IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 16, 2018, accepted November 10, 2018, date of publication November 21, 2018,
date of current version December 19, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2882624

Integrating Deep Learning Approaches for Identifying News Reprint Relation

YIN LUO[1], FANGFANG WANG[2,3], JUN CHEN[4], LEI WANG[3,5], AND DANIEL DAJUN ZENG[3,6,7], (Fellow, IEEE)

[1]School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China
[2]Beijing Wenge Technology Co., Ltd., Beijing 100080, China
[3]The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[4]Communication Technology Bureau, Xinhua News Agency, Beijing 100803, China
[5]The State Information Center, Beijing 100045, China
[6]School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100049, China
[7]Department of Management Information Systems, The University of Arizona, Tucson, AZ 85721, USA

Corresponding author: Fangfang Wang (fangfang.wang@ia.ac.cn)

This work was supported by the National Key R&D Program of China under Grant 2016QY02D0305, in part by the National Natural Science Foundation of China under Grant 61671450 and Grant 71621002, and in part by the Key Research Program of the Chinese Academy of Sciences under Grant ZDRW-XH-2017-3.

ABSTRACT With the rapid development of big data and new media technologies, a large amount of original news is generated and reprinted on the Internet via news portals. Identifying news reprint relations is of great importance for the analysis of news diffusion patterns and copyright protection. However, the amount of news data on the Internet creates a huge challenge for efficiently identifying news reprint relation. Some existing studies focus on computing the similarity of the full text of news reports, which is not always effective, because some reprints only excerpt some sentences of the original news reports. The core challenge of improving identification accuracy is excavating the potential semantic relevance between news articles at the sentence level. Inspired by deep learning and semantic-based text representation models, this paper proposes an approach for identifying news reprint relation by integrating deep learning approaches. First, news reports that are not related to the topic of the original news report are removed via topic correlation mining. Then, the potential semantic relevance is excavated at the sentence level through the integration of semantic analysis methods, and reprint relations are identified between news reports. The performance of the approach is empirically evaluated using a real-world dataset. Experimental results show that the semantic analysis model integration allows us to mine in-depth semantic associations between news stories and accurately identify news reprint relations. These results benefit news diffusion pattern analysis and copyright protection.

INDEX TERMS Deep learning, diffusion pattern, news reprint relation identification, semantic relevance, word embedding.

I. INTRODUCTION

The rapid development of big data and new media technologies has prompted revolutionary changes to media distribution channels and news forms. Much high-quality original news is produced on the Internet. Meanwhile, by virtue of its interactivity, high efficiency and low cost, many online news portals publish news stories that are reprints of original news stories generated by other sources, which results in the wide and rapid diffusion of the news. By identifying news reprinting relations and constructing the reprinting network of news portals, news diffusion patterns can be mined, which could have important applications in policy-making, crisis management and band imaging [1], [2]. Moreover, many media outlets reprint original news stories but do not indicate the original source, which diverts news readers from the original news sites to the reprinted news sites and affects the initiative of the originators. This behavior is not conductive to healthy news diffusion patterns. Therefore, it is necessary to identify reprint relations between news items to improve service for decision-makers and protect the copyright of the originator.

However, identifying news reprint relations is challenging in real-world scenarios because the reprint relations would be identified by comparing original news items with all the news published on the Internet. However, the number of news items on the Internet is huge. Additionally, there are various forms of news item reprints. Generally, news sources should reprint

an original news item by indicating the source or reference of the original news item by retaining electronic headings, specifying the media report (e.g., ''according to a certain media report'') or providing a dedicated news source field on the webpage. However, Internet news media formats are inconsistent, and many news sources reprint original news articles but do not indicate the original news source or even misprint its content, which also creates a great challenge for accurate reprint relation identification [3].

Existing studies have proposed approaches for identifying reprint relations among news stories; however, they have limitations. While some researchers extract the links between news portals using the source acknowledgement of news items, which does not work on reprinted news stories that do not indicate the source of the original story [1], [2], other researchers focus on computing the similarity between the full text of news items [3], [4], which is not always effective because some reprints only excerpt some sentences of the original news reports. Thus, the key technical challenge of improving the accuracy of identifying news reprint relation is excavating the potential semantic relevance between the original news story and the candidate news story at the sentence level by integrating semantic analysis models. Recently, many semantic-based text representation models, such as the topic model [5], Word2Vec [6], doc2vec [7] and fastText [8], have been proposed. These models consider the potential semantic information of text and represent the semantic features of text in low-dimension distributed vectors. Therefore, in our approach, we apply semantic-based text representation models to deeply excavate the potential semantic relevance between sentences found in news.

In this paper, we focus on the task of identifying reprint relations among large-scale news outlets based on news content and propose an approach for identifying news reprint relation by integrating deep learning approaches. First, the similarity between the full texts of news stories is computed using a TFIDF model to find the topic correlations; only news stories with topics related to the topic of the original news story will be retained. Then, the potential semantic relevance between the original news story and the candidate news story is deeply excavated at the sentence level using a Word2Vec model. Finally, reprint relations can be identified. Experiments using a real-world news dataset were conducted to empirically evaluate the performance of our proposed approach. The results show that the approach can efficiently identify reprint relations among large-scale news sources, and excavating the potential semantic relevance between news stories at the sentence level allows us to accurately identify reprint relations. With this accurate identification, a news dissemination network can be conveniently and accurately constructed, which benefits domains such as news influence estimation, public hotspot discovery and policy-making.

Our contributions are summarized as follows:

- This work is a first step of integrating semantic analysis models to mine the semantic relevance of news items and identify news reprint relations regarding large-scale news sources.
- We propose an approach for identifying news reprint relation by integrating deep learning approaches. It enables efficient and accurate news reprint relation identification by roughly determining topic correlations and deeply excavating potential semantic relevance at the sentence level.
- We demonstrate the efficacy of the approach using a real-world dataset that we collected from 3183 online news portals and conduct several case studies to explain how the results of news reprint relation identification can be used in real-world scenarios.

The remainder of this paper is organized as follows. Section 2 summarizes the related work of previous research. Section 3 provides the problem formulation and elaborates our proposed framework and model in detail. Experiments and evaluations are presented in Section 4. Section 5 concludes this research, and Section 6 discusses future work.

II. RELATED WORK

This section investigates the literature from three research fields: (1) news reprint relation identification, (2) text similarity computation and (3) text representation.

A. NEWS REPRINT RELATION IDENTIFICATION

In previous research, reprint relation identification between news stories has been applied in news influence estimation, online public opinion mining and decision-making dissemination. Driven by these applications, several approaches for identifying reprint relationships have been proposed. Wang *et al*. [1], [2] automatically identified the links between news portals using the source acknowledgement of news and constructed the reprinting network based on identified links; however, this method does not work on unregulated reprint relation identification. Yang *et al*. [4] proposed a co-occurrence word-based approach that constructs a word set of news items using high-frequency entities and keywords that are extracted from the content; if the number of co-occurrence words between two articles is greater than a threshold, the two articles are said to have a reprint relation. However, the accuracy of this approach is not high because high-frequency words are usually similar. Chen *et al*. [3] constructed a vector space model using the TFIDF method and computed the text similarity of two articles using common similarity computation methods, such as the Pearson correlation coefficient, cosine similarity and Euclidean similarity; however, these methods are not always effective because some reprints include only partial text or revised excerpts.

This paper aims to achieve accurate and efficient news reprint relation identification of large-scale news sources by roughly calculating news topic correlations using a TFIDF model and deeply excavating the potential semantic relevance at the sentence level using a deep learning and semantic analysis model.

B. TEXT SIMILARITY COMPUTATION

Accurately measuring the similarity of sentences plays a fundamental role in the process of identifying news reprint relation. The diversity of linguistic expressions creates a great challenge for text similarity computations. Most prior work on text similarity computation relied on feature extraction. According to their feature type, text similarity computation approaches can be divided into three categories [9]: (1) string-based approaches, (2) corpus-based approaches and (3) knowledge-based approaches.

String-based approaches employ common functions, such as the edit distance or hamming distance, to calculate similarity over string sequences extracted from a text, e.g., lemma, stem or n-gram sequences. Wan *et al*. [10] extracted sentence n-grams and measured the similarity of sentences using n-gram overlap. Madnani *et al*. [11] identified paraphrases based on machine translation metrics. String-based approaches are simple, but their efficacy is restricted because they ignore the semantic information of the text. Corpus-based approaches derive distributional vectors of words from a large corpus using distributional models (e.g., VSM [12], LSA [13] and LDA [5]) and then compute text similarity by summing the distributional vector of each word or using TFIDF to weight the vector of each word and then summing them. Then, the cosine similarity or Pearson coefficient of two vectors is used to measure the similarity of texts. Corpus-based approaches can retain the semantic information of texts; however, they are usually time consuming. Knowledge-based approaches compute text similarity with the aid of external resources, include ontology-based knowledge such as WordNet [14] and HowNet [15] and network-based knowledge such as Wikipedia [16] and Baidu Encyclopedia [17]. Knowledge-based approaches retain the semantic information of text well; however, they suffer from problems of knowledge completeness and algorithmic complexity.

Our proposed approach leverages text similarity computation strategies to efficiently and accurately identify news reprint relations. The topic-level model of the approach identifies news stories with topics similar to that of an original news story using a time-inexpensive TFIDF model. Then, the model excavates the potential semantic relevance between news items at the sentence level and identifies parts of candidate news item that are reprinted from original news item by representing sentences as low-dimensional distributional vectors using Word2Vec and computing the similarity of the original news item and candidate sentences.

C. TEXT REPRESENTAION

Text representation, which aims to numerically represent unstructured text as something mathematically computable, is a fundamental process of text mining tasks. Text representation approaches can be divided into three methods: (1) VSM-based representation approaches, (2) text enrichment representation via external knowledge incorporation and (3) text representation through the exploration of internal semantic relations in the corpus [18].

In VSM-based representation approaches, texts are mapped into a vector space, and each dimension of the vector space corresponds to a particular term and reflects the weight of the term in the document. The VSM paradigm is widely used due to its flexibility and effectiveness; however, its accuracy has been limited by the loss of adjacent words and semantic relations [19]. To overcome the limitations of VSM-based approaches, various approaches to incorporate word correlations and semantic information into text representation models have been proposed. One such method, text enrichment, usually leverages external knowledge for advanced VSM representation. External resources include ontology-based knowledge such as WordNet [20] and network-based knowledge such as Wikipedia [21]. Text enrichment representation approaches acquire the rich semantic relations between words; however, they are not flexible enough and are difficult to update in real time due to the rapidly changing availability of external knowledge [22]. Some approaches attempt to explore the information embodied in a large corpus and directly learn the vector representation of a text. Recently, word embedding techniques, such as Word2Vec [7] and GloVe [23], which encode the semantic properties of a word into a low-dimensional vector, have been successful in many natural language processing tasks.

In our approach, text representation is the key process of excavating the potential semantic relevance between news items at the sentence level. Due to its good performance in capturing synaptic and semantic information, we adopt a Word2Vec model to represent news content. We first learn the dense vectors of words in a large corpus using Skip-gram model and then apply pre-trained vectors of words to represent the news.

III. INTEGRATING DEEP LEARNING APPROACHES FOR IDENTIFYING NEWS REPRINT RELATION

In this section, we first introduce how the data from the web have been acquired in detail. Then we formulate the problem of news reprint relation identification and provide a detailed introduction of our proposed approach for identifying news reprint relation, which consists of two models, i.e., a topic-level model and a sentence-level model.

A. DATA COLLECTION

This paper studies news reprint relation identification approach. We crawled more than 10 million news articles from 3183 online news portals, from which news information could be crawled according to the Robot Exclusion Protocol, on a daily basis from January 1st, 2018 to June 30, 2018 and selected 30 popular original news items in the field of finance, sports and technology from the news articles. Then we separately constructed a keyword set for each original news. 25899 news items that were related to the original news item by topic, as determined using keywords matching, were chosen as candidate news items.

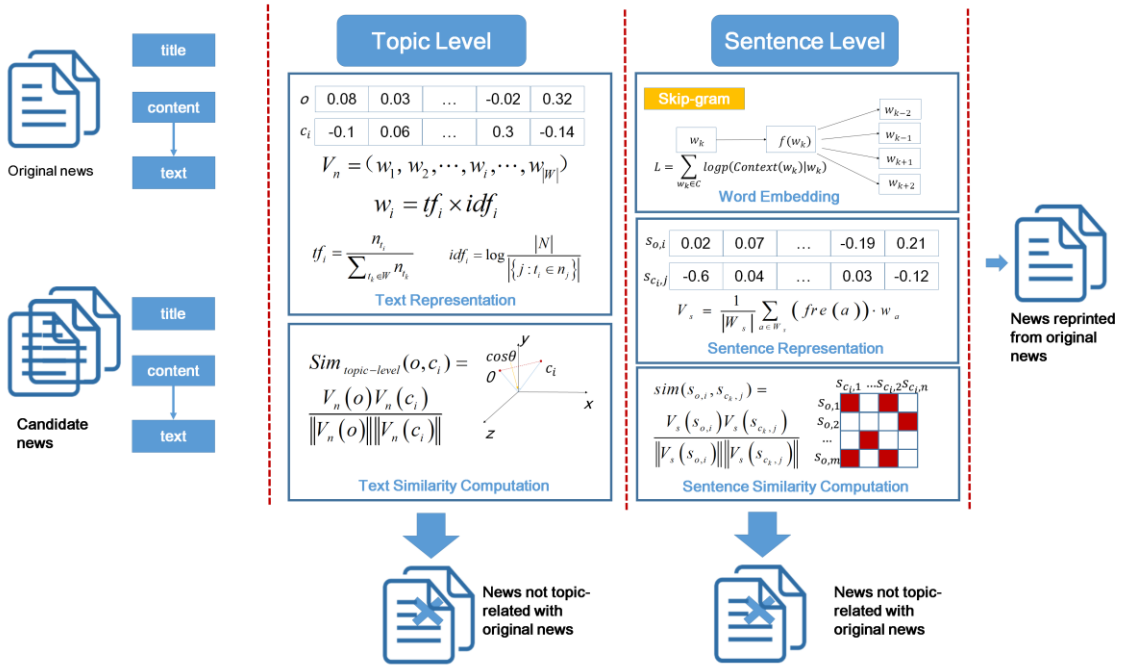


FIGURE 1. System architecture of our proposed approach.

B. PROBLEM FORMULATION

The problem of news reprint relation identification can be formally described as follows. Given the original news item and some candidate news items from the Internet, our goal is to identify which candidate news items reprinted the original news item.

Here, we define some notations that will be used throughout this paper. Let o denote the original news item and $C = \{c_1, c_2, \ldots, C_{|C|}\}$ denote the set of candidate news items. Our goal is to identify whether candidate news item c_i reprints original news item o.

C. SYSTEM ARCHITECTURE OF THE PROPOSED APPROACH

The overview of our proposed approach is shown in Fig. (1); it consists of a topic-level model and a sentence-level model. In the topic-level model, we compute the similarity of the full text of the original news item and candidate news items to identify articles that are related with the original news by topic with low time consumption. In the sentence-level model, we first learn the dense vectors of words on a large corpus using Skip-gram model [6]. Then, we represent news sentences by summing the frequency weighted pre-trained vectors of each word in the sentences and compute the similarity of sentences from the original news item and candidate news items using cosine similarity. Based on the similarity results, systems could accurately and efficiently identify news reprint relations. Next, we describe the models in detail.

1) TOPIC-LEVEL MODEL

Considering the huge amount of candidate news items, we design the topic-level model to remove news items that are not related to the original news item by topic to improve the efficiency of identifying news reprint relation. The topic-level model measures the topic correlation of the original news item and candidate news items by computing the similarity of the full text of the news items using TFIDF.

For specific news item n, we represent the news content as follows:

$$V_n = (w_1, w_2, \cdots, w_i, \cdots, w_{|W|}), \tag{1}$$

where W is the set of words of a news corpus and w_i is the weight of the word term t_i in news item n. w_i is calculated as follows:

$$w_i = tf_i \times idf_i, \tag{2}$$

where tf_i measures the frequency of term t_i that occurs in the content of news item n and idf_i measures the amount of information the word provides, i.e., if the word is common or rare across all news content. tf_i and idf_i are calculated as follows:

$$tf_i = \frac{n_{t_i}}{\sum_{t_k \in W} n_{t_k}}, \tag{3}$$

$$idf_i = \log \frac{|N|}{|\{j : t_i \in n_j\}|}, \tag{4}$$

where n_{ti} is the number of times term t_i appears in the content of news item n, n_{tk} is the number of times term t_k appears in the content of news item n, N is the set of the news corpus; and $\{j{:}t_i \in n_j\}$ is the set of news items that contains term t_i.

After representing the original news item and the candidate news items, the topic correlation between the two is measured as follows:

$$Sim_{topic-level}(o, c_i) = \frac{V_n(o)\, V_n(c_i)}{\|V_n(o)\|\, \|V_n(c_i)\|}, \tag{5}$$

where $V_n(o)$ is the vector representation of original news item o and $V_n(c_i)$ is the vector representation of candidate news item c_i. If $Sim_{topic-level}(o, c_i) < \alpha$, the topic of candidate news item c_i is not related to that of original news item o. Thus, candidate news item c_i will be removed by the system. Candidate news items that are related to the original news item by topic will be input into the sentence-level model to identify the reprint relation.

2) SENTENCE-LEVEL MODEL

In real-world scenarios, news portals reprint an original news item by using either the full text or excerpting parts of the original news item. The key point of news reprint relation identification is mining the potential semantic relevance between news items at the sentence level. Many studies have shown that a low-dimensional vector of text can generate better semantic representations and improve text similarity computation results. Hence, we integrate semantic-based text representation models in the sentence-level model.

In the sentence-level model, we first learn the semantic vector representation of words on a large corpus using Skip-gram model [6] and then apply the dense pre-trained vectors of words to represent the sentences of the original news item and candidate news items by summing the frequency weighted vector of each word that appears in the sentence. News sentences are represented as follows:

$$V_s = \frac{1}{|W_s|} \sum_{a \in W_s} (fre\,(a)) \cdot w_a, \tag{6}$$

where W_s is the set of words in a sentence, w_a is the K-dimensional vector of word a, $fre(a)$ is the word weight determined according to the frequency that the word appears in the sentence and V_s is the vector representation of the sentence.

The similarity between original news sentences and candidate news sentences is calculated as follows:

$$sim(s_{o,i}, s_{c_k,j}) = \frac{V_s\left(s_{o,i}\right) V_s\left(s_{c_k,j}\right)}{\left\|V_s\left(s_{o,i}\right)\right\| \left\|V_s\left(s_{c_k,j}\right)\right\|} \tag{7}$$

where $V_s(s_{o,i})$ is the vector representation of the ith sentence of original news item o and $V_s\left(s_{c_k,j}\right)$ is the vector representation of the jth sentence of candidate news item c_k.

After calculating the similarity between sentences of original news item o and candidate news item c_k, a sentence similarity matrix $R = \{r_{i,j}\}_m \times_n$ is constructed, where m is the number of sentences from original news item o, n is the number of sentences from candidate news item ck and $r_{i,j} = 1$ if sim$sim(s_{o,i}, s_{c_k,j}) > \delta$.

Based on the sentence similarity matrix R, we determine whether candidate news item c_k reprinted original news item o. Candidate news item c_k reprinted original news item o if such a sequence $L = \{l_{i,j,1}, l_{i+1,j+1,2}, \ldots, l_{i+|L|-1,j+|L|-1,|L|}\}$ exists in matrix R that satisfies the following conditions:

(1) $\forall l_{i,j,k}, r_{i,j} = 1$,
(2) $\gamma_1 m \leq |L| \leq \min(m, n)$, and
(3) $\gamma_1 m \leq |L| \leq \min(m, n)$,

where i denotes the raw coordinate in the text similarity matrix R, j denotes the column coordinate in the text similarity matrix R, k denotes the coordinate of element $l_{i,j,k}$ in L and $|L|$ is the length of sequence L.

IV. EXPERIMENTATION

In this section, we first present the experimental dataset and evaluation metrics. Then, comprehensive and systematic experiments are conducted on the dataset to verify the efficacy of our proposed approach. Finally, case studies are presented to show how the identification results can be applied in real scenarios.

A. DATASET

The experiment is conducted using the real-world dataset collected from more than 3000 news portals. We invited 3 annotators to manually label the reprint relations between original news and its candidate news. If the candidate news reprints the original news, the reprint relation will be labelled as 1, otherwise the reprint relation will be labelled as 0. The annotators, who worked independently and were not aware of one another, achieved a Fleiss' kappa of 0.89, which indicates that the annotated results of three annotators have high consistency. We take the annotated results as the golden evaluation criteria to evaluate the proposed new reprint relation identification approach quantitatively. The dataset is summarized in Table (1).

TABLE 1. Detailed dataset information.

Statistics	Values
# of original news	30
# of candidate news	25899
# of reprinted news (no source label)	4234 (537)

B. EVALUATION METRICS

Precision, Recall and F-Measure are three popular evaluation metrics that have been widely used in information retrieval systems, classification, and information identification. In this paper, we apply these metrics to evaluate our proposed method.

For original news item o, Precision and Recall are calculated as follows:

$$Precision = \frac{|R\,(o) \cap I\,(o)|}{|I\,(o)|}, \tag{8}$$

$$Recall = \frac{|R\,(o) \cap I\,(o)|}{|R\,(o)|}, \tag{9}$$

where $R\,(o)$ is the set of news items that have been manually labeled as reprinted news of original news item o, and $I\,(o)$ is the set of news items identified as the reprinted news of original news item o by our proposed method.

F-Measure is the weighted harmonic mean of Precision and Recall; it is calculated as follows:

$$F - Measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}, \tag{10}$$

After calculating the Precision, Recall and F-Measure of each original news item, we average them as the final assessment result.

C. EXPERIMENTAL SETTINGS

We applied the grid search method to select a reasonable value for each parameter of our proposed method. The parameters are explained and set to follows.

Parameter α: As the layer to filter candidate news items, we should set a low similarity threshold in the topic-level model to ensure a high recall rate. The similarity threshold α between news content is set to 0.5.

Parameter δ: In the sentence-level model, the similarity δ between sentences is set to 0.9.

Parameters γ_1 and γ_2: In the sentence-level model, these parameters, which are used to adjust the minimum length limitation of sequence L, are set to 0.6.

Parameter K: In the sentence-level model, we set the latent space dimension K to 200. The number of iterations is set to 20. The context window is set to 8. We use the Chinese dataset SogouCA as the corpus to learn the dense vector of words [24].

D. EXPERIMENTAL ANALYSIS

Our goal is to accurately and effectively identify reprint relations among large-scale news sources based on content information. We compare the performance of our proposed approach with that of baseline methods using the real-world dataset and the above evaluation metrics. According to existing work [3], [4], baseline methods that can be used to compute the similarity between the full text of original news items and the full text of candidate news items using text similarity computation methods include TFIDF [3], LDA [5], Simhash [25] and Word2Vec [6]. The similarity threshold for the full text of news items in the baseline methods is set to 0.9. Our proposed approach excavates the potential semantic relevance between original news items and candidate news items at the sentence level. Based on the results shown in Table (2), our proposed approach performs better than the baseline methods across all metrics, which indicates that excavating the potential semantic relevance between original news items and candidate news items at the sentence level can improve news reprint relation identification.

TABLE 2. Performance comparison results using the entire dataset.

Method	Precision	Recall	F-Measure
TFIDF	0.249	0.861	0.387
LDA	0.319	0.940	0.477
SimHash	0.732	0.868	0.794
Word2Vec	0.511	0.929	0.659
Our proposed approach	0.907	0.882	0.895

We also investigated sites that reprint original news items without citing the source, as this impacts the content creators. Thus, we compare the performance of our proposed approach against those of the baseline methods on news items without source information. The results in Table (3) show that our proposed approach is also more effective than the baseline approaches across all metrics when using this dataset.

TABLE 3. Performance comparison results using a dataset with implicit reprint.

Method	Precision	Recall	F-Measure
TFIDF	0.032	0.676	0.061
LDA	0.055	0.918	0.104
SimHash	0.232	0.758	0.355
Word2Vec	0.114	0.907	0.203
Our proposed approach	0.507	0.732	0.599

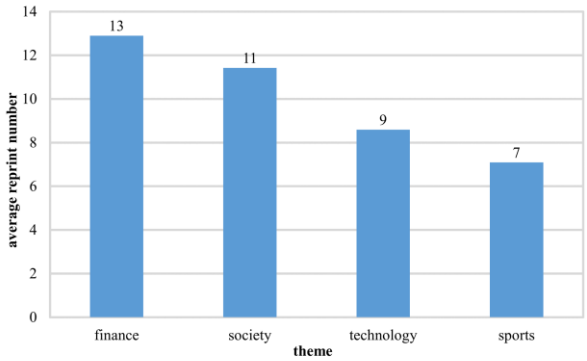


FIGURE 2. Average reprint number of original news items by theme.

E. CASE STUDIES

We conducted two case studies to explain how the results of news reprint relation identification can be used in real-world scenarios. The two case studies are: (1) news reprint relation identification for news diffusion pattern analysis, (2) news reprint relation identification for copyright protection.

1) NEWS REPRINT RELATION IDENTIFICATION FOR NEWS DIFFUSION PATTERN ANALYSIS

Online news reprint networks reveal how information spreads on the Internet. By utilizing our proposed news reprint relation identification approach, news reprint networks could be conveniently and accurately constructed. In this paper, we identify the reprint relations of 356 original news articles and construct a news reprint network. The 356 original news items are divided into four themes, i.e., technology, finance, sports and society, and the number of original news items in each theme is 75, 97, 91, and 93, respectively. We identified 3602 news reprint relations for the 356 original news items. Based on the reprint network, we analyzed the patterns of news diffusion from three perspectives: reprint number, reprint time and reprint medium.

a: ANALYSIS OF REPRINT NUMBER

The reprint number of the original news item indicates the scope of information spread. The average reprint number of the original news item is statistically analyzed for several themes, as shown in Fig. (2). The average reprint number of

original news items varied by themes: the average number of financial news reprints is the highest, followed by the average number of societal news reprints; the average numbers of technological news reprints and sports news reprints are less than that of financial news and societal news, which indicates that reprint number is associated with news theme. By observing data, we found that the reprint numbers of financial news and societal news are volatile and the reprint numbers of technological news and sports news are relatively stable.

b: ANALYSIS OF REPRINTING TIME

The reprinting time after publishing an original news item indicates the speed of news spread. The reprinting number of each hour after the original news item is published is counted, as shown in Fig. (3). Most (87.73%) reprinted news items are published within 36 hours of the publication of the original news item. The reprinting number increases sharply between hours 8 and 10. Because many original news items are published with only the date information, we set the hour of publishing as 00:00. Thus, high reprinting numbers in hours 8 and 10 after the publication of the original news item indicate that reprinted news is primarily published

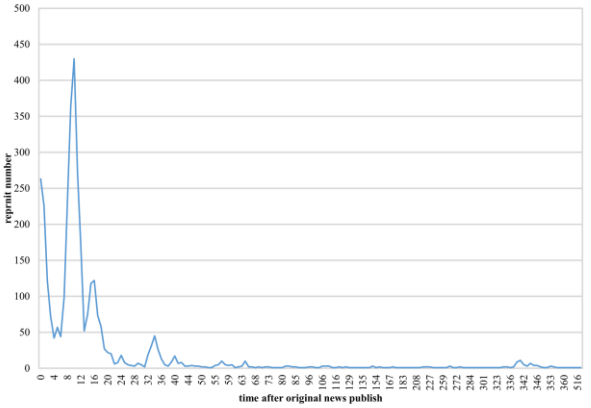


FIGURE 3. Reprinting number of each hour after the original news item is published.

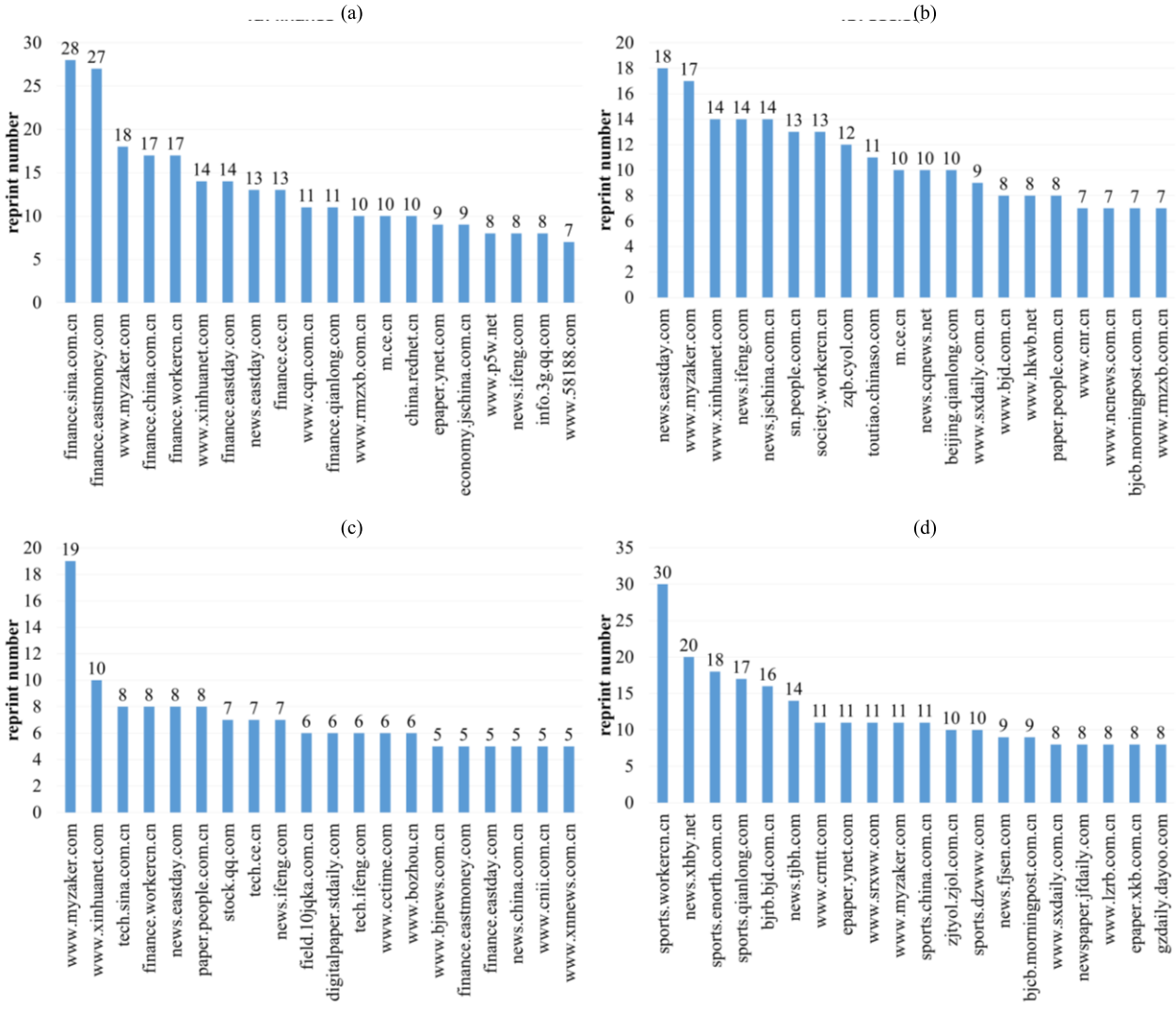


FIGURE 4. Numbers of original news items that media have reprinted by theme. (a) Finance. (b) Society. (c) Technology. (d) Sports.

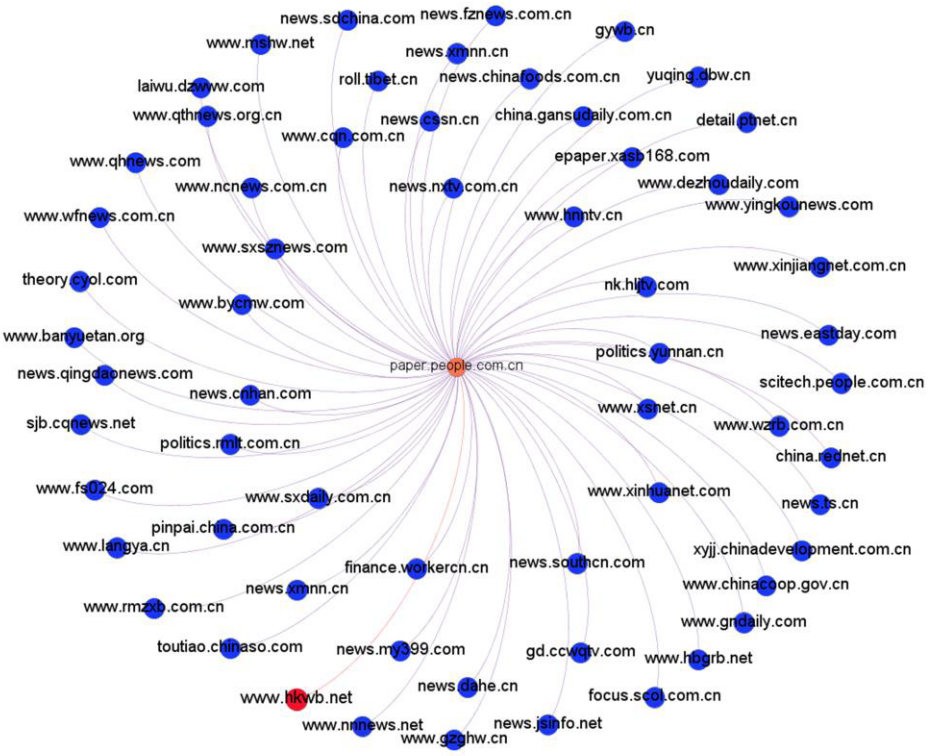


FIGURE 5. The diffusion network of an original news item.

between 8 and 10 a.m. on the day that the original news item is published.

c: ANALYSIS OF REPRINTING MEDIUM

Reprinting medium plays an important role in news information spread. We statistically analyze the numbers of original news items that each medium reprinted for several themes. The top 20 media are presented in Fig. (4). Some comprehensive news portals, such as myzaker.com, sina.cn, eastday.com and xinhua.net reprint the most original news items on all themes. Besides, it can be seen that 11 news media from the top 20 new media on finance theme are finance –related media, which indicates that news on one themes is mostly to be reprinted by the domain-related media.

2) NEWS REPRINT RELATION IDENTIFICATION FOR COPYRIGHT PROTECTION

News reprint relation identification can be applied to protect copyrights and contribute to the healthy development of the news industry. Fig. (5) represents the reprinting network of original news item, where the center node (in orange) is the news medium of an original news item[1]; the other nodes are the media that reprinted the original news item on different channels, and the nodes in red are the media that published news items without providing the source information of the original news. The original news items was reprinted by 62 news media. Reprinted news items from ''hkwb.net'' (in red) do not indicate the source of the original news. The URL of the original news items could be tagged on the webpage that the reprinted news item is published on, which would be beneficial for regulating the news industry and protecting the rights of original sites and authors.

[1]http://paper.people.com.cn/rmrb/html/2018-06/26/nw.D110000renmrb_20180626_6-09.htm

V. CONCLUSION

In this paper, we proposed an approach for identifying news reprint relation by integrating deep learning approaches. First, the full text similarity is computed to find topical correlations between news items and news items that are not related to the original news by topic are removed. Then, the potential semantic relevance between candidate news items and the original news item is excavated at the sentence level by integrating semantic analysis methods and the reprint relations between news items are identified. Extensive experiments on a real-world news dataset show the good performance of our proposed approach. This work is the first step towards using semantic analysis models to excavate the potential semantic relevance between news items at the sentence level and perform news reprint relation identification. The accurate and effective identification of reprint relations among large-scale news sources will benefit areas such as news diffusion pattern analysis, news copyright protection and news influence estimation. Using our proposed approach, we conducted a comprehensive analysis of news diffusion patterns and found that the reprinting number of original news items is associated

with theme, most reprint news are published within 36 hours of the publication of the original news item and news on different themes are all usually reprinted by some specific news portals. Mined knowledge would further contribute to the design of news reprint relation identification and news reprinting prediction models in the future.

VI. LIMITATIONS AND FUTURE RESEARCH

Our experimental dataset consists of only 30 original news items. We collected the candidate news items using keywords related to the original news topic and assumed that the candidate news items covered the reprinted news items.

In future research, we would like to expand the number of original news items to construct a more comprehensive dataset. We designed the news reprint relation identification approach based on news content; however, some news items reprint the pictures, videos, etc. of the original news item. In the future, such information could be integrated to improve the accuracy of reprint relation identification.

ACKNOWLEDGMENT

This work was supported by the National Key R&D Program of China under Grant No. 2016QY02D0305, the National Natural Science Foundation of China under Grant Nos. 61671450 and, 71621002, and the Key Research Program of the Chinese Academy of Sciences under Grant No. ZDRW-XH-2017-3.

REFERENCES

[1] Y. Wang, D. Zeng, X. Zheng, and F. Wang, "Propagation of online news: Dynamic patterns," in *Proc. IEEE Int. Conf. Intell. Secur. Inform.* Piscataway, NJ, USA: IEEE Press, Jun. 2009, pp. 257–259.

[2] Y. Wang, D. Zeng, B. Zhu, X. Zheng, and F. Wang, "Patterns of news dissemination through online news media: A case study in China," *Inf. Syst. Frontiers*, vol. 16, no. 4, pp. 557–570, 2014.

[3] X. Chen, J. Chen, and Y. Wang, "News reprint and citation analysis for Internet data," *Sci. Technol. China's Mass Media*, vol. 11, pp. 89–91, 12, 2017, doi: 10.19483/j.cnki.11-4653/n.2017.11.029.

[4] W. Yang, R. Dai, and X. Cui, "Model for Internet news force evaluation based on information retrieval technologies," *J. Softw.*, vol. 20, no. 9, pp. 2397–2406, 2009.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[6] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 3111–3119.

[7] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.

[8] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2017, pp. 427–431.

[9] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," *Int. J. Comput. Appl.*, vol. 68, no. 13, pp. 13–18, 2013.

[10] S. Wan, M. Dras, R. Dale, and C. Paris, "Using dependency-based features to take the 'para-farce' out of paraphrase," in *Proc. Australas. Lang. Technol. Workshop*, 2006, pp. 131–138.

[11] N. Madnani, J. Tetreault, and M. Chodorow, "Re-examining machine translation metrics for paraphrase identification," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2012, pp. 182–190.

[12] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[13] T. K. Landauer and S. T. Dutnais, "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge," *Psychol. Rev.*, vol. 104, no. 2, pp. 211–240, 1997.

[14] M. Batet, D. Sánchez, and A. Valls, "An ontology-based measure to compute semantic similarity in biomedicine," *J. Biomed. Inform.*, vol. 44, no. 1, pp. 118–125, 2011.

[15] B. Ge, F. Li, S. Guo, and D. Tang, "Word's semantic similarity computation method based on hownet," *Appl. Res. Comput.*, vol. 27, no. 9, pp. 3329–3333, 2010.

[16] M. Strube and S. P. Ponzetto, "WikiRelate! Computing semantic relatedness using Wikipedia," in *Proc. Nat. Conf. Artif. Intell.*, 2006, pp. 1419–1424.

[17] K. Yin, H. Yin, Y. Yang, and Z. Jia, "Semantic similarity computation of Baidu encyclopedia entries based on SimRank," *J. Shandong Univ.*, vol. 44, no. 3, pp. 29–35, 2014.

[18] J. Bai, L. Li, D. Zeng, and Q. Li, "Associated activation-driven enrichment: Understanding implicit information from a cognitive perspective," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 12, pp. 2655–2668, Dec. 2017.

[19] B. S. Harish, D. S. Guru, and S. Manjunath, "Representation and classification of text documents: A brief review," *Int. J. Comput. Appl.*, vol. 8, no. 2, pp. 110–119, 2010.

[20] T. Wei, Y. Lu, H. Chang, Q. Zhou, and X. Bao, "A semantic approach for text clustering using WordNet and lexical chains," *Expert Syst. Appl.*, vol. 42, no. 4, pp. 2264–2275, 2015.

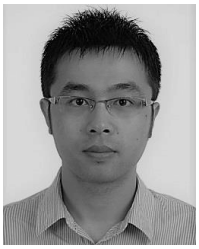
[21] H.-J. Kim, K.-J. Hong, and J. Y. Chang, "Semantically enriching text representation model for document clustering," in *Proc. ACM Symp. Appl. Comput.*, 2015, pp. 922–925.

[22] D. Zeng, H. Chen, R. Lusch, and S.-H. Li, "Social media analytics and intelligence," *IEEE Intell. Syst.*, vol. 25, no. 6, pp. 13–16, Nov./Dec. 2010.

[23] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

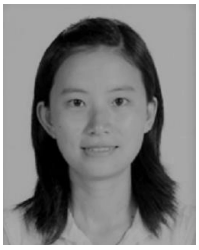
[24] C. Wang, M. Zhang, S. Ma, and L. Ru, "Automatic online news issue construction in Web environment," in *Proc. Int. Conf. World Wide Web*, 2008, pp. 457–466.

[25] G. S. Manku, A. Jain, and A. D. Sarma, "Detecting near-duplicates for Web crawling," in *Proc. Int. Conf. World Wide Web*, 2007, pp. 141–150.



YIN LUO received the B.S. degree in mathematics and applied mathematics from the Hefei University of Technology, Hefei, China, and the M.S. degree in software engineering from the University of Electronic Science and Technology of China, Chengdu, China. He is currently pursuing the Ph.D. degree with the Beijing Institute of Technology, Beijing, China. His research interests include machine learning, social computing, and information management.

FANGFANG WANG received the B.S. degree in information management and information system from Northeastern University, Shenyang, China, and the M.S. degree in management science and engineering from Tianjin University, Tianjin, China. She is currently an Assistant Engineer with the Institute of Automation, Chinese Academy of Sciences, China. Her research interests include machine learning and social computing.



JUN CHEN received the B.S. degree in mathematics from the Information Engineering University of the PLA, Zhengzhou, China, and the M.S. degree in computer science and technology from the Beijing University of Posts and Telecommunications, China. She is currently a Deputy Senior Engineer with the Communications Technology Bureau, Xinhua News Agency. Her research interests include journalism and communication, and social computing.

LEI WANG received the Ph.D. degree in management science and engineering from Tianjin University, Tianjin, China. He is currently the Chairman of Beijing Wenge Technology Co., Ltd., Beijing, China. He is also a Post-Doctoral Fellow with the State Information Center and an Associate Research Fellow with the Institute of Automation, Chinese Academy of Sciences, China. His research interests include machine learning and social computing.

DANIEL DAJUN ZENG (F'15) received the B.S. degree in economics and operations research from the University of Science and Technology of China, Hefei, China, and the M.S. and Ph.D. degrees in industrial administration from Carnegie Mellon University. He is currently a Gentile Family Professor with the Department of Management Information Systems, The University of Arizona. He also holds a research fellow position at the Institute of Automation, Chinese Academy of Sciences. His research interests include intelligence and security informatics, infectious disease informatics, social computing, recommender systems, software agents, and applied operations research and game theory. He has published more than 300 peer-reviewed articles. He currently serves as the Editor-in-Chief of *ACM Transactions on MIS* and the President of the IEEE Intelligent Transportation Systems Society.