Hindawi
Scientific Programming
Volume 2023, Article ID 9851726, 1 page
https://doi.org/10.1155/2023/9851726

Retraction

Retracted: Study on the Influence of Wuthering Heights Characters Based on Web Analysis and Text Mining

Scientific Programming

Received 18 July 2023; Accepted 18 July 2023; Published 19 July 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] R. Wang and L. Deng, "Study on the Influence of Wuthering Heights Characters Based on Web Analysis and Text Mining," *Scientific Programming*, vol. 2022, Article ID 4326551, 11 pages, 2022.

Hindawi
Scientific Programming
Volume 2022, Article ID 4326551, 11 pages
https://doi.org/10.1155/2022/4326551

Research Article

Study on the Influence of Wuthering Heights Characters Based on Web Analysis and Text Mining

Rui Wang[1] **and Lin Deng**[2]

[1]*Department of Basic Courses, Shanghai SIPO Polytechnic, Shanghai 201399, China*
[2]*Xingzhi College, Zhejiang Normal University, Jinhua 321004, China*

Correspondence should be addressed to Rui Wang; wangrui@sicfl.edu.cn

Received 3 March 2022; Revised 17 March 2022; Accepted 24 March 2022; Published 14 April 2022

Academic Editor: Jie Liu

The rapid development of network technology and the popularity of the Internet have made people rely more and more on the exchange and sharing of network information, and the demand for obtaining information about people from the Internet has gradually increased, but the massive amount of network data has made the information about people fragmented and disorganized, and the existing work on portraits has mainly focused on the extraction of people's attributes. In this paper, we examine the artistic construction and characterization of Wuthering Heights on the basis of a tasting of the book, with the aim of presenting the "alienated personalities" hidden in the depths of the characters' consciousness and showing the author's unique creative art through an analysis of thematic ideas, the use of temporal elements, and the tracing of the creation of the characters' original forms. In the aspect of character social relationship extraction, first, the method of expanding the seed dictionary by means of synonym word forest is used to build a character relationship lexicon, which avoids the inefficiency caused by manual lexicon collection; second, a character relationship extraction algorithm based on the combination of rule matching and syntactic tree is proposed, which effectively overcomes the disadvantage of low recall rate caused by rule matching, and the average F-value of this algorithm reaches 82.61% in the experiment The algorithm achieves an average F-value of 82.61% in the experiment, which is a significant advantage over other methods.

1. Introduction

Emily Brontë's novel Wuthering Heights died of depression and low critical acclaim until the 20th century, when her reputation grew and the novel was considered a "complete and profound insight into human nature and life." The novel is a unique artistic construction, based on the love-hate relationship between two generations and the distorted human nature of life in the aberrant English society of the time. This article explores and examines the artistic construction and characterization of Wuthering Heights from the perspective of a taster [1, 2].

The theme is clearly stated. Set in 18th-century Yorkshire, Wuthering Heights tells the story of Heathcliff, an orphaned boy who is adopted by Earnshaw, the old owner of the cottage, and goes out to earn money after being humiliated and losing his love. When he becomes rich, he returns to take revenge on the landowner Linton and his children, who have married his girlfriend Catherine. Heathcliff's transformation from an orphan to avenger is a reflection of the perverse society of the time [3]. Heathcliff is discriminated against by the society he describes, and this is the trigger for his fierce rebellion. His frenzied revenge against Wuthering Heights and the Painted Hills is an indictment of a heartless society [4].

When it comes to the organization of character information, traditional methods often use manual editing and collation, which can achieve a high accuracy rate but is inefficient, and users are eager to get global information about the target character in a simple and quick search. If global information about people could be automatically extracted and collated from people data scattered all over the

Internet, and the scattered and fragmented data could be brought together to form a portrait of people and stored in a structured manner, this would greatly improve the efficiency of users in obtaining global information about people and facilitate human work and life [5].

Of course, in addition to search engines, users can also obtain information about people through specific people search systems. The most mature people search engines in the market today are Youku (https://www.ucloo.com/), Yahoo People Search (https://people.yahoo.com/), Microsoft People Cube (https://renlifang.msra.cn/), etc [6]. These people search engines for people are mainly biased toward the basic information of the person's gender, age, and place of origin, and do not present the trajectory of the person's activities, i.e., the Wuthering Heights incident that the person was involved in as reported on the Internet. Generally, when people are learning about a character object, they not only want to get basic information about the object but are more eager to get information about when and where the character was involved in what events with whom, the emotional evaluation of the character on the Internet and the size of the character's hotness, to grasp information about a character as a whole.

In this paper, we study the character portrait mining technique of text data represented by Wuthering Heights, focusing on three aspects: extraction of characters' social relations, tracking of characters' participation in events, and analysis of characters' hotness and emotions. First, the text data are divided into words and lexical annotations to extract the character entities, and at the sentence level, the social relations of characters are extracted using shallow syntactic analysis; second, the features of individual texts are extracted according to the corresponding feature extraction algorithm, and the clustering algorithm is used to achieve the aggregation of similar events, taking characters and time as clues to form a time-based character event activity; finally, the combination of Wuthering Heights' coverage. The hotness value and emotional tendency of the characters are calculated by combining various factors such as the amount of reports, comments, readings, and time span of Wuthering Heights. The above analysis results are combined to form a character portrait, and the research results can be applied to character search systems, specific target tracking, and online celebrity detection.

1.1. Related Work

1.1.1. Personality Information Extraction. In the area of character information organization, existing work has focused on the field of character biography and character search. After the authors of [7] proposed the concept of biography, many scholars have worked on this area, resulting in a variety of methods for the extraction of biographies, mainly including methods based on multidocument summarization techniques, ontology-based and character-tracking-oriented. The authors of [8] realized the extraction of biographies using multi-document summarization techniques. The method combines linguistic knowledge and statistical theory to extract the basic information of an object, including name, gender, education, etc., from multiple documents to form a biographical text. The authors of [9] implemented a multidocument biographical summary system, mainly using the idea of classification to divide sentences into corresponding clusters of classes, first formulating a classification of sentences about the biography: social relations, educational background, place of origin, work, etc., and then using a classification algorithm to obtain the sentences that best characterize the characteristics of the person, which are eventually combined to form the biography. The authors of [10] proposed the concept of "meta-events" and applied it to the field of character information extraction, where "meta-events" are acts consisting of three named entities: people, time, and place. The authors of [11] proposed constructing an ontology of events for people and to realize this through an ontology description language.

Personal information extraction on the other hand is a research work for personal search engines, which started late. The ArnetMiner system [12] developed by [13] mainly targets experts in academic fields and mines personal information from their personal homepages, published papers, social networks, and other data. The authors of [14] proposed a personal information mining tool for social media such as Linkedln and Facebook. The authors of [15] proposed a rule-based algorithm for extracting personal information, focusing on summarizing rules such as place of origin, date of birth, and political appearance and developed a semistructured personal information extraction system. The authors of [16] proposed a person information extraction based on trigger words, realizing the extraction of person attribute information from Baidu's encyclopedia web pages, first by developing a trigger word list through linguistic analysis and second by automatically discovering candidate rules based on the word field around the person's name using statistical principles. The authors of [17] for information extraction in the field of teachers, first used SVM to classify the crawled down web pages and selected those containing person information, second developed a rule base for person attribute extraction, and finally used the rules to achieve information extraction of computer teachers in universities. The authors of [18] proposed a method based on double-layer cascaded text classification to extract personal information from resumes. The authors of [19] extracted personal information from personal homepages and CVs by a method based on trigger words and rules. The authors of [7] proposed a personal information extraction algorithm based on semantic context analysis, which incorporates the theory of hidden Markov model, semantic analysis, natural language processing, and information extraction.

1.2. Organization of Character Events. The current research on the organization of personal information is mainly focused on the extraction of personal information, and there is still a need to go further into the character activity events or the tracking of character events. In the work of [2], the concept of

"personal tracking" is proposed, which applies topic recognition and tracking to the extraction of personal events and proposes that personal events consist of three elements: time, place, and event description. The authors of [3] proposed a single-pass-based topic recognition algorithm, which is simple and fast, but the biggest drawback is that it is sensitive to the order of text arrival; the authors of [4] proposed a cohesive hierarchical clustering algorithm to solve the problem of hierarchical topic detection for the situation that multiple topics may exist in a text (i.e., there may be intersection between texts at multiple levels). The authors of [5] used the K-Means clustering algorithm to achieve topical recognition, clustering K points in the text as the centers of class clusters and dividing all texts into the closest class clusters at once, and then by continuous iteration.

To address the shortcomings of the single-pass algorithm, which is sensitive to the input order of the text, the concept of batch processing is introduced to improve the accuracy of the clustering algorithm by first clustering a batch of arriving text, then comparing it with existing classes and introducing the process of adjustment and "resurrection." In the study of [3], the automatic prediction of the number of clusters is investigated to address the shortcomings of K-Means, which requires the prior determination of the number of clusters and is sensitive to noise points and initial points. The authors of [8] proposed a new treatment for the determination of the initial centroids.

2. Main Characterization

2.1. Paranoid and Brutal Heathcliff. Heathcliff was tough, rude and rebellious, but a man with a passion for love. He was a childhood friend of Catherine's, but after the death of old Earnshaw, Heathcliff was reduced to being a servant of Sindre, deprived of an education, and the chance to become a civilized man. However, Heathcliff's love for Catherine would have allowed him to endure any torment from Hindley, something that would have been difficult for ordinary people to do. Heathcliff, who was from a humble background in a materialistic society at the time, was filled with a deep sense of depression and inferiority and was acutely aware of the inequality of treatment brought about by the disparity in status and money. His hatred for Sindre, who is a member of the upper class, is so great that he can only endure it when he is unable to retaliate, but deep down, he has already developed the idea and intention of revenge. Admittedly, the change of role from outcast to avenger suggests that he was not born that way, but because of Sindre's servitude and abuse and the disappointment of love [12]. When Catherine married Linton of Painted Hills, he was completely broken by the absence of that powerful love in his heart. Faced with an awkward situation, the darker side of his character begins to surface and he leaves with anger and hatred. He returned a few years later, rich and burning with the fire of revenge, believing from the beginning that he would be happy with the torment of the past as long as he could succeed in his revenge. It is not difficult to explain his subsequent desperate attempts to acquire the two great estates and his vast family fortune.

3. Selfish and Wild Catherine

Wild and untamed, Catherine and Heathcliff are a matched made in heaven. Catherine is the only person in Wuthering Heights who gives Heathcliff a solace of mind, and both young are spontaneous, brave, and strong in their approach to love. However, when they break into Wuthering Heights, Catherine begins to waver between the naked idea of true love and the rich and noble reality, secretly comparing the rich, gentle, and civilized Linton and the poor, rude, and primitive Heathcliff. After her marriage, Catherine also tried to be a graceful, polite lady, creating the illusion of true love for Linton, but this inevitably involved concealing and repressing herself and hiding a different spiritual world from Linton. The prolonged separation of spirit and body caused the otherwise lively and energetic Catherine to suffer and spend her days in depression.

4. Mainstream and Secular Linton

Catherine's husband, Linton, was a wealthy, gentle, and generous man, a typical heroic figure in the Victorian fiction. His gentlemanly manners and unique charm had fascinated Catherine, and his upper class label of wealth, civility, and elegance made Heathcliff jealous. Linton was accommodating, courteous, and caring toward Catherine. He treats his relatives and servants with the same kindness and humanity that shine through at all times. He is the embodiment of the prevailing values of society, the guardian of the secular moral order, and the bearer of civilized order and decency for generations. He is also weak, introverted, and hypocritical.

5. Ruthless Cowardly Sindre

As a result of the accidental appropriation of his father's love, Sindre becomes cold, brutal, and even self-absorbed. The absence of his father's love is the origin of Hindley's hatred, and Heathcliff's arrival gives him an object of hate. After his father's death, he uses all means to abuse and mistreat Heathcliff, which is the key point of the entire novel.

6. Innocent Helpless Second Generation

Hindley's son Harington could have been a gentleman in mainstream society, but his mother died soon after his birth and without his mother's care he went from innocent and sweet boy to the rebellious and disobedient brat. Moreover, all this change comes at the hands of Heathcliff, who torments and persecutes Harington in the same way that Hindley persecuted himself. Fortunately, his love for young Catherine makes Harington aware of his own shortcomings, awakening his self-respect and good nature, and eventually, through his own efforts, he becomes a handsome, civilized young man.

The entire act of revenge carried out by Heathcliff, involving different degrees of human distortion shown by each individual on different occasions, is also a true and objective reflection of the distortions and struggles of human nature in the social context of the work.

7. Character Relationship Extraction based on a Combination of Rule Matching and Syntactic Trees

This section deals with the automatic extraction of social relations of characters from texts. The social relations of characters are the general term for the mutual relations formed by human beings in the material conditions of their common life, e.g., "father," "daughter," "friend," "colleague," etc. "The term "relationship" or "character relationship" is used later to refer to the social relationships of people. As people are in a large community, they are inevitably connected to other people in a variety of ways, and if relationships can be automatically extracted from Wuthering Heights reports on the Internet about characters, this is essential to the portrayal of characters. Phrases describing character relationships tend to be fragmented, and if the character relationships implicit in the sentences can be extracted through syntactic analysis, this will help to improve the effectiveness of the relationship extraction.

The paper proposes a character relationship extraction algorithm based on a combination of rule matching and syntactic trees, which broadly consists of the following steps:

(1) Data preprocessing: first, the text is divided into words and lexical annotation; second, the personal names and relationship words are identified, and finally, candidate sentences with possible personal relationships are filtered out in terms of sentences.

(2) Rule matching-based personal relationship extraction: first, building a rule base, second, using the rules to match the candidate sentences obtained in the data preprocessing step to achieve the first extraction of personal relationships.

(3) Character relationship extraction based on a syntactic tree: syntactic analysis is performed on candidate sentences that do not satisfy the rules, a syntactic tree is built, and the second extraction of character relationships is performed through the path distance between the relationship words and the two character entities in the syntactic tree.

(4) Character relationship determination: after obtaining all relationship words between two characters through the extraction of the large-scale corpus, the relationship words with the highest weight are selected as the final character relationship determination by merging the relationship synonyms.

The specific flowchart of the character relationship extraction algorithm based on the combination of rule matching and syntactic tree is shown in Figure 1, and the details of these parts are described below.

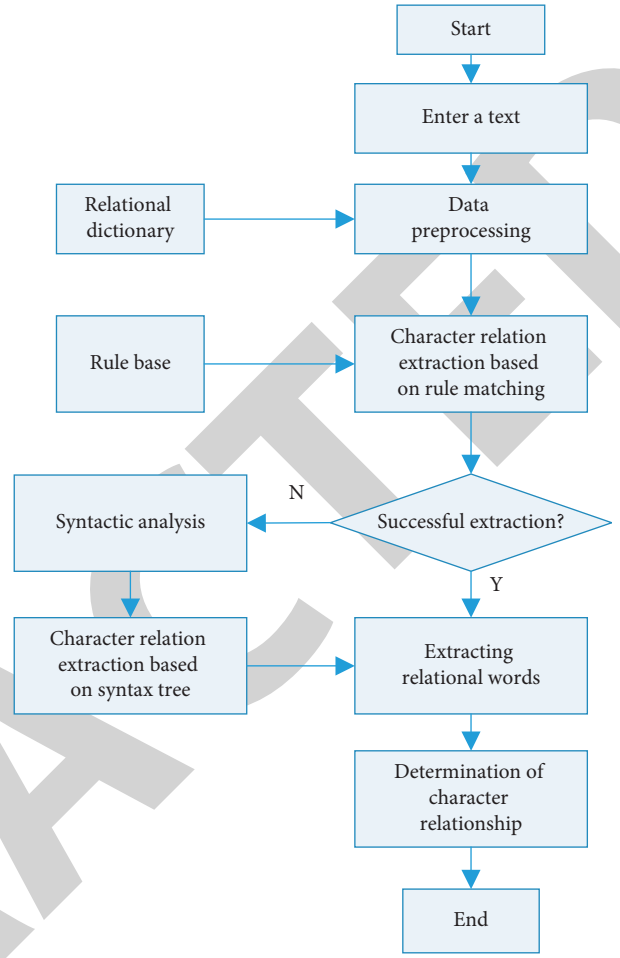

Figure 1: Flowchart of the character relationship extraction algorithm.

8. Data Preprocessing

Before proceeding to the subsequent relationship extraction algorithms, the data first need to be preprocessed, including word separation and personal identification. Second, subsequent rule-based matching and syntactic tree-based relationship extraction methods are all based on sentences. Since a large number of sentences in the text do not contain person relationships, it is necessary to first perform sentence separation (the sentences are separated according to the sentence separator in the Chinese grammar) and then perform sentence screening, and the valid sentences selected are used as an input for the subsequent algorithm. This can filter a large number of irrelevant statements and improve the efficiency of the algorithm. The flowchart of data preprocessing is shown in Figure 2.

8.1. Rule-Based Matching for Character Relationship Extraction. The person relationship candidate sentences have been obtained through the data preprocessing stage, this stage will elaborate on the extraction of relationships using rule matching on person relationship candidate sentences, this part of the work is mainly to improve the accuracy rate and to prepare for the later extraction of person relationships based on syntactic trees. The first step in the rule-based matching approach is to develop a complete and accurate rule base. As the social relationships between people are described in a relatively fixed way in the text, the thesis adopts a manual collection approach for the

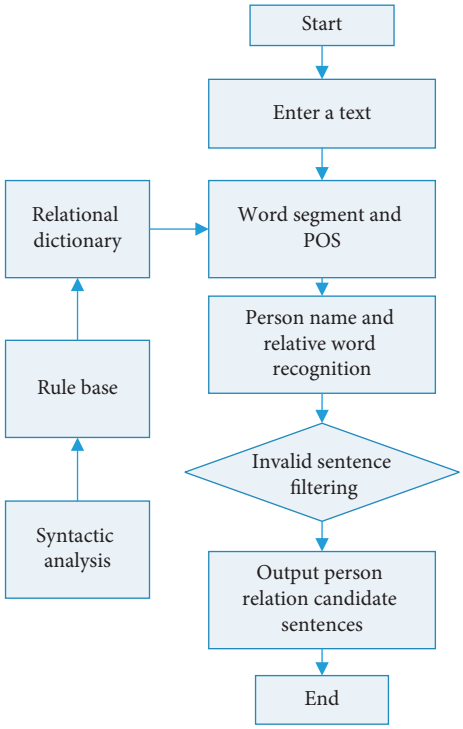

Figure 2: Flowchart of data preprocessing.

development of the rule base, which can ensure the accuracy of the rules.

Candidate sentences for character relations that have been split are matched separately using each regular expression in the rule base, in the order of highest to lowest frequency of rule occurrences when matching, which will play a role in improving the accuracy of the algorithm, while reducing the number of irrelevant sentences matched. If a sentence cannot be matched by all rules, it is used as the input to the subsequent syntactic tree-based algorithm for the next extraction step.

9. Syntactic Tree-Based Character Relationship Extraction

The syntactic tree based person relationship extraction includes three steps as follows: (1) constructing a syntactic tree by syntactic analysis of person relationship candidates; (2) pruning the syntactic tree to eliminate a large number of non-joint points to form the shortest path containing tree (SPT tree); and (3) calculating the weight of each relationship word based on the path distance between the relationship word and person pair, (p_i, p_j) in the SPT tree, and finally determining the relationship word that best represents person pair, (p_i, p_j) by using the weight of the relationship word.

The flowchart of the syntactic tree based character relationship extraction is shown in Figure 3.

The thesis uses the Stanford Parser syntactic parser for syntactic tree construction, a context-independent syntactic parser based on probabilistic statistics developed in Java by the Stanford University NLP Research Group, which is fully open source and supports English, Chinese, German, and French. Stanford Parser can obtain the dependencies between components in a sentence and the syntactic tree of the sentence. For the processing of Chinese, Stanford Parser provides five training models.

In this paper, we focus on the processing of simplified mainland texts and therefore use the Xinhua corpus. In terms of model selection, the Factored and PCFG were compared in terms of time and space consumption when processing sentences of different lengths. Stanford Parser version 3.5.2.1 was used for the comparison [20].

Once the SPT tree has been obtained, the structure of the SPT tree needs to be parsed to determine the best relational words to describe the character pairs (p_i, p_j). In this case, it is necessary to obtain the distance of each relation to each character pair. The smaller the distance, the more relevant the relation is to describe the relationship between the character pairs. The following definitions are given first:

Define a SPT tree in which the node corresponding to person p_i is node_i, the node corresponding to relation r_k is node_k, and the nearest common parent of node_k and node_i is root, then the distance $\text{dis}(r_k, p_i)$ of relation r_k to person p_i is defined as follows, where d denotes the shortest path length of node_i back up to root.

$$\text{dis}(r_k, \text{p}_\text{i}) = d. \tag{1}$$

Define the distance r_k of the relation (p_i, p_j) to the character pair, $\text{dis}(r_k, \langle p_i, p_j \rangle)$ defined by the following:

$$\text{dis}\left(r_k, \langle p_i, p_j \rangle\right) = \text{dis}(r_k, p_i) + \text{dis}\left(r_k, p_j\right). \tag{2}$$

In the SPT tree shown in Figure 4, the nearest common parent of the node corresponding to the relation "Dad" and the character "Li Ping" is the "NP" node numbered 1 in the diagram, then the shortest path of the character "Li Ping" up to the nearest common parent node is NP (NR) ⟶ NP ⟶ DNP ⟶ NP, i.e., dis (Dad, Li Ping) = 3, and similarly dis (Dad, Li Jiantao) = 1. Therefore, using the definition, the distance dis (Dad, Li Jiantao) is for the character pair Li Ping and Li Jiantao. The distance of dis (Dad, Li Ping, Li Jiantao) = 4.

The process of the SPT tree based person relationship extraction algorithm is as follows: let the set of relationship words in a person relationship candidate sentence S be $Q = \{r_1, r_2, \ldots, r_m\}$ and the set of person names be $P = \{p_1, p_2, \ldots, p_n\}$; first, the distance $dis(r_k, \text{p}_\text{i})$ between each relationship word $r_k\,(1 \le k \le m)$ and each person name $p_i\,(1 \le i \le n)$ is calculated, second, the weight $\text{dis}(r_k, \langle p_i, p_j \rangle)$ of the relationship words to the person pair (p_i, p_j) is calculated, and $\text{dis}(r_k, \langle p_i, p_j \rangle)$ in ascending order of size are arranged, and finally the relationship description of the person pair, in order of the smallest distance and the relational word that fits within the threshold is selected as the relational description of the person pair (p_i, p_j). If a relation is already identified as the relation description of a character pair [21], it will not be used as the relation description of other character pairs in the sentence. The flow of the extraction of the character pair, (p_i, p_j) relations is shown in Figure 5.

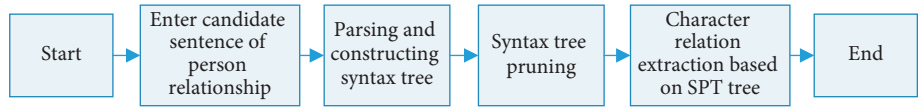

Figure 3: Flowchart of the syntactic tree based character relationship extraction algorithm.

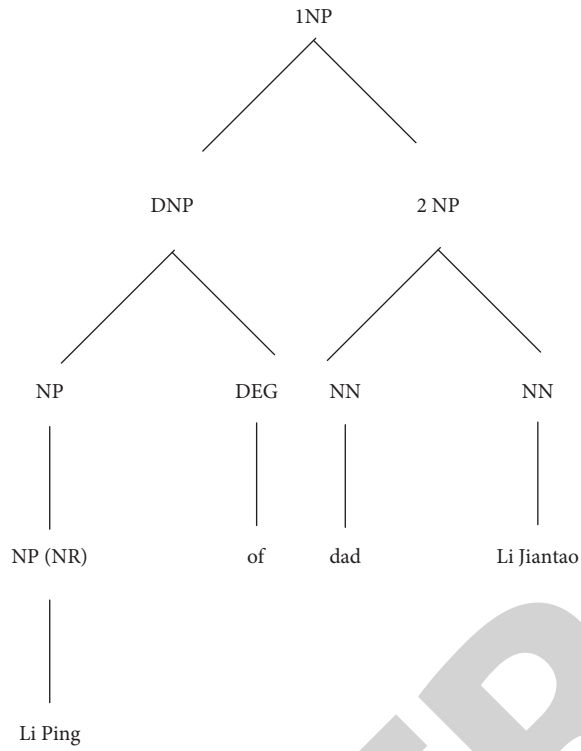

Figure 4: Schematic diagram of the SPT tree.

10. Results and Analysis of Character Relationship Extraction Experiments

In order to verify the performance of the personal relationship extraction algorithm, the 10,000 experimental data were divided into sentences and the personal relationship candidates (i.e., sentences containing at least two names and a relationship word) were selected as the input to the algorithm. 4685 sentences were personal relationship candidates, of which 2,437 contained personal relationships and 2,248 did not contain personal relationships.

10.1. Experiments on the Selection of Path Thresholds. In syntactic tree-based character relationship extraction, the choice of path thresholds is crucial to the algorithm results; therefore, this section first selects different path thresholds for experimentation and compares the results to select the best threshold The results are compared and the best threshold is selected. The paper first selects 200 relationships out of 4685 candidate sentences as training data, and manually annotates 107 relationships and 93 relationships with no relationships [22, 23]. Using the syntactic tree based relationship extraction algorithm, different path thresholds $\theta = 2, 3, \ldots, 14$ are chosen., 14 and the accuracy, recall, and F-value for each threshold are shown in Table 1.

Figure 6 shows the accuracy, recall, and F-value of the syntactic tree based personal relationship extraction algorithm for different path thresholds. As can be seen from the figure, the recall rate increases with increasing, but the accuracy rate decreases, with a maximum value of 78.19% being achieved at = 6. In the subsequent experiments, the paper chose = 6 as the path threshold.

10.2. Accuracy and Performance Comparison of Character Relationship Extraction Results. For the extraction of character relations, the thesis first performs the first step of character relation candidate sentence extraction by rule matching. If rule matching is not successful, a syntactic tree is built and the second step is performed by the syntactic tree based personal relationship. The second step of extraction is carried out by a syntactic tree-based character relationship extraction algorithm. The rule matching algorithm yields a very high accuracy rate, but a low recall rate. The syntactic tree-based approach is able to obtain a higher recall, but the accuracy is correspondingly lower [24, 25]. The combination of the two can give relatively good results. In the following, the rule matching, syntactic tree-based, and combined methods are investigated for different datasets. The following experiments compare the accuracy and performance of rule-based matching, syntactic tree based matching and the combination of the two methods. The results of each of the three algorithms are presented in Table 2.

The experimental results of the rule-based matching character relationship extraction algorithm with different data sets are shown in Table 2.

The experimental results of the syntactic tree-based character relationship extraction algorithm with different data sets are shown in Table 3.

The experimental results of the personal relationship extraction algorithm based on a combination of rule matching and syntactic trees Figure 7for different data sets are shown in Table 4.

Figures 7–9 show the comparison of the results of the three algorithms in terms of accuracy, recall, and F-value on different datasets, respectively. From the above three figures, it can be seen that the rule-based matching algorithm is able to obtain a high accuracy rate but a relatively low recall rate, and the syntactic tree-based algorithm is able to obtain a high recall rate but a very low accuracy rate. The algorithm based on the combination of rule matching and syntactic tree is able to achieve a compromise between accuracy and recall and is able to obtain a high F-value, indicating that the algorithm of the thesis is effective.

Figure 10 shows a comparison of the time consumed by the three methods on different datasets. As syntactic analysis involves the process of sentence component analysis, the relationship between the components and the construction

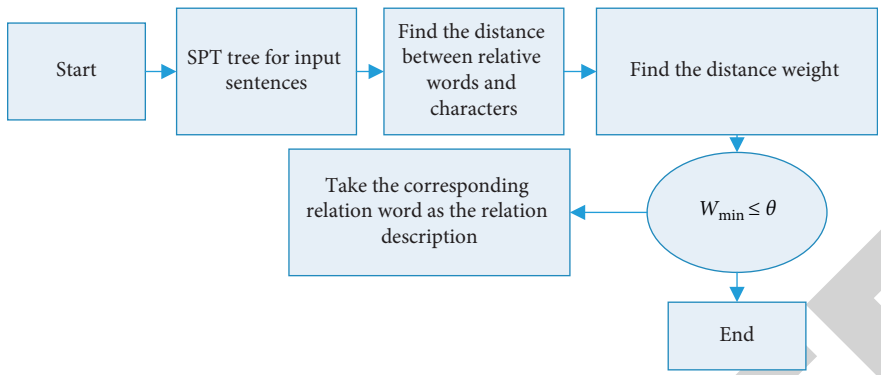

Figure 5: Flowchart of character relationship extraction based on syntactic tree.

Table 1: Metrics for syntactic tree-based character relationship extraction under different path thresholds.

Path threshold θ	Algorithm recognition median	Correct identification number	Number of false identifications	Accuracy (%)	Recall (%)	F value (%)
2	32	32	0	100	29.91	46.04
3	63	52	12	80.95	47.66	60
4	112	79	33	70.54	73.83	72.15
5	116	82	34	70.69	76.64	73.54
6	136	95	41	69.85	88.79	78.19
7	143	95	48	66.43	88.79	76
8	165	99	66	60	92.52	72.29
9	176	102	74	57.95	95.33	72.08
10	181	103	78	56.91	96.26	71.53
13	194	107	87	55.15	100	71.1
14	197	107	90	54.31	100	70.39

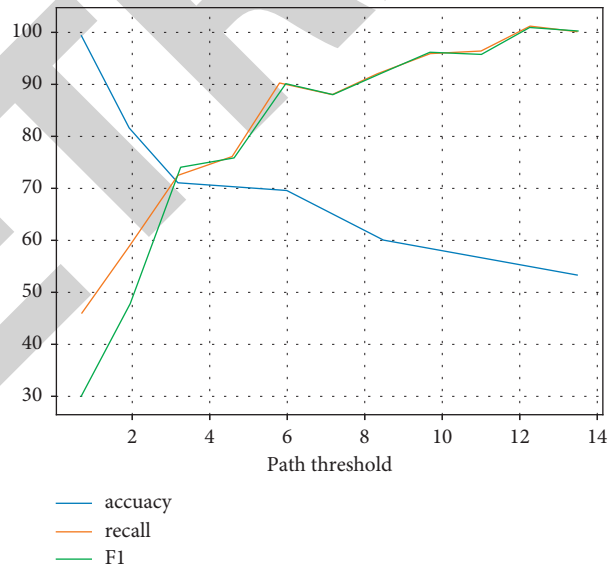

Figure 6: Metrics for syntactic tree-based character relationship extraction under different path thresholds.

Table 2: Experimental results of the rule-based matching algorithm with different data sets.

Dataset size	Total number of algorithm recognition	Correct identification number	Number of all relationships	Accuracy (%)	Recall (%)	F value (%)
40	17	15	23	88.24	65.22	75
80	30	27	41	90	65.85	76.06
120	46	41	65	89.13	63.08	73.87
160	72	63	94	87.50	67.02	75.90
200	84	76	112	90.48	67.86	77.55

Table 3: Experimental results of the Table 3syntactic tree-based algorithm with different data sets.

Dataset size	Total number of algorithm recognition	Correct identification number	Number of all relationships	Accuracy (%)	Recall (%)	F value (%)
40	33	21	23	63.64	91.3	75
80	56	37	41	66.07	90.24	76.29
120	97	58	65	59.79	89.23	71.60
160	128	82	94	64.06	87.23	73.87
200	143	98	112	68.53	87.5	76.86

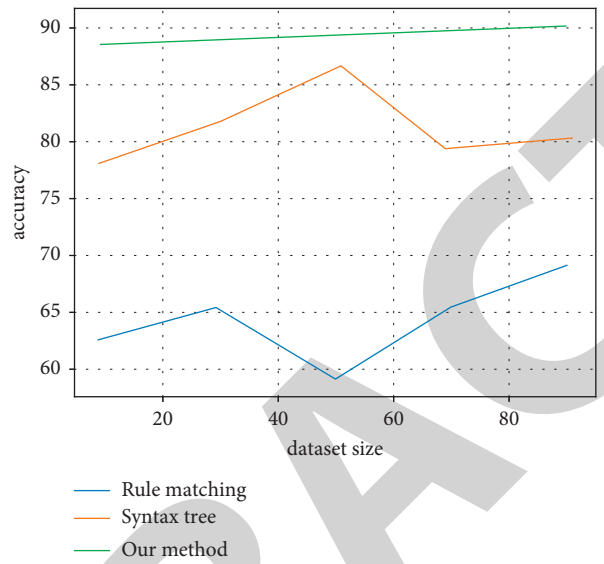

Figure 7: Comparison of the accuracy of the three methods on different datasets.

Table 4: Experimental results of the combined rule-Table 4 based matching and syntactic tree algorithm with different data sets.

Dataset size	Total number of algorithm recognition	Correct identification number	Number of all relationships	Accuracy (%)	Recall (%)	F value (%)
40	24	19	23	79.17	82.61	80.85
80	40	33	41	82.5	80.49	81.48
120	61	52	65	85.25	80	82.54
160	99	79	94	79.80	84.04	91.87
200	118	95	92	80.51	84.21	80.61

of a syntactic tree, whereas rule matching is simply a string comparison, syntactic analysis consumes a very high amount of time, whereas the rule-based matching algorithm is very fast. In the algorithm combining rule and syntactic tree, as part of the character relationship candidate sentences that can be matched by the rule are already extracted in the first step, there is no need to build a syntactic tree, so there is some improvement in time consumption compared to the method based on syntactic tree only, but the time consumption is also much more than that of the rule-based method.

10.3. Comparison with Other Methods. The final results of the paper were averaged from the experimental results of the different data sets and compared with other methods in the literature compared with other methods in the literature, as shown in Table 5. The literature [53] used a feature extraction algorithm based on character annotation. The literature [54] classified character relationships into six categories, selected character pairs of contextual features, distance features, and syntactic features as feature vectors, and finally used support vector machine classification methods to identify the relationships. The literature [55] used a convolutional tree kernel function to extract character relationships. Comparison with other methods shows that the method proposed in the thesis and the method based on convolutional tree kernel have no obvious advantage in terms of accuracy, but they all have significant improvement in terms of recall compared with the other three methods, and finally, the comprehensive evaluation index F-value is higher than the remaining three methods. Therefore, the character relationship extraction algorithm based on the combination of rule matching and syntactic trees proposed in the thesis can improve the relationship extraction effect to a certain extent. The time element is properly used.

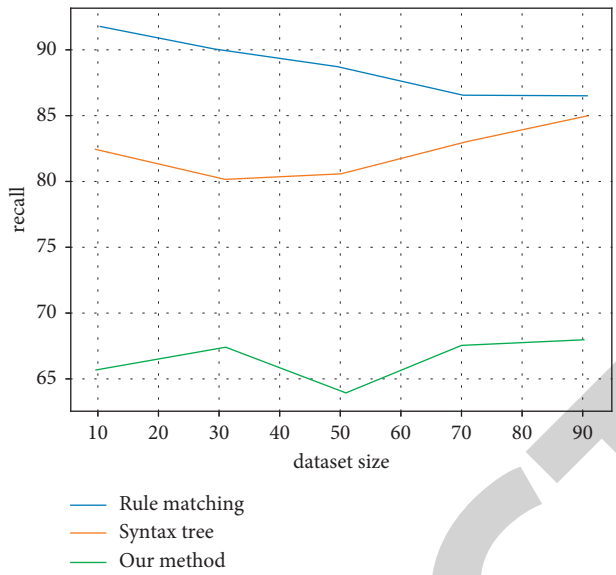

Figure 8: Comparison of the recall rates of the three methods on different datasets.

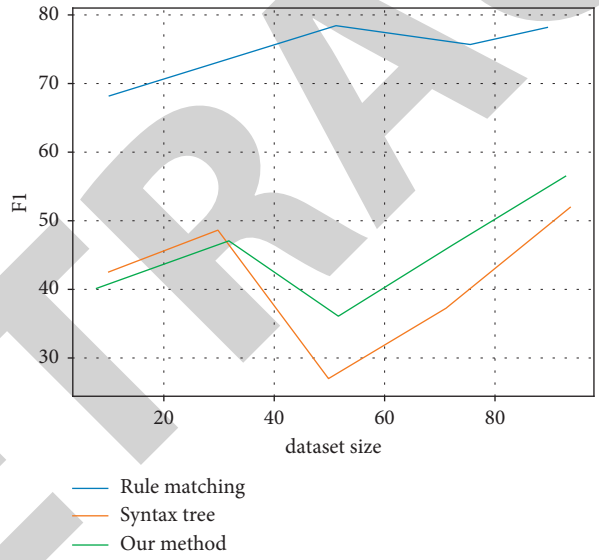

Figure 9: Comparison of the F-values of the three methods on different data sets.

Wuthering Heights condenses all the elements of a rich and rigorous scene, and author Emily presents a sophisticated and detailed chronology of text time and story time in a clever way, unfolding the plot in a staggering reversal of time and highlighting Table 5 the theme in a complex interweaving of chronology. The descriptions of the weather and seasons bring the emotions and actions of the novel's characters to life, making the scenes more vivid and dramatic, and greatly enhancing the lively nature and mystery of this thrilling and original novel.

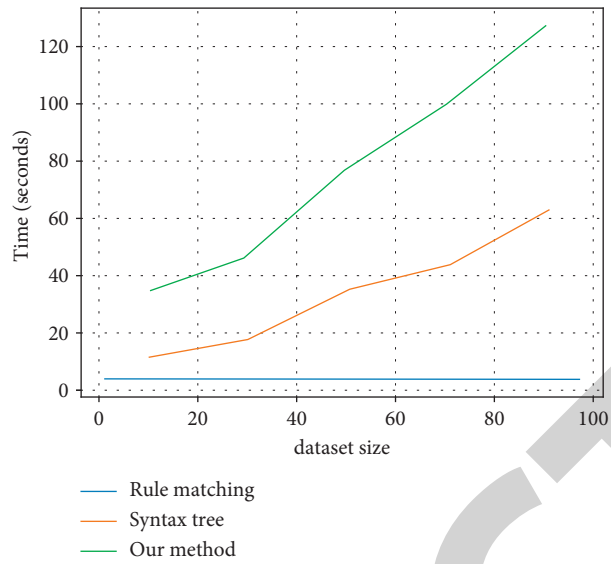

Figure 10: Comparison of the temporal performance of the three methods on different datasets.

Table 5: Comparison of experimental results with other literature.

Method used	Accuracy (%)	Recall (%)	F value (%)
Rule matching + syntax tree	81.44	82.39	81.87
Semantic role annotation	81.17	81	81.03
Based on SVM	60.8	61.82	61.33
Based on convolution tree kernel	85.8	71.1	61.33

11. Conclusions

This paper mainly describes the algorithm of person relationship extraction based on the combination of rule matching and syntactic tree, which is divided into four parts: the first part is data preprocessing, as the basic preparation part, first, it describes the use of ICTCALS for word separation and person name recognition and gives the relevant definitions to be used subsequently; it describes the person relationship extraction based on rule matching, including the establishment of rule base, regular expression. The algorithm of character relationship extraction based on the syntactic tree is proposed. First, the syntactic tree is built using Stanford Parser, second, the syntactic tree is transformed into an SPT tree by pruning out the non-joint points, and then the character relationships are extracted according to the SPT tree.

Data Availability

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding this work.

References

[1] S. Ho. Han, "A review of research on restaurant brand personality: a focus on the hospitality and tourism journals listed at korea research foundation," *Journal of Tourism Sciences*, vol. 35, no. 2, pp. 337–353, 2011.

[2] C. H. Cao, Y. N. Tang, and D. Y. Huang, W. Gan, C. Zhang, IIBE: An Improved Identity-Based Encryption Algorithm for WSN Security," *Security and Communication Networks*, pp. 1–8, 2021.

[3] S. Sohangir and D. Wang, "Improved sqrt-cosine similarity measurement," *Journal of Big Data*, vol. 4, no. 1, pp. 1–13, 2017.

[4] D. Wu, C. Zhang, L. Ji, R. Ran, H. Wu, and Y. Xu, "Forest fire recognition based on feature extraction from multi-view images," *Traitement du Signal*, vol. 38, no. 3, pp. 775–783, 2021.

[5] L. Wang, C. Zhang, Q. Chen et al., "A Communication Strategy of Proactive Nodes Based on Loop Theorem in Wireless Sensor Networks," in *Proceedings of the 2018 Ninth International Conference on Intelligent Control and Information Processing (ICICIP)*, pp. 160–167, IEEE, Wanzhou, China, November 2018.

[6] T. Tsao, "Postcolonial life and death: a process-based comparison of emily bronte's wuthering Heights and ayu utami's saman," *Comparative Literature*, vol. 66, no. 1, pp. 95–112, 2014.

[7] J.-M. Chen, M.-C. Chen, and Y. S. Sun, "A novel approach for enhancing student reading comprehension and assisting teacher assessment of literacy," *Computers & Education*, vol. 55, no. 3, pp. 1367–1382, 2010.

[8] I. Defant, "Inhabiting nature in emily Brontë's wuthering Heights," *Brontë Studies*, vol. 42, no. 1, pp. 37–47, 2017.

[9] N. F. Newman, "Workers, gentlemen and landowners: identifying social class inThe ProfessorandWuthering Heights," *Brontë Society Transactions*, vol. 26, no. 1, pp. 10–18, 2001.

[10] B. A. O. Xiaoli, "Paradoxes concerning the love in wuthering Heights," *Cross-Cultural Communication*, vol. 11, no. 6, pp. 89–93, 2015.

[11] Z.-wan Zhang, Di Wu, and C.-jiong Zhang, "Study of cellular traffic prediction based on multi-channel sparse LSTM," *Computer Science*, vol. 48, no. 6, pp. 296–300, 2021.
[12] F. S. Basirizadeh, M. Soqandi, N. Raoufzadeh, N. Zarei, and A. Adisaputera, "A study of wuthering Heights from the perspective of eco-criticism," *Budapest International Research and Critics in Linguistics and Education (BirLE) Journal*, vol. 3, no. 4, pp. 1623–1633, 2020.
[13] C. Zhang, X. Wu, Z. Niu, and W. Ding, "Authorship identification from unstructured texts," *Knowledge-Based Systems*, vol. 66, pp. 99–111, 2014.
[14] C. Van Der Meer, "Interrogating Brontë sequels: anna L'estrange,The return to wuthering Heights," *Brontë Studies*, vol. 30, no. 1, pp. 41–52, 2005.
[15] S. Vellay, N. Miller Latimer, and G. Paillard, "Interactive text mining with Pipeline Pilot: a bibliographic web-based tool for PubMed," *Infectious Disorders - Drug Targets*, vol. 9, no. 3, pp. 366–374, 2009.
[16] P. An, Z. Wang, and C. Zhang, "Ensemble unsupervised autoencoders and Gaussian mixture model for cyberattack detection," *Information Processing & Management*, vol. 59, no. 2, Article ID 102844, 2022.
[17] N. Collier, "Uncovering text mining: a survey of current work on web-based epidemic intelligence," *Global Public Health*, vol. 7, no. 7, pp. 731–749, 2012.
[18] T. Atanasova, M. Kasheva, S. Sulova, and J. Vasilev, "Analysis of the possible application of data mining, text mining and web mining in business intelligent systems," in *Proceedings of the In The 33rd International Convention MIPRO*, pp. 1294–1297, IEEE, Opatija, Croatia, May 2010.
[19] W. Jicheng, H. Yuan, W. Gangshan, and Z. Fuyan, "Web mining: knowledge discovery on the Web," in *Proceedings of the IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 99CH37028)*, pp. 137–141, Tokyo, Japan, October 1999.
[20] F. S. Gharehchopogh and Z. A. Khalifelu, "Analysis and evaluation of unstructured data: text mining versus natural language processing," in *Proceedings of the In 2011 5th International Conference on Application of Information and Communication Technologies (AICT)*, pp. 1–4, IEEE, Baku, Azerbaijan, October 2011.
[21] X. Lin, J. Wu, S. Mumtaz, S. Garg, J. Li, and M. Guizani, "Blockchain-based on-demand computing resource trading in IoV-assisted smart city," *IEEE Transactions on Emerging Topics in Computing*, vol. 9, no. 3, pp. 1373–1385, 2021.
[22] J. Li, Z. Zhou, J. Wu et al., "Decentralized on-demand energy supply for blockchain in internet of things: a microgrids approach," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 6, pp. 1395–1406, 2019.
[23] Z. Zhengwan, Z. Chunjiong, L. I. Hongbing, and X. I. E. Tao, "Multipath transmission selection algorithm based on immune connectivity model," *Journal of Computer Applications*, vol. 40, no. 12, p. 3571, 2020.
[24] W. Duan, J. Gu, M. Wen, G. Zhang, Y. Ji, and S. Mumtaz, "Emerging Technologies for 5G-IoV Networks: Applications, Trends and Opportunities," *in IEEE Network*, vol. 34, pp. 283–289, 2020.
[25] R. Nikhil, N. Tikoo, S. Kurle, H. S. Pisupati, and G. R. Prasad, "A survey on text mining and sentiment analysis for unstructured web data," *Journal of Emerging Technologies and Innovative Research*, vol. 2, no. 4, pp. 1292–1296, 2015.