, a simple community division result is obtained, that is, two community subnets. However, since \( E \) does not contain emotional polarity information, these two subnets are the final division results, and the positive and negative polarities need to be manually matched. This method has flaws because it only considers the co-occurrence network structure and does not consider the context semantics of the co-occurrence of character nodes (i.e. co-occurring words, the emotional tendency of co-occurrence paragraphs) [37]. For example, the polarities are not clearly marked, and the division of polar communities is not accurate. We will extract more context and emotional information modularly in the following methods to improve the performance.
Algorithm 1a GN Community Discovery

Input: $G(V, E), E_W = \{w(e_i) | e_i \in E\}$
Output: $G_1, G_2, \ldots, G_C, C \geq 2$

1: def GNI($G(V, E), E_W$)
2: repeat
3: for $e_i \in E$:
4:     $B(e)$ = edge betweenness function of $e_i$
5:     $C_B(e) = B(e)/w[e(v_i, v_j)]$
6:     if $C_B(e)$ is max
7:         $(C_B(e_1), C_B(e_2), \ldots, C_B(e_M))$:
8:             remove($e_i$);
9:             $E = \{e_1, e_2, \ldots, e_i-1, e_{i+1}, \ldots, e_M\}$
10:            mod_all.append(modularity of $G(V, E)$)
11:        write_treestructure($G(V, E)$)
12:     if $E = \emptyset$:
13:         break
14:     for modularity in modular_all:
15:         if $C = 2$:
16:             modular_polar.append (modularity)
17:             when mod_all.index (max(modular_polar))
18:         $G_1, G_2 = \text{read_tree}(G(V, E))$
19: return ($G_1, G_2, \text{max(modular_polar)}$)

Algorithm 1b CDC

Input: $T_V$
Output: $G(V, E), E_W = \{w(e_i) | e_i \in E\}$

1: for $tv$ in $T_V$:
2:     if the length of $tv < 2$:
3:         $H(X) \propto 0$
4:     else if the length of $tv \geq 2$:
5:         for $(v_i, v_j)$ in $Tv$:
6:             $w[e(v_i, v_j)] + = 1$

Algorithm 2 CDVP

Input: $T_V, T'_V, V' = \{v'_i\}$, $v'_i$ is not identified
Output: $V' = \{v'_i\}$, $v'_i$ is in $\{v^+, v^-\}$

1: for $tv$ in $T_V$ and $t'$ in $T'_V$:
2:     for $c'$ in $t'$:
3:         $p(c')$ = Polarity judgment value of $c'$
4:     for $tv$ in $T_V$:
5:         $p(tv) + = p(c')$
6:     for $v'$ in $V'$:
7:         if $p(v'_i) > \text{thresh}$:
8:             $v'$ is $v^+$
9:         else if $p(v'_i) < \text{thresh}$:
10:            $v'$ is $v^-$

Algorithm 3 CDEP

Input: $T_V, T'_V$
Output: $E' = \{e(v'_i, v'_j)\}$, $p[e(v'_i, v'_j)]$ is a real number

1: for $tv$ in $T_V$ and $t'$ in $T'_V$:
2:     for $c'$ in $t'$:
3:         $p(c')$ = Polarity judgment value of $c'$
4:     for $e(v'_i, v'_j)$ in $T_V$:
5:         $p[e(v'_i, v'_j)] + = p(c')$

The set of polar links indicates the polarity interaction between nodes in the network of polar literary characters. Although it contains emotional polarity information, these two communities are the final division results, and their positive and negative polarities still need to be matched to complete the evaluation of the algorithm model. CDEP uses the PoW to measure the distance between nodes so that the communities in the network can be identified, and the community polarities can be divided by simple matching. The magnitude of the link polarity value indicates the interaction strength of the attraction or repulsion of the node. The
link polarity value is also a kind of co-occurrence number. What has been improved is that it is the number of co-occurrences after emotion indication space filtering from $T'$.

**D. COMMUNITY CLASSIFICATION BASED ON THE POLARITY OF THE ENTIRE NETWORK**

The integrated network community divided polarity, called polar community division (CDNP) takes into account both the link polarity and the node polarity.

In addition, when coupling $T_V$ and $T'$ in the corresponding windows in the time sequence, each emotion-indicating word in $T'$ is directed to the vertices in $T_V$ to calculate $p(v)$ and is also directed to the links in $T_V$ to calculate $p(e)$.

**Algorithm 4 CDNP**

**Input:** $T_V$, $T'$

**Output:** $V' = \{v'_i\}$, $v'_i$ is in $\{v^+, v^-\}$

$E' = \{e(v'_i, v'_j)\}$, $e(v'_i, v'_j)$ is a real number

1: for $t'_V$ in $T_V$ and $t'$ in $T'$:
2: for $c'$ in $t'$:
3: $p(c') = \text{Polarity judgment value of } c'$
4: for $t_V$ in $T_V$:
5: $p(v^+_V) = p(c')$
6: for $v$ in $V'$:
7: if $p(v^+_V) > \text{thresh}$:
8: $v'$ is $v^+$
9: else if $p(v^+_V) < \text{thresh}$:
10: $v'$ is $v^-$
11: for $(v'_i, v'_j)$ in $E'$:
12: if $p(v'_j)$ is the same as $p(v'_i)$:
13: $p[e(v'_i, v'_j)] = -$
14: else if $p(v'_j)$ is not the same as $p(v'_i)$:
15: $p[e(v'_i, v'_j)] = +$

The appropriate polarity threshold $\text{th}$ is determined. Each PoV can be pre-confirmed. Then, based on the assumption that nodes with the same polarity repel each other and nodes with different polarities attract each other, this interaction force relationship is fused to the polar interaction. Finally, the community classification will be handled by Algorithm 1a.

The comprehensive polarity interaction between the nodes in $G'$ is shown, which comes from the node polarity and link polarity information. The addition of node polarities into CDEP not only contains emotional polarity information but also automatically matches their positive and negative polarities as the final result of community division. The addition of the link polarity makes the result have the graph structure of the network.

**IV. EXPERIMENTS AND ANALYSIS**

**A. DATA ACQUISITION AND CHARACTER DICTIONARY**

In our experiments, seven novels with characters that have obvious positive and negative polarities in the social network of the text were selected. The text and text expansion materials were used to determine the node dictionary and to label the node polarities. Depending on the length of the text and the complexity of the plot, the number of nodes is between 6 and 22.

Text-based statistical results can help retrieve most of the information about the person’s name in the text [39] and avoid omissions when constructing the entries, but a small amount of information may be duplicated or even wrong. The text expansion material helps us match the main role of the character and other roles into the same character entry [40].

In literary works, the main character’s tendency to depreciate is obvious, and the polarity is easy to distinguish; secondary characters are mostly used to assist the protagonist and the main plot. During the extraction of $S_V$ from $S'_V$, the secondary characters corresponding to nonpolar nodes are removed and protagonists are labeled. Finally, the character dictionary used in the experiment is obtained. Taking RenMinDeMingYi as an example, the selection of the primary and secondary characters and the polarities of the primary characters are shown in Table 1.

**Table 1. Primary and secondary characters in RenMinDeMingYi.**

| Positive protagonists | Negative protagonists |
|-----------------------|-----------------------|
| Hou L.P.              | Li D.K.               |
| Sha R.J.              | Chen H.               |
| Lu Y.K.               | Ji C.M.               |
| Chen Y.S.             | Zhao D.L.             |
| Yi X.X.               | Wang D.L.             |

Secondary characters (nondipolar):

| Gao X.F. | You R.X. | Sun L.C. | Lv L. |
|----------|----------|----------|-------|
| Du B.Z.  | Wang F.Z. | Zheng Q. | Lin H.H. |
| Wu C.L.  | Zhou G.C. | Zhong X.A. | Zhang B.B. |
| Chen Q.F. | Zhang S.L. | Qian D.Z. | Wu H.F. |
| Lin J.G. | Liu S. | Sun H.P. | Bi J.Y. |
| Li L.    | Liang Q.F. | Hou H.R. | Tan X.Z. |
| Zhang T.F. | Bai C.Z. | He A.S. | Ma W.M. |
| Zhou Z.  | Wang J.L. | Jin M.S. | Wang W.G. |
| Cheng D. | Wang G.F. | Chang C.H. | Tang C.L. |
| Wu X.Y.  | Chen W.Q. | Liu S. |

Characters in the table are sorted according to the frequencies they show up in the literature, where the character on the top and left plays a more important role.

**B. SELECTION OF THE GRANULARITY OF THE LITERARY INTERACTION WINDOW**

The CDC, CDVP, CDEP, and CDNP algorithms all need to complete the extraction of $T_V$ and $T'$ from $T_V$ and $T'$, and then continue to complete the division based on these windows. The selection of the window granularity has a large impact on the performance of the algorithm.

We use the text space entropy to evaluate the information content of windows obtained by dividing the text space $T$....
with different $Q$. In particular, we choose the number of $v^+$ and $v^-$ in $t_V$ as the variable of interest $X$. Figure 1 shows the probability mass function distribution graph of $X$ with different $Q$. It maps the probability mass function on the left axis, which is the overall distribution of $X$. The right axis and the line graph represent the frequency cumulative percentage of the information entropy calculation. The sum of the corresponding frequencies of all the $X$ values is the total number of windows in the entire literary work.

As shown in Figure 1, most short-sentence windows cannot be extracted to indicate the appearance of the character and only play a role in plot development and logical connections. However, for the research of this paper, the information in the window is missing and is an invalid window because the length of the window is too short. A small number of windows can be extracted with one character-indicating morpheme, and vertex polarity may be calculated but cannot be used to calculate the link polarity, as there is a small proportion of active window. In summary, the information content contributed by the short-sentence granularity is not sufficient, and it is not suitable for selection in this work.

The chapter window regards each chapter as the smallest unit for analyzing the status of the characters and the literary polarities. After the literary space extractor of the chapter granularity is implemented, the amount of contributed information is sufficient, but at the same time, the accuracy of the polarity indication in the literary interaction window and the window error are considered. If the length of the interaction window is too large, the data are noisy and inappropriate. Table 2 further illustrates the calculated indexes when different windows are selected for 7 novels: the text length $l$, the number of windows $n$, the window word length $l'$, the text space entropy, and the maximum of morphemes in $t_V$ and $t'$. The information content measured by the text space entropy and the polar action accuracy rate measured by the number of character-indicating morphemes and emotion-indicating morphemes contained in the window cannot be optimized simultaneously. In terms of the text space entropy and polarity accuracy, long-sentence windows and paragraph windows are suitable. Taken together, the percentage of invalid windows in paragraphs is low, and there are more valid windows in long-sentence windows, which can ensure that the accuracy of literary interactions in the window and the information content in the window are not very sparse. This can better ensure network construction and the accuracy of the model. Long sentence granularity is eventually obtained with the particle extractor to get $T_N$, $T_V$ and $T'$.

### C. CHARACTER-INDICATING MORPHEME RECOGNITION

In the data acquisition and processing phase, we have constructed the node dictionary, and the extraction of nodes is based on this dictionary. The word segmentation results of natural morphemes are compared with the node dictionary, keeping the matching morphemes, and these morphemes are arranged in the original order to call a new sequence, which
TABLE 2. Index and entropy of long, short sentence and paragraph windows.

| Natural language datasets | short-sentence windows | Long-sentence windows | Paragraph windows |
|---------------------------|------------------------|-----------------------|-------------------|
|                           | $n$ $l$ $H(X)$ $i_r$ $t'$ | $n$ $l$ $H(X)$ $i_r$ $t'$ | $n$ $T$ $H(X)$ $i_r$ $t'$ |
| RenMinDeMingYi            | 258428                 | 92962                 | 11641             |
| JinGuangDaDao             | 1722512                | 206817                | 78423             |
| ShanXiangJuBian           | 32778                  | 49519                 | 23961             |
| HongQiYu                  | 291759                 | 41172                 | 16352             |
| ZhiQuWeiHuShan            | 342181                 | 41376                 | 18535             |
| YanYangTian               | 969445                 | 147624                | 67502             |
| GuoJiaGongSu              | 327185                 | 53036                 | 15904             |

TABLE 3. Examples of character morpheme recognition in RenMinDeMingYi.

| Character instruction sequence |
|--------------------------------|
| Ding Y.Z. / Zhao D.H. / Hou L.P. / Chen H. / Chen H. / Ding Y.Z. / Chen H. / Hou L.P. / Zhao D.H. |

Character-indicating space

Ding Y.Z. / Zhao D.H. / Hou L.P. / Chen H. / Chen H. / Ding Y.Z. / Chen H. / Hou L.P. / Zhao D.H.

is the CIS. Before word segmentation, a long-grained literary text space extractor generates a natural morpheme space from a natural morpheme sequence, then we perform word segmentation on the natural morpheme window in the natural morpheme space, and then matches the node dictionary. Finally, a high-confidence character-indicating text space is obtained.

D. EMOTION-INDICATING MORPHEME RECOGNITION

The analysis and recognition of polar morphemes belong to the field of sentiment analysis, and databases of Chinese sentiment dictionaries are relatively scarce. To improve the accuracy of polar morpheme recognition, we synthesize several Chinese polarity sentiment dictionaries and build a special polarity sentiment thesaurus. The NTU Sentiment Dictionary [41] and the Hownet Sentiment Dictionary [42] are Chinese emotional polarity analysis keyword sets with a total of 6110 positive words and 11152 negative words.

The morphemes that match the natural morphemes are retained, and these morphemes are arranged in the original order to call a new sequence, which is the EIS. Correspondingly, the NMS can obtain the EMS.

E. COMMUNITY DISCOVERY AND COMMUNITY MATCHING

Regardless of the emotion-indicating morphemes, CDC builds a simple literary character network. Figure 2 shows the visual result of the CDC method and it will work as the baseline for performance comparison. Without emotional knowledge being extracted, the result of the CDC is 50.00% (avg. weight) and 44.86% (link weight), see Table 7. Li D.K. (李大康) is classified as the biggest positive node in pink surrounded by other good characters such as Sha R.J. (沙瑞金) and Wang D.L. (王大路). Hou L.P. (侯亮平) has remarkable co-occurrences with those negative shown as green nodes, so it’s divided into negative community though he is a good character without polar information at this step. Gao Y.L. (高育良) is the head of negative community, other bad characters like Qi T.W. (祁同伟) and Ding Y.Z. (丁义珍) are distributed and emanated around him.

The CDVP algorithm considers the emotion-indicating morphemes. In this paper, the ratio of the number of negative...
words to the number of positive words in all the matching polarity indication windows of the node is selected as the polarity quantization value of the node. Tables 6 and 7 list the node polarity quantization values and the node polarity statistics in all the novels, including the number of positive nodes, the number of negative nodes, the maximum and minimum node polarity quantization values, and the threshold values.

The CDEP algorithm matches the link in the character-indicating window and the polarity word in the corresponding emotion-indicating window. Figure 3 shows a typical timing diagram of the link polarity for the RenMinDeMingYi. The cumulative result is used as the weight of the link in the polar literary character network. In Algorithm 1a, the weight of the network link only accepts positive values. Considering that the negative link indicates the node’s repellent relationship, weakening the repellent relationship to a link disconnection has no effect on the community discovery results.

Finally, the scattered nodes discovered by the community weakened as a result of the disconnection and were classified into the negative community. The average weight method does not consider the importance of the character in the text and comprehensively evaluates the overall effect of the model.

### E. EVALUATION OF THE METHODS

The node weight method takes the frequency of a character’s appearance in the entire literary space as a weight to evaluate the importance of the node and uses the weight to assign different discrimination scores to the nodes.

Table 8 shows the accuracy of the results of the seven novels when the four methods are used to divide the polar community, and the average weight method and the link weight method are used to evaluate the overall method. The accuracy of CDC is calculated from the common community detection method and general co-occurrence calculation. It is used as a benchmark for comparing the accuracy of the methods proposed in this paper. This method can reflect the character network in the literature to some extent, but it is not suitable for polar community division. The CDVP algorithm first determines the PoV and then builds the network structure from the positive and negative communities. The CDEP algorithm divides the community by using the common community discovery algorithm to weaken the negative link weight. The CDNP algorithm takes the node and links polar polarity. The statistical test (t-test) results given at the bottom of Table 8 show that CDNP performs better compared with basic CDC with both average and link weight. By performing statistical tests on our findings, we find that it validates the experimental results to an infinite possible number of experiments.

The performance of the CDEP algorithm is a bit lower than CDNP because of the following reasons: When the polarity indicator arrives at the link, it points to all the possible character pairs. The indication relationship is not clear, although the influence of the polarity effect is eliminated as much as possible through the choice of the window granularity. However, the links are combined with the corresponding characters without regard to the true existence and high reliability of the linking relationship. The accuracy of the sentimental polarity
FIGURE 3. The interaction data between characters by the timeline in RenMinDeMingYi. The red lines show the positive interactions between the two nodes, which means how friendly they are to each other. While the green ones represent negative and hostile interactions. The gray lines are the superimposed of all their interactions. Through the entire timeline, the accumulated value will be calculated as PoE in the graph.

The keyword set may not be adaptable to the link polarity context. Most of the existing sentimental polarity keywords are attached to a single subject, that is, a node. While these emotional indicators should have been attached to several subjects but there are fewer words that can have a dual-agent or multiagent relationship. With all these words pointing to the link together, there is considerable noise in the link polarity calculation.

The CDVP calculation does not face the two problems above and achieves better test performance. In the CDNP algorithm, the influence weight of the node polarity is set to a higher value, which can include the network structure information while the polar community is better divided so that the positive and negative polar nodes can be connected to form a complete network structure. For instance, in the graph of RenMinDeMingYi in Figure 4, Zhao L.C. (赵立春) and Gao Y.L. (高育良) are regarded as bad characters shown as square nodes because they both have polarity values smaller than the threshold. And they are closely connected to each other because their interactions always indicate that they have positive emotions towards each other. Ouyang J. (欧阳晋) and Ding Y.Z. (丁义珍) are also bad characters but have a closer connection, while Lu Y.K. (陆亦可) and Ji C.M (季昌明) are connected good characters shown as round nodes because of polarity value are higher than the threshold. As Hou L.P. (侯亮平) and Gao Y.L. (高育良) occurred next to each other frequently and EISs indicate negative polarity, there are a few edges are confusing and Hou L.P. (侯亮平) results as a false negative node in the network. In the graph of ZhiQuWeiHuaShan (see Figure 4), Zuo S.D. (座山雕), Ma X.S. (马希山) and Xu D.M.B. (许大马棒) are in the negative clustered community, while Shao J.B. (少剑波), Li Y.Q. (李勇奇) and Yang Z.R. (杨子荣) are connecting in the positive community.
By analyzing the accuracy results, it can also be found that in the novels of the 1970s and 1980s, the emotional tendency was more obvious, the polarity of the natural language nodes became stronger, and the model discrimination accuracy was higher. At the same time, when constructing the novel character network, the longer the novel is, the greater the number of nodes, the more information and the higher the accuracy.

As for the efficiency of the methods, the classification result of the graph is relatively simple, but a big difficulty with this method is that as the number of emotional words increases, the detection time of these words will increase linearly. We chose basic and naive filter for searching emotional morphemes, which took relatively more time but perform easily and equally. The searching process can be substituted by deterministic finite automaton (DFA) or Bayesian spam filter (BSF) to extract graph knowledge from the huge amount of literature materials.

V. CONCLUSION

Chinese literary works are precious cultural treasures and important datasets for natural language processing. However, the data labeling and cleaning phase is complex and time-consuming, and there is a lack of sufficiently labeled sets to use machine learning methods for analysis.

This paper identifies the characters and emotional morphemes from natural text data and divides the literary interaction window. Based on the concept of literary sentiment polarity, the graph model of polar literary characters is built, and the polar communities are divided based on the extraction of polar nodes and polar links. The literary space entropy clearly defines the process of extracting information from the natural text in long texts, and the literary interaction window defines the range of nodes and polar morpheme indicators. The graphs built with CDC are the baseline for comparing the performance. Most of the previous works on network...
analysis are based on co-occurrence only, so the analysis is limited in interaction frequency regardless of how emotional interactions affect the relationships. Therefore, the results of those community division methods only describe some parts of the exact original literature. In our methods, we attach emotional polar information to nodes and edges, so that characters and their relationships contain abundant knowledge about how they act and interact with each other. This multidimensional information has a significant contribution to graph building, social network computing and social network analysis such as community division. The experimental results show that the accuracy of the CDNP algorithm is much higher than the baseline. The key work of this paper is focused on the network structure using a dictionary to achieve the effective division of opposite characters. Though the method based on the link polarity has some limitations, it contributes valuable information of network structure which can be combined with the method based on node polarity. The CDNP method takes the advantages of them and can automatically match the positive and negative communities as well as include the complete link structure of the network. It clearly reflects the connotation of literary works from the perspective of polarity.

This systematic model provides general standard operation steps for the machine to understand complex Chinese literature and extract graph data. The experiments conducted on seven benchmark Chinese novel datasets demonstrate that the method based on emotional polarity shows a significant improvement compared to baseline performance. It is an unprecedented and beneficial effort to analyze polar characters in literature. This work plays a role model for further work on long-text literature understanding and is a meaningful reference to researches on natural language processing in the future.

As an interdisciplinary problem, it took impressive efforts to complete the fusion and the connection of semantics, network theory, emotion quantification recognition. And it is a really complicated and difficult topic to extract knowledge from literatures. Especially for Chinese literature, rhetoric sentences, polysemic base, lack of labeled data and dynamic personalities for characters make it even more laborious to convert the natural language to semantic data and calculate the relationships among the entities. We manage to overcome these difficulties, simplify and model this problem into a standard operation algorithm. The practical application of the method is that readers can quickly find the positive and negative nodes in the literary work to understand the characters and plots with a small amount of manual assistance and assist decision making for writing and publishing [43]. It is a bold experiment and novel attempt for computers to understand natural language literary works. These methods can be used for reference when integrating short texts and analyzing polar networks of public opinion. Though it is always a tricky and troublesome problem to extract emotional information, recognize character graph and divide community into specific sub-communities, we get a satisfying result for these problems with the methods in this paper.

There is a saying in Chinese, *In the book, there is a house of gold; On the book, there is a shade of jade*, which is used to say there is tremendous value in text database remained untouched. The identification of positive and negative characters is a hard topic because the character relationships are intricate and complex, and even human readers are difficult to identify. This paper has completed the analysis of novels with distinctive personalities. We hope these works can help researchers analyze the more subtle and delicate novels in the future. It provides a benchmark for graph knowledge extraction from literature which aims to improve the efficiency of understanding. Though the experiments are conducted in Chinese datasets, we believe that the proposed model and methods are referential for researches in other languages.

For further research, time can be considered as a variable in a character graph analysis [44], which would capture the character-spaces as the narrative is unfolded [45]. Some well-defined temporal centrality measures will categorize the actors in different time spans. More stylistic analysis according to specific literature of different languages will enhance flexibility of our methods [46], [47]. It can also be used with online datasets to prioritize the responses and better manage numerous posts [48]. For the basic and supportive process of the natural language, customized entity recognition method will make it more efficient [49] and multi-polar emotion calculation will represent multi-dimensional knowledge for tremendous literature. With the character recognition method [50], our work will contribute to its promising performance, provide a good benchmark that can assist investigations of intelligent heritage in historical document images. Among with methods that identify and analyse other types of relationships, it will assist extracting knowledge from films because adaptation of a film is the transfer of a novel to its visual medium [50]–[52]. We believe that in the next few years NLP will be a promising technology and extract more interesting results from huge volumes of works of literature in human history.

REFERENCES

[1] J.-S. Zhao, L. Zhang, Q.-M. Zhu, and G.-D. Zhou, “Extracting and analyzing social networks from Chinese literary,” J. Chin. Inf. Process., vol. 31, no. 2, pp. 99–106, 2017.
[2] F. L. Aldama and P. C. Hogan, Conversations on Cognitive Cultural Studies: Literature, Language, and Aesthetics. Columbus, OH, USA: The Ohio State Univ. Press, 2014, pp. 93–94.
[3] D.-Z. Zhou, “Structure of text community based on complex networks,” M.S. thesis, School Econ. Manage., Xi’an, China, 2014.
[4] X. Yang and C.-Y. Huang, “The overview of text mining technique,” Sci. Tech. Inf., vol. 33, no. 55, pp. 82, Nov. 2009.
[5] G.-Q. Guo, “Web text mining technology,” Comput. Netw., vol. 1, no. 77, pp. 114–116, Jan. 2004.
[6] J. Li, C. Zhang, H. Tan, and C. Li, “Complex networks of characters in fictional novels,” in Proc. IEEE/ACIS 18th Int. Conf. Comput. Inf. Sci. (ICIS), Beijing, China, Jun. 2019, pp. 417–420.
[7] L. Pan, “Research on several community discovery algorithms,” Ph.D. dissertation, School Electron. Sci. Eng., Nanjing Univ., Nanjing, China, 2014.
[8] R. Lotfian and C. Busso, “Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings,” IEEE Trans. Affect. Comput., vol. 10, no. 4, pp. 471–483, Oct. 2019.
[9] L. Pan, B. Wu, and Y.-Z. Liu, “Annual report on social network group behavior and group emotion theory,” Sci. Technol. Innov. Herald, vol. 13, no. 12, pp. 163–164, 2016.

[10] X. Zhang, C. Wang, Y. Su, L. Pan, and H.-F. Zhang, “A fast overlapping community detection algorithm based on weak cliques for large-scale networks,” IEEE Trans. Comput. Social Syst., vol. 4, no. 4, pp. 218–230, Dec. 2017.

[11] P. Suppa and E. Zimeo, “A clustered approach for fast computation of betweenness centrality in social networks,” in Proc. IEEE Int. Congr. Big Data, New York, NY, USA, Jun. 2015, pp. 47–54.

[12] C.-L. Yao and N. Di, “A large-scale Web-based social relationship extraction method for people,” Pattern Recognit. Artif. Intell., vol. 20, no. 6, pp. 740–744, 2007.

[13] X. Zhao, S. Guo, and Y. Wang, “The node influence analysis in social networks based on structural holes and degree centrality,” in Proc. 7th IEEE Int. Conf. Comput. Sci. Eng. (CSEE) IEEE Int. Conf. Embedded Ubiquitous Comput. (EUC), Guangzhou, China, Jul. 2017, pp. 708–711.

[14] S. Muhuri, S. Chakraborty, and S. K. Sett, “An edge contribution-based approach to identify influential nodes from online social networks,” in Proc. IEEE Int. Symp. Nanoelectron. Inf. Syst. (iNIS), Gwalior, India, Dec. 2016, pp. 155–160.

[15] J. Chen, Y. Wu, X. Xu, H. Zheng, Z. Ruan, and Q. Xuan, “PSO-ANE: Adaptive network embedding with particle swarm optimization,” IEEE Trans. Comput. Social Syst., vol. 6, no. 4, pp. 649–659, Aug. 2019.

[16] M. John, M. Baumann, D. Schuetz, S. Koch, and T. Ertl, “A visual approach for the comparative analysis of character networks in narrative texts,” in Proc. IEEE Pacific Visualizat. Symp. (PacificVis), Bangkok, Thailand, Apr. 2019, pp. 247–256.

[17] S. Muhuri, S. Chakraborty, and S. N. Chakraborty, “Extracting social network and character categorization from Bengali literature,” IEEE Trans. Comput. Social Syst., vol. 5, no. 2, pp. 371–381, Jun. 2018.

[18] T. J. Akiki, C. L. Averill, J. C. Scott, L. A. Averill, S. Muhuri, S. Chakraborty, and S. N. Chakraborty, “Extracting social network and character categorization from Bengali literature,” IEEE Trans. Comput. Social Syst., vol. 5, no. 2, pp. 371–381, Jun. 2018.

[19] J. H. Krystal, and C. G. Abdallah, “Default mode network abnormalities in posttraumatic stress disorder: A novel network-restricted topology approach,” Neuroimaging, vol. 176, pp. 489–498, Aug. 2018.

[20] J. Zhu and J.-J. Hong, “Research on character relationship extraction method based on SDAs,” Comput. Sci., vol. 44, no. 1, pp. 141–145, 2017.

[21] D. Damjanovic and K. Bontcheva, “Named entity disambiguation using linked data,” in Proc. 9th Extended Semantic Web Conf., 2012, pp. 231–240.

[22] F. Tian, B. Gao, Q. Cui, E. Chen, and T.-Y. Liu, “Learning deep representations for graph clustering,” in Proc. 28th AAAI Conf. Artif. Intell., 2014, pp. 1293–1299.

[23] R. A. Rossi, R. Zhou, and N. K. Ahmed, “Deep inductive graph representation learning,” IEEE Trans. Knowl. Data Eng., vol. 32, no. 3, pp. 438–452, Mar. 2020.

[24] H. Cai, V. W. Zheng, and K. C.-C. Chang, “A comprehensive survey of graph embedding: Problems, techniques, and applications,” IEEE Trans. Knowl. Data Eng., vol. 30, no. 9, pp. 1616–1637, Sep. 2018.

[25] X. Zhou, X. Liang, X. Du, and J. Zhao, “Structure based user identification across social networks,” IEEE Trans. Knowl. Data Eng., vol. 30, no. 6, pp. 1178–1191, Jun. 2018.

[26] Q. Wang, Z. Mao, B. Wang, and L. Guo, “Knowledge graph embedding: A survey of approaches and applications,” IEEE Trans. Knowl. Data Eng., vol. 29, no. 12, pp. 2724–2743, Dec. 2017.

[27] B.-X. Xu and W.-H. Cai, “Discussion on the author of the story of the stone from the perspective of information theory,” J. Chin. Inf. Process., vol. 4, no. 2, pp. 3–7, 1990.

[28] J.-L. Zhang, Y.-L. Chang, and W. Shi, “Review of theories and applications of tag propagation algorithms,” J. Comput. Appl., vol. 30, no. 1, pp. 21–25, Dec. 2013.

[29] L. Shi, “Good people go bad and bad people go good: An investigation of the plot mode in Chinese modern and contemporary novels,” Masterpiece Appreciation, vol. 17, no. 44, pp. 122–124, Jun. 2010.

[30] G.-H. Tan, “Research on the classification of novel characters,” J. Jiangxi Normal Univ. (Philosophy Social Sci. Ed.), vol. 50, no. 5, pp. 85–91, 2017.

[31] X.-P. Yang, Chinese Morpheme Theory. Nanjing, China: Nanjing Univ. Press, 2003.

[32] L. Chen, Y.-Z. Hu, and W. Ai, “Extraction of power relationship and its corresponding social network in the story of stone,” J. Chin. Inf. Process., vol. 29, no. 5, pp. 185–194, 2015.

[33] Y.-B. Wan, J.-S. Yu, and C.-Y. Zhao, “Research on application of co-word analysis on relationships of characters in the romance of the three kingdoms,” Inf. Explor., vol. 7, no. 10, pp. 52–56, Jul. 2017.

[34] M. Tapaswi, M. Bauml, and R. Stiefelhagen, “Book2Movie: Aligning video scenes with book chapters,” in Proc. IEEE Int. Conf. Comput. Vision (ICCV), Santiago, Chile, Dec. 2015, pp. 19–27.
XIAOLEI ZHAO is currently pursuing the master’s degree with the School of Information, Renmin University of China and a Product Manager Trainee with the Department of Cloud and Smart Industries Group, Tencent Technology (Shenzhen) Company Ltd. His main research areas are data mining, text mining, natural language processing, social computing, and artificial intelligence.

XUN LIANG (Senior Member, IEEE) focused his research on machine learning and social computing, hosted several projects from the NSFC, and led the Business Intelligence Laboratory, Renmin University of China. His research interests include machine learning, Web mining, and social computing.

FENG TANG is currently with the Social Intelligence Computing Laboratory, Renmin University of China. His main research interests include text mining and social computing.

XUAN ZHANG is currently pursuing the Ph.D. degree with the School of Information, Renmin University of China. His main research interests include text mining and social computing.

XIAOPI NG ZHOU (Member, IEEE) is currently an Associate Professor with the Beijing University of Civil Engineering and Architecture. His research interests include data mining, artificial intelligence, and building information modeling (BIM).

XINYU GUO is currently a Database Engineering with the Department of Relational Database Service, Alibaba Corporation.

***