Integrating Multisourced Texts in Online Business Intelligence Systems

Jianping Cao, Senzhang Wang, Benxian Li, Xiao Wang, *Member, IEEE*, Zhaoyun Ding,
and Fei-Yue Wang, *Fellow, IEEE*

***Abstract*—Online business intelligence systems often collect the texts from different sources, such as social media and news websites that can be heterogeneous in practice. These collections bring the difficulties of managing and organizing the comprehensive information hidden in different texts of the system. To more effectively organize the multisourced texts and help online users acquire wider knowledge, we propose a business intelligence system which integrates the multisourced texts from multisources. Regarding in many occasions, multisourced texts share some common contents with respect to the same topics. For example, a tweet and a news report may talk about the same event. Therefore, our goal is to correlate such texts of different sources with respect to the similar topics and get integrated more comprehensive information to facilitate other data mining tasks as well as online applications. To handle the problem, we propose a heterogeneous information network-based text aligning (HINTA) framework in this paper. HINTA applies meta-paths to calculate the text similarities, and constructs correlated pairs between the two types of texts. Next, HINTA first applies anchored pairs as bridges to combine the different types of texts. Finally, three different inference methods are employed to align the multisourced texts. Experimental results on real-world dataset show the effectiveness and efficiency of the framework in addressing the texts alignment problem.**

***Index Terms*—Integrate, intelligence system, multisourced, text.**

I. Introduction

WITH the rapidly development of social media and modern information technology, the construction of online business intelligence systems (BISs) that can fuse the information from multisources has been paid great attention by numerous researchers [1]–[3]. Modern BIS often contain the text information from various sources, and we have to clean, correlate, and mine the data, and to store and utilize the data for real world applications [4], [6], [13], [17]. One of the key tasks here is to integrate the information from various sources to get richer and cleaner information.

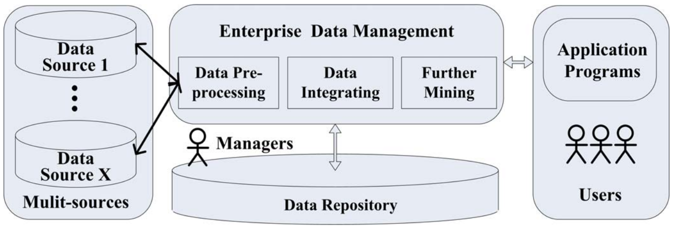


Fig. 1. Architecture of online BISs with multisourced data.

To do the task effectively and efficiently, one of the popular methods is to align the data from various sources first and then extract useful information collaboratively, that is correlating the data from multisources for further mining [10], [19], [21]. Since modern BIS often contain the multisourced text information that can be highly correlated to each other, researchers tend to combine them together through certain objects, words, and labels (e.g., ratings [5], [12]). The general architecture of online BIS is shown in Fig. 1 with four key parts: 1) the collection of multisourced data; 2) the management (or processing) of data; 3) the repository; and 4) application of the processed data. From the figure we can also observe that, data integration, as a widely used method in data processing, is a key step in the processing of multisourced data. Specifically, as a step after data collecting, the integration of multisourced data is a collaboratively processing of the data that can more effectively reduce the redundancy. For the next step of data storing, the integration of multisourced data can get more valuable information to reduce the load of storing. Moreover, in real-world applications the integrated data can be easily utilized for other data mining tasks, such as the detection of events in BIS. Generally, the integration of multisourced texts is critically important for BIS.

A. Motivation

Among the methods of text integration, the approach that correlates the multisourced texts with respect to certain topics

Manuscript received October 12, 2016; accepted May 8, 2017. This paper was recommended by Associate Editor A. Bargiela. *(Corresponding author: Jianping Cao.)*

J. Cao, B. Li, and Z. Ding are with the College of Information System and Management, National University of Defense Technology, Changsha 410073, China (e-mail: caojianping@nudt.edu.cn; libenxianxian@163.com; zyding@nudt.edu.cn).

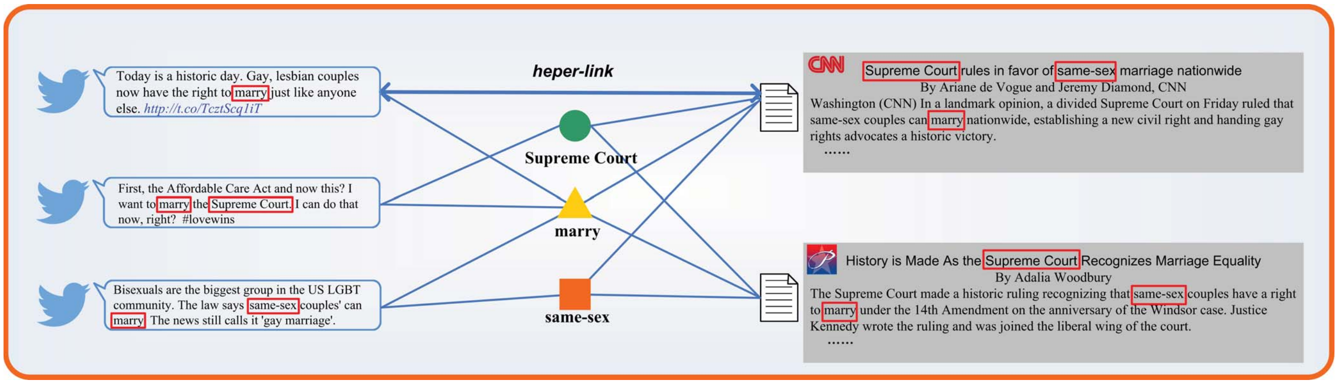
S. Wang is with the College of Computer Science, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China, and also with the Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 211106, China.

X. Wang is with the State Key Laboratory of Management and Control for Complex Systems, CAS, Beijing 100010, China.

F.-Y. Wang is with the College of Information System and Management, National University of Defense Technology Changsha, 410073 China, and also with the State Key Laboratory of Management and Control for Complex Systems, CAS, Beijing 100010, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2017.2710161



Today is a historic day. Gay, lesbian couples now have the right to marry just like anyone else. http://t.co/TcztScq1iT

First, the Affordable Care Act and now this? I want to marry the Supreme Court. I can do that now, right? #lovewins

Bisexuals are the biggest group in the US LGBT community. The law says same-sex couples' can marry. The news still calls it 'gay marriage'.

Supreme Court rules in favor of same-sex marriage nationwide
By Ariane de Vogue and Jeremy Diamond, CNN
Washington (CNN) In a landmark opinion, a divided Supreme Court on Friday ruled that same-sex couples can marry nationwide, establishing a new civil right and handing gay rights advocates a historic victory.
……

History is Made As the Supreme Court Recognizes Marriage Equality
By Adalia Woodbury
The Supreme Court made a historic ruling recognizing that same-sex couples have a right to marry under the 14th Amendment on the anniversary of the Windsor case. Justice Kennedy wrote the ruling and was joined the liberal wing of the court.
……

Fig. 2. Toy example of the HINTA framework. Different types of texts can be transformed into HINs, and then the correlations between them can be inferred by certain algorithms.

has rarely been studied and used for comparison to get better business intelligence [7]. Given the texts from different sources that are related to the same topic or event, it is very likely that these different texts discuss different aspects of the topic or event. The combination of different types of texts will help us obtain comprehensive information and get a better understanding on certain issues. Therefore, taking tweets and news articles as an example, we propose to construct an online BIS which can integrate the multisourced texts by aligning them with respect to certain topics. For instance, there are tweets and news articles related to the topic of #*Obamacare* and #*iphone*, respectively. The studied problem is how to match such texts from different sources. As a novel idea in the construction of online BIS, the integration of multisourced texts will facilitate the users, managers, and developers in the following aspects.

1) For the users of online BIS, the aligned multisourced texts can be read directly and used to find the more valuable information. For example, if comments from different sources on the competitive products are well aligned, the readers can find more valuable information quickly, and thus to improve the users' satisfactory.
2) For the managers of online BIS, the alignment of multisourced texts is another form of classification of texts and can better organize the information from various sources [22]. The aligned texts can be recognized as more important data and stored in the data warehouse for requiring and searching.
3) For the developers of online BIS, the aligned texts can be used directly as an indicator of certain issues. As the texts types can be different and become more complex, the well aligned texts are proved to be more influential and of valuable for developers to use.

B. Contribution and Outline

Although the integration of multisourced texts shows numerous advances, it is nontrivial to align different types of texts together. First, it is challenging to find a suitable method to represent the different types of texts that can also be employed for mining the correlation among them. The texts of news articles are usually formally published with explicit semantic meanings, while tweet texts are quite short, informal and fuzzy of semantic meanings. Second, since the contents of both tweets and news can change rapidly, it is hard for an unsupervised method to correlate them without labels. The hidden semantic meanings of different types of texts can be various, and thus traditional unsupervised method may connect the texts of similar features out-of-events, which may lead to undesirable results.

In this paper, we propose a framework of HINTA to address the above challenges. As shown Fig. 2, HINTA applies heterogeneous information networks (HINs) to represent the texts. HINs use the collections of different types of objects to represent texts, and they can also be easily applied for the calculation of text similarity in all the cases of text types. Second, HINTA uses the hyper-linked texts as "bridges" to connect tweets and news, and then infer the correlation relationships among the tweet and news texts. Different from the previous methods that focus on the interactions between nodes, we utilize semantic correlations for the computation of similarities across the two types of networks. Specifically, HINTA addresses the problem in three steps. First, it transfers the tweet and news texts to HINs with respect to semantic meanings, respectively. Then, it calculates the semantic similarities between the two HINs to obtain a few hyper-linked pairs for further correlations. Finally, it uses three matching algorithms [the extended meta-path-based similarity algorithm (MSA), the similarity flooding algorithm (SFA), and the Cross-Network Anchoring-Based Algorithm] to openly discuss the matching of the tweet and news pairs, and predict the relationships between two types of texts. Our contributions in this paper can be summarized as follows.

1) We study a novel problem of multisourced text alignment which could be used for the integration of text information from different sources in online BIS.
2) We propose a framework of HINTA to address the challenges of the proposed problem.
3) We conduct extensive experiments on three real-world datasets. The experimental results show that the proposed framework outperforms the baselines.

The rest of this paper is organized as follows. Section II will take a retrospect of the related works. In Section III, we will give a formulation of the proposed problem. In Section IV, we will introduce the framework of HINTA and the basic representation of texts. Section V will discuss the matching problem with three different algorithms. Then, Section VI shows the experimental results on three real-world datasets. Finally, Section VII concludes this paper and looks forward to future works.

II. Related Work

A. Text Representation

Text clustering is a widely studied problem in data mining. One of the key challenges is how to represent texts [8]. A typical method is using bag-of-words (BoWs) for representation [24], and the well-known approach TF-IDF is a classical approach based on it [16]. Other representations include *n*-grams (NG) model [20], [23], and vector-space models [16], etc. As for the representation of tweets, except the above methods, previous works usually applied existing networks to form HINs, e.g., the users' (authors') social network, retweet network, etc. [25], or simple term-tweet correlation network [26], or from the perspective of whole Twitter network level [34]. Since none of them can bridge different types of texts, these representations of tweets can not be directly applied in this paper.

B. Heterogeneous Information Networks

HINs are used to describe the multiattributes of nodes and edges of networks [8], [9]. In reality, it can be applied on texts [11]. The author-topic model proposed by Steyvers *et al.* [30] can be viewed as the early study of HINs applied in texts mining. Deng *et al.* [31] proposed a topic model with biased propagation algorithm to incorporate heterogeneous information network with topic modeling in a unified way. Many works have tried to combine the sufficient information on heterogeneous networks (e.g., the text and links) to detect communities, to analyze the evolution of networks and to model relational learning [35]–[37], [49]. In addition, some researchers studied the problem of information diffusion over networks [32], [33], [38], [39].

III. Preliminary

In this section, we will first provide the formal definitions of tweet, news, and the anchor links with respect to HINs, and then formally define the multisourced text alignment problem.

A. Terminology Definition

Definition 1 (Tweet/News): A tweet/news can be represented by such a heterogeneous information network $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where vertexes $\mathcal{V}$ contain objects with the following types, text (T), word ($\mathcal{O}_w$), named entities ($\mathcal{O}_e$), and other types like mention ($\mathcal{O}_m$) and hashtag ($\mathcal{O}_h$) (only for tweet); and edges $\mathcal{E}$ represent the multiple types of relations connecting the tweet objects. Note that to discern the two types of texts, we use superscript "(1)" and "(2)," e.g., $T^{(1)}$ and $T^{(2)}$, to represent tweet and news and corresponding objects, respectively.

Definition 2 (Anchor Link): Given a tweet $t_i^{(1)} \in T^{(1)}$ and a news article $t_j^{(2)} \in T^{(2)}$, if they are highly correlated with respect to semantic meanings, e.g., $t_i^{(1)}$ contains a hyper-link directing to $t_j^{(2)}$ and $t_i^{(1)}, t_j^{(2)}$ have a considerable objects in common. The link between the two texts is referred as an *anchor link*, denoted as $t_i^{(1)} \rightarrow t_j^{(2)}$, and all these pairs form an anchor link collection $\mathcal{R}$.

B. Problem Definition

The studied problem targets at detecting the semantically corresponding relationships between tweet $T^{(1)}$ and news articles $T^{(2)}$ with respect to certain topics. It can be formally defined as follows.

Given two types of text corpora $T^{(1)}$ and $T^{(2)}$ which can be semantically correlated, the text alignment problem targets at obtaining the corresponding relationship set $\mathcal{R}$, where the corresponding relationship set consists of anchor links $r_{ij} = (t_i^{(1)} \rightarrow t_j^{(2)})$, $i = 1, \ldots, |T^{(1)}|; j = 1, \ldots, |T^{(2)}|$. Note that for each tweet $t_i^{(1)}$, there is only one news $t_j^{(2)}$.

IV. HINTA Framework

We introduce a framework of HINTA to address the multisourced text alignment problem. Fig. 3 shows the general framework of HINTA based on the commonly used KDD process model of extracting knowledge [18]. One can see that the framework of aligning the texts is outlined. First, we extract features from the multisourced texts to construct HINs for tweets and news, respectively. Then, we calculate the similarities between the two types of texts to get two similarity graphs (homogeneous networks). In such a step, we will use all the possible anchor text pairs between tweets and news articles, and then calculate the similarities among the two types of texts, respectively. The last step is the inference of correlations on the two networks using different algorithms.

In this section, we introduce the first two parts of the framework. First, we model the tweet and news texts as HINs for further calculation. Second, we select features from the two parts for the calculation of the similarity, and construct the possible pairs according to text similarities.

A. Transform Texts to HINs

In this part, we transform the texts into HINs to represent the two types of texts in a unified way.

1) Texts to HINs: A tweet HIN consists of nodes belonging to the following object types as shown in the left part of Fig. 4: 1) tweet ($T^{(1)}$); 2) word ($O_w^{(1)}$); 3) hashtag (O_h); 4) mention (O_m); and 5) named entities ($O_e^{(1)}$). Similarly, A news network consists of nodes belonging to the following object types: 1) news ($T^{(2)}$); 2) word ($O_w^{(2)}$); and 3) named entities ($O_e^{(2)}$). The topological structure of tweet and news information network is shown in Fig. 4, which forms two star network schemas, where the tweet and news are in the centers and all other objects are linked via them, respectively. Links between the objects denote the semantical correlation among them, which will contribute to the similarity calculation of the texts. Note that the weight of an object is defined as c if it appears c times.

We extract the entities from tweets or news using the tools developed by [43]. Note that the named entities are extracted

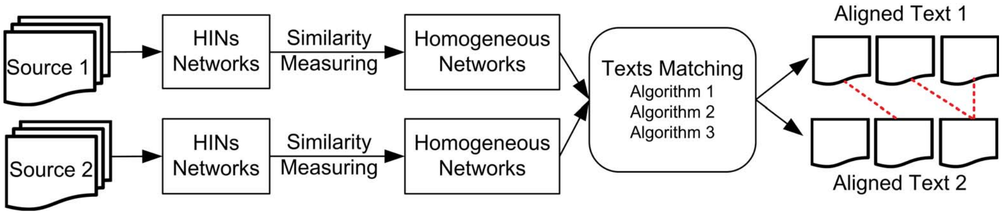


Fig. 3. Illustration of the HINTA framework.

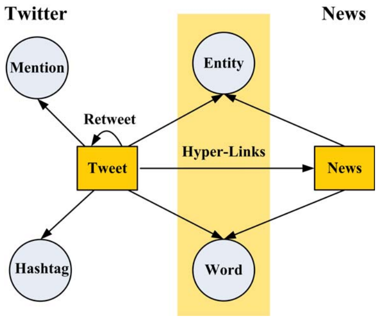


Fig. 4. Schema of the HINs used in HINTA. It contains the tweet part (left) and news part (right). In the tweet part, we have objects with the type of entity, word, mention, and hashtag. In the news part, we have objects of entity and word. The two types of texts are correlated by the objects of entity and word, and hyper-links.

from “words,” “hashtags (#),” and “mentions (@)” of the tweet. In order to emphasize such information, we do not move the corresponding terms out from the above three types of objects. Compared to tweets, news articles are usually much longer and contain more words. To more effectively extract the representative words of news, we utilize both the title and LDA topic words to obtain the identical words in this paper.

B. Similarity Matrix Construction

1) Network Schema and Meta-Path: Based on the above introduction, we can form the network schema of both the tweet and news texts as shown in Fig. 4.

Based on the network schema, we can build the similarity matrix based on the concept of meta-path introduced in [47] to measure the semantic similarity among the texts.

2) Meta-Paths for the Similarity Calculation of Tweets: There are two meta-paths for the calculation of the similarity of tweets in this paper.

(1) *Retweet:* Tweet $\xrightarrow{\text{links}}$ Tweet, whose notation is $T^{(1)} \to T^{(1)}$.

(2) *Common Objects:* Tweet $\xrightarrow{\text{contains}}$ Word/Entity/Mention/Hashtag $\xleftarrow{\text{contains}}$ Tweet, whose notation is $T^{(1)} \to \mathcal{O}_w^{(1)}/\mathcal{O}_e^{(1)}/\mathcal{O}_m/\mathcal{O}_h \leftarrow T^{(1)}$.

3) Meta-Paths for the Similarity Calculation of News: For the similarity of news, there are two meta-paths used in this paper.

(1) *News:* News $\xrightarrow{\text{links}}$ News, whose notation is $T^{(2)} \to T^{(2)}$.

(2) *Common Objects:* News $\xrightarrow{\text{contains}}$ Word/Entity $\xleftarrow{\text{contains}}$ News, whose notation is $T^{(2)} \to \mathcal{O}_w^{(2)}/\mathcal{O}_e^{(2)} \leftarrow T^{(2)}$.

4) Meta-Paths for the Calculation of Semantic Similarity Between Tweet and News Texts: In this paper, we apply two meta-paths for the calculation of the similarity between tweet and news texts.

1) *Hyper-Links:* Tweet $\xrightarrow{\text{links}}$ News, whose notation is $T^{(1)} \to T^{(2)}$.

2) *Common Objects:* Tweet $\xrightarrow{\text{contains}}$ Word(Entity, Mention, Hashtags)/Entity $\xleftarrow{\text{contains}}$ News, whose notation is $T^{(1)} \to \mathcal{O}_w^{(1)}(\mathcal{O}_e^{(1)}, \mathcal{O}_m, \mathcal{O}_h) \leftarrow T^{(2)}$.

Note that named entities contain multiple types (e.g., person, location, and organization names), and we calculate the similarity by distinguishing the name entity type following the work [45]. In reality, the news texts do not contain mentions and hashtags, and if the mention or hashtag phrases are appeared in news, we consider that the news texts contain such objects.

5) Similarity Matrices: PathSim is an effective meta-path-based similarity measurement [47]. Following this paper we also introduce a meta-path-based similarity measure “HINT similarity” (HINTS) to calculate the text similarity.

Definition 3 (HINTS): Let $P_i(x \rightsquigarrow y)$ and $P_i(x \rightsquigarrow \bullet)$ be the sets of path # i instances of HINT going from x to y and those going from x to other nodes in the network. The semantic closeness between two text nodes can be defined as follows:

$$\mathrm{Sim}(x, y) = \sum_i w_i \left(\frac{|P_i(x \rightsquigarrow y)| + |P_i(y \rightsquigarrow x)|}{|P_i(x \rightsquigarrow \bullet)| + |P_i(y \rightsquigarrow \bullet)|} \right) \tag{1}$$

where w_i is the weight of the ith meta-path, and we have $\sum_i w_i = 1$.

As the similarity of each text pairs has been defined, we develop the two similarity matrices according to the numbers of the two types of texts. Let A_i be the adjacency matrix of a kind of texts with respect to the ith meta-path. $A_i(m, n) = k$ denotes that there are k concrete path instances between nodes m and n corresponding to the ith meta-path. Then, the similarity score matrix among all the texts can be represented as

$$S = \sum_i w_i S_i = \sum_i w_i \cdot \mathrm{Norm}\left(A_i + A_i^T\right) \tag{2}$$

where Norm$(\cdot)$ is the normalization of matrix. Thus, the similarity matrix of all possible connections among tweets is represented as above.

In the above, we propose to represent the texts and calculate their similarities via HINs. To compute the similarities between all the texts, in this paper we correlate tweet and news texts through anchored pairs.

6) Anchor Pairs: The tweet and news texts are usually correlated by such hyper-links with respect to semantic meanings

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE I

SUMMARY OF BIASED META-PATHS FOR THE CALCULATION OF SEMANTIC SIMILARITIES BETWEEN TWEET AND NEWS TEXTS

ID	Notation	Heterogeneous Network Meta-Path	Semantic Meaning
1	$T^{(1)} \to \mathcal{O}_w^{(1)} \to T^{(1)} \to T^{(2)}$	$Tweet \xrightarrow{contains} Word \xrightarrow{contains^{-1}} Tweet \xrightarrow{links} News$	Common Words
2	$T^{(1)} \to \mathcal{O}_e^{(1)} \to T^{(1)} \to T^{(2)}$	$Tweet \xrightarrow{contains} Entity \xrightarrow{contains^{-1}} Tweet \xrightarrow{links} News$	Common Entities
3	$T^{(1)} \to \mathcal{O}_m^{(1)} \to T^{(1)} \to T^{(2)}$	$Tweet \xrightarrow{contains} Mention \xrightarrow{contains^{-1}} Tweet \xrightarrow{links} News$	Common Mentions
4	$T^{(1)} \to \mathcal{O}_h^{(1)} \to T^{(1)} \to T^{(2)}$	$Tweet \xrightarrow{contains^{-1}} Hashtag \xrightarrow{contains^{-1}} Tweet \xrightarrow{links} News$	Common Hashtags
5	$T^{(1)} \to \mathcal{O}_w \to T^{(2)} \to T^{(2)}$	$Tweet \xrightarrow{contains} Word \xrightarrow{contains^{-1}} News \xrightarrow{similar} News$	Common Words
6	$T^{(1)} \to \mathcal{O}_e \to T^{(2)} \to T^{(2)}$	$Tweet \xrightarrow{contains} Entity \xrightarrow{contains^{-1}} News \xrightarrow{similar} News$	Common Entities

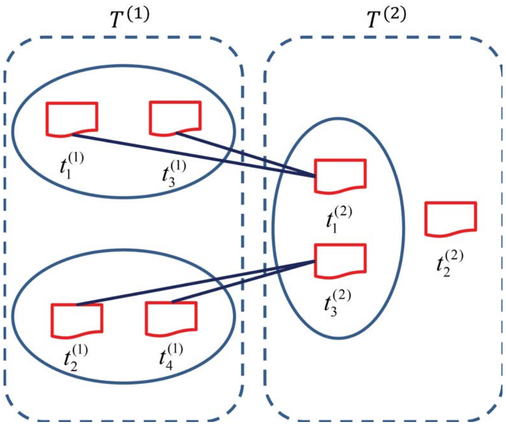


Fig. 5. Illustration of the "one-to-many" corresponding relationships between the two types of related texts. One tweet text ($t_1^{(1)}$) can only be corresponded to one news text ($t_1^{(2)}$) while one news ($t_1^{(2)}$) can be associated with several tweets ($t_1^{(1)}$ and $t_3^{(1)}$).

(as shown in Fig. 4). With this observation, and for the purpose of information transfer among the comparative texts, in this paper we first "align" these tweets with news in terms of semantic meaning and referred these texts as anchored texts.

We first obtain the anchor texts through linked tweets and news pairs following [50]. Since there are still a considerable amount of tweets (about 7.9% in our dataset) which are not closely related to the semantic meaning of news, we filter the anchor texts by common entities ($\mathcal{O}_e$) and topic words ($\mathcal{O}_w$). In this way, we develop a reliable correlation network across the two types of texts for further parameter estimation.

The semantic meanings of tweet and news are not totally identical. One observation is that: a piece of news usually corresponds to many tweets, while a piece of tweets only corresponds to a piece of news article. Thus, the constraint on this type of correlation is "one-to-many." Fig. 5 gives an example of anchor texts and pairs, one can learn that there are four "anchor pairs" across the two types of texts.

V. ALGORITHMS FOR ALIGNING TWEET AND NEWS TEXTS

Once the meta-paths for similarity calculation within the two corpora are obtained, we calculate the similarity between the two types of texts for the inference of correlated pairs, and it serves as the second step of HINTA. We first introduce the sample pairs as bridges between the two types of texts. Then, based on it, we use three different algorithms to calculate the similarities between the two types of texts. Finally, we present the complete algorithm for the alignment of tweet and news texts.

Since we consider such similarity as correlation between the two types of texts. We employ the extended MSA, the SFA [42], and the multinetwork anchoring (MNA) to compute the similarity between tweets and news articles, and thus we obtain the promising pairs as shown in our results.

A. Method 1: Extended Meta-Path-Based Similarity Algorithm

In this part, we introduce the extended meta-path to correlate the tweet and news texts. As mentioned above, traditional meta-paths are symmetric that the object type sequences of a path are the same whatever from which end [47]. For example, meta-path $T^{(1)} \to \mathcal{O}_w^{(1)} \leftarrow T^{(1)}$, the object type sequence is "Tweet-Word-Tweet." We extend the concept of meta-path to "biased path" for the calculation of tweet and news correlation. Our biased meta-paths are from tweet to news in only one direction, and the object sequences are not symmetric. We confine the biased meta-paths in three steps. Different from HINTS, the correlation between tweet and news is calculated by combining a symmetric meta-path and an anchored pair. The correlation $\text{Cor}(x, y)$ of object x and y is calculated as follows:

$$\text{Cor}(x, y) = \text{Sim}(x, x') * \text{Sim}(x', y) \tag{3}$$

where x' is an object with the same type to x, $\text{Sim}(x, x')$ is calculated following (1), and $\text{Sim}(x', y)$ is an anchored pair as mentioned above. The biased meta-paths used in this paper are list in Table I.

B. Method 2: Similarity Flooding Algorithm

The correlation of tweet and news texts calculated of by SFA is based on the pairwise connectivity graph (PCG) which is originated from the similarity graph of the two types of texts. We use the method proposed in [42] to generate such a PCG, and compute the fixed points of the tweet-news pairs as the correlation between them.

Having calculated the similarities between the tweet and news datasets, we can develop a homogeneous graph with the texts (tweet or news) as nodes and the similarity connections among them as weighted edges. Then, we first define a PCG as follows: $(t^{(1)}, p, t^{(2)}) \in \text{PCG}(T^{(1)}, T^{(2)}) \leftrightarrow t^{(1)} \in T^{(1)}$

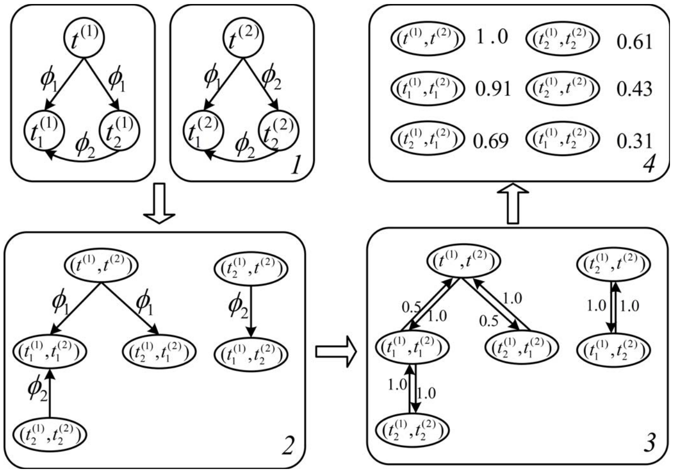


Fig. 6. Illustration of the SFA. We first transfer the two network into PCG (left), and then calculate the similarities by rules (bottom).

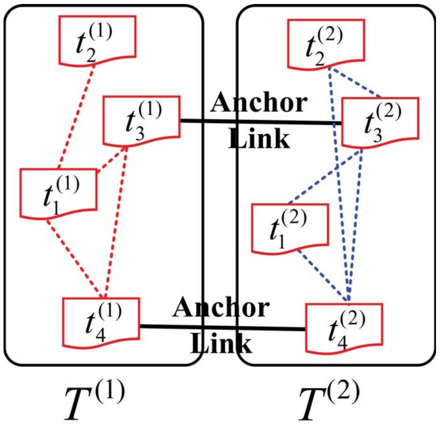


Fig. 7. Illustration of the extended similarity measures based on the cross-network anchoring.

and $t^{(2)} \in T^{(2)}$, where p is the correlation score. The PCG is generated by the method proposed in [42].

To illustrate how the propagation graph is computed, we illustrate the PCG and the calculating steps in Fig. 6. We utilize the anchored pairs as the root pairs (as shown in the part 2 of Fig. 6). The calculation steps can be illustrated as follows. Let $\sigma(t^{(1)}$ and $t^{(2)}) \geq 0$ be the similarity measure of nodes $t^{(1)} \in T^{(1)}$ and $t^{(2)} \in T^{(2)}$ defined as a total function over $T^{(1)} \times T^{(2)}$, where σ is referred as a mapping. The SFA is based on an iterative computation of σ-values. Let σ^i denote the mapping between $T^{(1)}$ and $T^{(2)}$ after the ith iteration. Mapping σ^0 represents the initial similarities among all the node pairs of $T^{(1)}$ and $T^{(2)}$. In our example, we assume that no initial mapping between A and B is available, for example, $\sigma^0(t^{(1)}, t^{(2)}) = 1.0$ for all $t^{(1)} \in T^{(1)}$ and $t^{(2)} \in T^{(2)}$ as shown in Fig. 6. In every iteration, the σ-values for a map pair $\sigma^0(t^{(1)}, t^{(2)})$ are incremented by the σ-values of its neighbor pairs in the propagation graph multiplied by the propagation coefficients on the edges going from the neighbor pairs to $(t^{(1)}, t^{(2)})$. For example, after the first iteration $\sigma^1(t_1^{(1)}, t_1^{(2)}) = \sigma^0(t_1^{(1)}, t_1^{(2)}) + \sigma^0(t^{(1)}, t^{(2)}) * 0.5 + \sigma^0(t_2^{(1)}, t_2^{(2)}) * 1.0 = 2.5$. Analogously, $\sigma^1(t^{(1)}, t^{(2)}) = \sigma^0(t^{(1)}, t^{(2)}) + \sigma^0(t_1^{(1)}, t_1^{(2)}) * 1.0 + \sigma^0(t_2^{(1)}, t_1^{(2)}) * 1.0 = 3.0$. Then, all the values in the propagation network are normalized, i.e., divided by the maximal σ-value (of current iteration) $\sigma^1(t^{(1)}, t^{(2)}) = 3.0$. After normalization, we get $\sigma^1(t^{(1)}, t^{(2)}) = 1.0$, $\sigma^1(t_1^{(1)}, t_1^{(2)}) = 0.833$, etc. In general, mapping σ^{i+1} is computed from mapping σ^i as follows (for clarity, the normalization is omitted):

$$\begin{aligned}\sigma^{i+1}\left(t^{(1)}, t^{(2)}\right) = \sigma^i\left(t^{(1)}, t^{(2)}\right) &+ \sum w\left(\left(t_u^{(1)}, t_u^{(2)}\right), \left(t^{(1)}, t^{(2)}\right)\right)\\ &+ \sum w\left(\left(t_v^{(1)}, t_v^{(2)}\right), \left(t^{(1)}, t^{(2)}\right)\right) \quad (4)\end{aligned}$$

where $(t_u^{(1)}, t^{(1)}) \in T^{(1)}$, $(t_u^{(2)}, t^{(2)}) \in T^{(2)}$, and $(t^{(1)}, t_v^{(1)}) \in T^{(1)}$, $(t^{(2)}, t_v^{(2)}) \in T^{(2)}$. We compute the equation iteratively till the residual vector $\delta(\sigma^n, \sigma^{n-1})$ becomes less than ϵ for some $n > 0$. If (4) does not converge, we will terminate the computation after some maximal number of iterations. The convergence of SFA is discussed in [42], and according to it, the algorithm will converge if the graph is well connected.

In this paper, we normalize σ^i by formula $\sigma^{i+1} = \text{normalize}(\sigma^0 + \sigma^i + \phi(\sigma^0 + \sigma^i))$, and we will discuss the normalization in Section VI.

C. Method 3: Cross-Network Anchoring-Based Algorithm

This method apply the homogeneous networks after the calculation of similarity between tweets in the tweet network and news in the news network. Following previous work [14], we extend the definitions of some widely used measures in link prediction to correlate the two networks. Fig. 7 show the basic relationships between such two types of texts, and the corresponding similarity measures are based on the cross-network anchor pairs $(t_3^{(1)}, t_3^{(2)})$ and $(t_4^{(1)}, t_4^{(2)})$. Here, we first defined the neighbors used in the following measures, we only consider the neighbors that are cross-network anchored.

1) Extended Common Neighbors: The common neighbors indicate the number of pairs of tweet or news texts that are semantically related to a particular tweet-news pair. We denote the number of common neighbors as $\text{CN}(t_i^{(1)}, t_j^{(2)})$ between $t_i^{(1)}$ in the tweet formed network and $t_j^{(2)}$ in the news formed network. Then the neighbors of $t_i^{(1)}$ in the tweet network can be denoted as $\Gamma_{(1)}(t_i^{(1)})$, and the neighbors of $t_j^{(2)}$ in the news network is $\Gamma_{(2)}(t_j^{(2)})$. The extended common neighbor measure is defined as follows:

$$CN\left(t_i^{(1)}, t_j^{(2)}\right) = \left|\Gamma\left(t_i^{(1)}\right) \cap_{\mathcal{A}} \Gamma\left(t_j^{(2)}\right)\right|. \quad (5)$$

Noted that $\cap_{\mathcal{A}}$ means the common neighbors are all included in the anchored pairs $\mathcal{A}$.

2) Extended Jaccard's Coefficient: The extended measure of Jaccard's coefficient using the similar method of extending common neighbors. The extended Jaccard's coefficient $\text{JC}(t_i^{(1)}, t_j^{(2)})$ is defined as a normalized version of common neighbors divided by the total number of distinct users in

$\Gamma(t_i^{(1)}) \cup \Gamma(t_j^{(2)})$

$$JC\left(t_i^{(1)}, t_j^{(2)}\right) = \frac{\left|\Gamma\left(t_i^{(1)}\right) \cap_{\mathcal{A}} \Gamma\left(t_j^{(2)}\right)\right|}{\left|\Gamma\left(t_i^{(1)}\right) \cup_{\mathcal{A}} \Gamma\left(t_j^{(2)}\right)\right|} \tag{6}$$

where

$$\left|\Gamma\left(t_i^{(1)}\right) \cup_{\mathcal{A}} \Gamma\left(t_j^{(2)}\right)\right| = \left|\Gamma\left(t_i^{(1)}\right)\right| + \left|\Gamma\left(t_i^{(2)}\right)\right| - \left|\Gamma\left(t_i^{(1)}\right) \cup_{\mathcal{A}} \Gamma\left(t_j^{(2)}\right)\right|. \tag{7}$$

3) Extended Adamic/Adar Measure: The adamic/adar measure is used to analyze the links within a social network. Here, to connect the two types of networks, we extend the measure into the cross text networks settings, where the common neighbors are weighted by their average degrees in both social networks

$$AA\left(t_i^{(1)}, t_j^{(2)}\right) = \sum_{\forall\left(t_p^{(1)}, t_q^{(2)}\right) \in \Gamma\left(t_p^{(1)}\right) \cup_{\mathcal{A}} \Gamma\left(t_q^{(2)}\right)} \log^{-1}\left(\frac{\left|\Gamma\left(t_p^{(1)}\right)\right| + \left|\Gamma\left(t_q^{(1)}\right)\right|}{2}\right). \tag{8}$$

VI. Experimental Results

In this section, we will evaluate the proposed framework HINTA on three real datasets and compare the MSA, SFA, and MNA algorithms with several typical baselines.

A. Dataset Description and Preprocessing

Dataset 1: Such a dataset is collected by Guo *et al.* [50], and contains tweets spanning over 18 days. Each tweet contains a URL linking to a news article of CNN or NYTIMES. All the news of CNN and NYTIMES are published during this period. The final dataset contains 34 888 tweets and 12 704 news articles.

Dataset 2: The original tweet corpora was crawled from the over four million followers of Hillary Clinton. We keep the tweets containing news URLs, and randomly select other tweets posted from June 1, 2015 to June 7, 2015 to make up the tweet part. The news part was crawled via the URLs appeared in the tweets that covered over 20 news sites. The final dataset contains 5628 tweets and 3653 news articles.

Dataset 3: Similar to DataSet 2, the tweet part consists of the tweets with URLs and the randomly selected ones posted from June 25, 2015 to July 2, 2015. The news part was crawled based on the URLs appeared in the tweets that covered over 20 news sites. The final dataset contains 3847 tweets and 2825 news articles.

TABLE II
Summary of Datasets

	Objects	Dataset 1	Dataset 2	Dataset 3
Tweets	#Hashtag	15,471	1,602	1,526
	#Mention	12,911	4,120	2,950
	#Keyword	13,322	7,325	4,527
	#Author	–	4,806	3,271
	#HyperLink	2,710	4,345	3,784
	#Entity	5,468	1,237	666
News	#Entity	8,325	8,696	18,696
	#Keyword	16,210	17,475	21,475

Following a similar approach as in [46], we preprocess tweets and news, respectively, by removing stopwords, extracting hashtags, mentions, keywords, and entities on tweet datasets, and also entities and keywords on news dataset. By solving short URLs to expanded identical URLs, we construct the substitution of out-of-vocabulary words, the moving out of long-tailed users, tweets. Following [45], we use three types of entities: 1) the person (P); 2) organization (O); and 3) location (L) in tweets and news. After preprocessing, the two corpora associated with useful properties are obtained and the description on them is summarized in Table II.

We let three annotators to annotate the text and obtain the ground truth of the one-to-many matching. The annotated matching results are with 98.3% consensus among the annotators.

B. Baselines and Assessment Methods

1) Baselines: In order to study the performances of our algorithms in the HINTA framework. We conduct the experiments by comparing our methods with eight baseline methods which are summarized as follows.

2) Proposed Correlation Algorithms: MSA, SFA, and MNA. The general idea of framework HINTA is transforming the texts into HINs, and using the hyper-linked texts as anchored pairs for semi-supervised learning. As in the correlating step, HINTA applies three different algorithms to correlate the texts. Specifically, the MSA correlates the two types of texts by extended meta-path-based similarities. The SFA models the two types of texts as pairs and correlates them by calculating the static similarities between them. The MNA calculates the similarities between the two types of texts by three extended measures.

3) Variants of the Correlation Algorithms: We apply three types of variants in the correlation algorithm.

(1) *Direct Calculation of MSA and MSA-dr:* We vary the MSA by viewing the object types of tweet and news as the same, and calculate the similarities of the meta-paths in Table III using HINTS directly.

(2) *Variants of SFA, SFA-1, SFA-2, and SFA-3:* We try three other normalization formulas of SFA to form the variants of SFA, which are denoted as SFA-1, SFA-2, and SFA-3, respectively. The normalization formulas are illustrated in Table III. Note that the general idea of function ϕ is to increment the similarities of each map pair based on the similarities of their neighbors in the propagation graph.

(3) *MNA Without Anchored Pairs, MNA-No:* We apply the MNA without using the anchored pairs, and denote it as MNA-no. Such a variant of MNA can be employed to investigate the utility of anchored pairs.

4) Commonly Used Text Similarity Measures: To the best of our knowledge, there is no specific method that can be directly

TABLE III
SUMMARY OF FORMULA USED IN THE SFA

Notation	Fixpoint Formula
SFA	$\sigma^{i+1} = normalize(\sigma^0 + \sigma^i + \phi(\sigma^0 + \sigma^i))$
SFA-1	$\sigma^{i+1} = normalize(\sigma^i + \phi(\sigma^i))$
SFA-2	$\sigma^{i+1} = normalize(\sigma^0 + \phi(\sigma^i))$
SFA-3	$\sigma^{i+1} = normalize(\phi(\sigma^0 + \sigma^i))$

employed to the proposed problem. We employ the following commonly used methods in text analysis to correlate the texts as baselines [8].

1) *Word Co-Occurrence (WCO):* The methods using WCOs to match the two types of texts.
2) *TF-IDF + K-Means (TIK):* We use the TIK on the tweets and news articles with equal, and choose the most closing pairs as correlation.
3) *NG:* We use the methods using NG for comparison, here we use the best performance with $n = 4$.

5) Assessment: We apply two measures to assess the quality of matching, the F-score and accuracy. We use the typical definition of the F-score measures. Following [42], we define the accuracy from the view of how much effort it saves the user to modify the proposed match result $P = \{(t_1^{(1)}, t_1^{(2)}), \ldots, (t_n^{(1)}, t_n^{(2)})\}$ into the labeled dataset $L = \{(a_1, b_1), \ldots, (a_m, b_m)\}$. For simplicity, we assume that deletions and additions of match pairs require the same amount of effort, and that the verification of a correct match pair is free.

We denote the number of accurate matching as $c = |P \cap L|$. Thus, the number of false positive matchings can be denoted as $n-c$, and the false negative matching number is $m-c$. From the above definition we can find that (c/m) and (c/n) correspond to recall and precision of matching. According to [42], the match accuracy as a function of recall and precision is defined as follows:

$$\begin{aligned}\text{Accuracy} &= 1 - \frac{(n-c)+(m-c)}{m} = \frac{c}{m}\left(2 - \frac{n}{c}\right)\\ &= \text{Recall}\left(2 - \frac{1}{\text{Precision}}\right).\end{aligned} \tag{9}$$

Note that in such a definition, the notion of accuracy only makes sense if precision is not less than 0.5, i.e., at least half of the returned matches are correct. Otherwise, the accuracy is negative.

C. Experiment Settings

In this part, we first introduce the parameter settings of the proposed methods. The framework contains two parts, the first part is the calculation of basic similarities among the tweet and news text datasets. The second part is the three kinds of different methods. As for the first part, we set the diagonal vector of tweet meta-paths as $[(1/4), \ldots, (1/4)]$, and that of news meta-paths $[(1/2), (1/2)]$. As for the second part, we first introduce the experiment settings of the three methods proposed in this paper, respectively.

1) The extended MSA and its variant MSA-dr contain six biased meta-paths which may contribute to the similarity between tweets and news. To be simplicity, we set the diagonal vector of the biased meta-path as $[(1/6), \ldots, (1/6)]$.

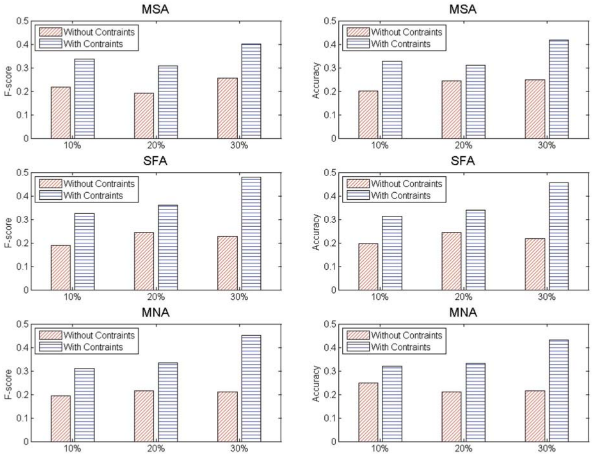


Fig. 8. Comparison on the methods with and without time constraints. The y-axis of all the subfigures are ratio of imbalance .

2) The fixed points computation contains four different iterative calculational formula on the similarity network. In this part, we study the four fix point formulas, respectively. According to the experience, the needed iterations is no more than 48 till the convergence, thus we set the max iterations as 50, and the θ is 0.05.

In real-world link prediction problems, the data samples are usually imbalanced. Such an issue also exists in the text alignment problem. Therefore, we test the performances of all the methods and baselines with imbalanced datasets. In each round of the cross validation, we sample pairs of user accounts as the data samples according to different imbalance ratios, i.e., #negative pairs and #positive pairs. The following experiments are conducted under the imbalanced ratio of 10%, 20%, and 30%, respectively. For each method on a dataset with a specific imbalance ratio, we run the experiments for 100 times.

D. Experimental Result Analysis

In this part, we first investigate the performance of the proposed algorithms with and without time constraints, and then we compare the matching results of our algorithms with the baselines mentioned above.

We investigate the performance of MSA, SFA, MNA with and without time constraints, respectively. Specifically, the time constraints mean that we only consider the news articles $t_j^{(2)}$ which are published ahead of a tweet $t_i^{(1)}$ within three days (72 h). It is denoted as time gap $T_G(t_i^{(1)}) - T_G(t_j^{(2)}) < 72$ h.

Fig. 8 shows the performances of the MSA, SFA, and MNA with and without time constraints. From the figure one can see that time constraints contribute significantly to the final performance of all the methods, which is approximately 30% on average to the final results of F-score and 35% of accuracy. Generally, the methods with constraints outperform the corresponding methods without constraints. It is necessary to apply the "time" as constraints to the methods. Therefore, we use the time constrained method in the following experiments.

TABLE IV
PERFORMANCE COMPARISON OF DIFFERENT METHODS FOR ALIGNING TWEET AND NEWS TOGETHER. USE DIFFERENT IMBALANCE RATIOS OF THE THREE DATASETS. (IMBALANCE RATION = # POSITIVE ACCOUNT PAIRS/ # NEGATIVE ACCOUNT PAIRS)

		Dataset 1			Dataset 2			Dataset 3		
		10	20	30	10	20	30	10	20	30
F1	MSA	**0.3129 ± 0.0150**	**0.3435 ± 0.0078**	**0.3519 ± 0.0024**	0.2778 ± 0.0021	0.3128 ± 0.0409	0.3342 ± 0.0297	0.3419 ± 0.0041	0.4095 ± 0.0427	0.4507 ± 0.0328
	MSA-dr	0.1828 ± 0.0031	0.2062 ± 0.0026	0.2317 ± 0.0034	0.2334 ± 0.0019	0.2243 ± 0.0038	0.2304 ± 0.0028	0.2424 ± 0.0021	0.3019 ± 0.0046	0.3384 ± 0.0061
	SFA	0.3073 ± 0.0068	0.3292 ± 0.0017	0.3401 ± 0.0168	**0.3243 ± 0.0028**	**0.3634 ± 0.0020**	**0.3952 ± 0.0047**	**0.4138 ± 0.0081**	**0.4973 ± 0.0060**	**0.5269 ± 0.0076**
	SFA-1	0.2874 ± 0.0093	0.3258 ± 0.0116	0.3284 ± 0.0124	0.2741 ± 0.0038	0.3299 ± 0.0112	0.3542 ± 0.0116	0.3871 ± 0.0031	0.4584 ± 0.0027	0.4875 ± 0.0101
	SFA-2	0.2819 ± 0.0082	0.3022 ± 0.0034	0.3308 ± 0.0073	0.2964 ± 0.0019	0.3345 ± 0.0042	0.3770 ± 0.0088	0.3921 ± 0.0024	0.4494 ± 0.0076	0.4937 ± 0.0040
	SFA-3	0.2684 ± 0.0042	0.2907 ± 0.0021	0.3118 ± 0.0061	0.3029 ± 0.0007	0.3488 ± 0.0135	0.3760 ± 0.0162	0.3945 ± 0.0171	0.4612 ± 0.0021	0.4893 ± 0.0098
	MNA-no	0.1608 ± 0.0025	0.1792 ± 0.0036	0.1650 ± 0.0047	0.1609 ± 0.0034	0.1824 ± 0.0020	0.2232 ± 0.0047	0.2349 ± 0.0061	0.1962 ± 0.0042	0.2604 ± 0.0046
	MNA	0.3070 ± 0.0075	0.3019 ± 0.0056	0.3250 ± 0.0160	0.2909 ± 0.0054	0.3482 ± 0.0046	0.3661 ± 0.0054	0.3831 ± 0.0049	0.4668 ± 0.0090	0.5024 ± 0.0064
	WCO	0.2377 ± 0.0045	0.2348 ± 0.0039	0.2461 ± 0.0023	0.1907 ± 0.0026	0.2407 ± 0.0019	0.2764 ± 0.0029	0.3182 ± 0.0040	0.3049 ± 0.0026	0.3926 ± 0.0016
	TIK	0.2368 ± 0.0046	0.2629 ± 0.0035	0.2150 ± 0.0060	0.1908 ± 0.0044	0.2301 ± 0.0024	0.2367 ± 0.0038	0.3027 ± 0.0015	0.3906 ± 0.0037	0.4067 ± 0.0086
	N-Grams	0.2208 ± 0.0025	0.2392 ± 0.0036	0.2450 ± 0.0047	0.2109 ± 0.0034	0.2624 ± 0.0020	0.2722 ± 0.0047	0.3049 ± 0.0061	0.3162 ± 0.0042	0.3904 ± 0.0046
Acc.	MSA	**0.2991 ± 0.0019**	**0.3304 ± 0.0010**	**0.3504 ± 0.0046**	0.2591 ± 0.0046	0.3216 ± 0.0039	0.3480 ± 0.0071	0.3587 ± 0.0073	0.4153 ± 0.0136	0.4762 ± 0.0151
	MSA-dr	0.2826 ± 0.0015	0.3084 ± 0.0057	0.3345 ± 0.0022	0.2724 ± 0.0066	0.3298 ± 0.0055	0.3547 ± 0.0168	0.3558 ± 0.0164	0.4460 ± 0.0011	0.4727 ± 0.0174
	SFA	0.2811 ± 0.0087	0.3157 ± 0.0057	0.3456 ± 0.0057	**0.2822 ± 0.0047**	**0.3536 ± 0.0021**	**0.3809 ± 0.0025**	**0.3894 ± 0.0019**	**0.4645 ± 0.0025**	**0.5118 ± 0.0057**
	SFA-1	0.2749 ± 0.0016	0.3116 ± 0.0040	0.3322 ± 0.0068	0.2788 ± 0.0083	0.3351 ± 0.0044	0.3540 ± 0.0106	0.3745 ± 0.0133	0.4500 ± 0.0049	0.4861 ± 0.0028
	SFA-2	0.2811 ± 0.0106	0.2964 ± 0.0077	0.3402 ± 0.0098	0.2786 ± 0.0033	0.3360 ± 0.0077	0.3620 ± 0.0174	0.3795 ± 0.0104	0.4388 ± 0.0114	0.4932 ± 0.0058
	SFA-3	0.2824 ± 0.0066	0.3020 ± 0.0025	0.3307 ± 0.0040	0.2662 ± 0.0050	0.3245 ± 0.0030	0.3566 ± 0.0120	0.3637 ± 0.0028	0.4290 ± 0.0039	0.4737 ± 0.0032
	MNA-no	0.1453 ± 0.0037	0.1669 ± 0.0023	0.1558 ± 0.0064	0.1419 ± 0.0024	0.1851 ± 0.0023	0.2074 ± 0.0027	0.2091 ± 0.0026	0.2003 ± 0.0065	0.2107 ± 0.0042
	MNA	0.2850 ± 0.0069	0.3259 ± 0.0016	0.3467 ± 0.0117	0.2775 ± 0.0006	0.3486 ± 0.0012	0.3670 ± 0.0095	0.3733 ± 0.0135	0.4409 ± 0.0039	0.4825 ± 0.0205
	WCO	0.2274 ± 0.0025	0.2681 ± 0.0042	0.2310 ± 0.0050	0.1874 ± 0.0024	0.2018 ± 0.0036	0.2409 ± 0.0041	0.3037 ± 0.0041	0.4128 ± 0.0034	0.4149 ± 0.0040
	TIK	0.2384 ± 0.0052	0.2308 ± 0.0037	0.2159 ± 0.0070	0.2250 ± 0.0062	0.2543 ± 0.0019	0.2408 ± 0.0047	0.3180 ± 0.0026	0.4107 ± 0.0042	0.4058 ± 0.0056
	N-Grams	0.2179 ± 0.0056	0.2382 ± 0.0043	0.2350 ± 0.0042	0.1908 ± 0.0034	0.2108 ± 0.0023	0.2607 ± 0.0024	0.3103 ± 0.0045	0.3884 ± 0.0053	0.3934 ± 0.0036

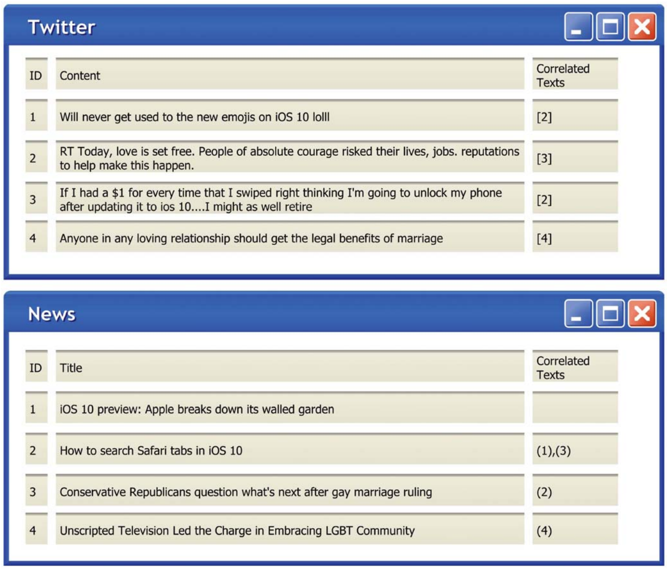
For simplicity, we will still use "MSA," "SFA," and "MNA" to denote them, respectively.

We also compare the performances of the proposed methods with the baselines as mentioned before. Table IV shows the final results in the form of mean and standard deviations of the two assessment measures. The conclusions of the experimental results are drew as follows.

First, we investigate whether HINs fit the representation of various types of texts. As one can see from Table IV, the proposed HIN-based methods (MSA, SFA, MNA) are generally performing better than the ones (WCO, TIK, NG) not using HINs in most cases. Specifically, according to our statistics, the three proposed HIN-based methods (MSA, SFA, MNA) averagely outperform the other non-HIN-based methods (WCO, TIK, NG) about 34.5% and 30.2% by F-score and accuracy, respectively. The results supports the intuition of this paper, HINs can be used to represent various types of texts and improve the alignment quality.

Second, we investigate the utility of the anchored pairs by comparing the performances of MNA with MNA-no. Remind that the method MNA uses the anchored pairs while MNA-no does not use. From the table one can see that the MNA outperforms MNA-no in all cases, especially in the assessment measure of accuracy (over 80% in the three datasets with various imbalanced ratios). Averagely, the method MNA outperforms MNA-no about 86.7% and 100.1% by F-score and accuracy on the three datasets with different imbalanced ratios, respectively. Obviously, the result suggests that one of our motivations is right that anchored texts can be used to improve the alignment quality.

Finally, there is no specific method performs the best in all cases. The method MSA outperforms all the other methods and baselines on the first dataset. However, when it comes to second and last dataset, the SFA performs the best among all the methods. That is because the MSA method can get higher precision than the other methods in most cases, and the SFA method can usually get higher recall than other methods. To be more detail, we can find that as for the MSA, the method using biased meta-path outperforms the one that does not use the first method. As for the second method the SFA, the method using the third normalization formula (SFA-3) outperforms all the other three variants, which confirms the conclusions drew by Melnik *et al.* [42].



Twitter

ID	Content	Correlated Texts
1	Will never get used to the new emojis on iOS 10 lolll	[2]
2	RT Today, love is set free. People of absolute courage risked their lives, jobs. reputations to help make this happen.	[3]
3	If I had a $1 for every time that I swiped right thinking I'm going to unlock my phone after updating it to ios 10....I might as well retire	[2]
4	Anyone in any loving relationship should get the legal benefits of marriage	[4]

News

ID	Title	Correlated Texts
1	iOS 10 preview: Apple breaks down its walled garden	
2	How to search Safari tabs in iOS 10	(1),(3)
3	Conservative Republicans question what's next after gay marriage ruling	(2)
4	Unscripted Television Led the Charge in Embracing LGBT Community	(4)

Fig. 9. Case study of the text alignment in the BISs.

E. Case Study

We also conduct a case study on dataset 2 which serves as a dataset of an online business intelligence system. The system integrates the multisourced texts by aligning the two types of texts from Twitter and news websites, respectively.

Fig. 9 shows the interface of the system. From the figure one can see that windows 1 and 2 contain the news and tweets that are semantically correlated. Specifically, the tweets in the first window are related to the topic of *apple* and *equal marriage right*. The second window is a news table which is about *apple operation system* and the *LGBT movement*. Each window has a column called *Correlated texts*, where tweet 1–4 are correlated to news texts 2, 3, 2, and 4, respectively, and news 1 has no correlation (although it is semantically related to 1, 3, due to the used algorithm, it has not been correlated), news 2 is correlated two tweet 1 and 3, and news 3 and 4 is correlated to tweet 2 and 4, respectively. From the figure, we can also see that the tweets contain more personal views from users, and the views are usually direct and emotional. By aligning the texts from different sources, we can obtain more comprehensive information from different perspectives.

VII. Conclusion

In this paper, we study the problem of integrating the texts from different sources in online BISs. We propose a framework HINTA to address the problem. The framework transforms different types of texts into a unified form of HINs, and applies three different methods to match the two types of texts based on anchored pairs. We conduct extensive experiments on three real-world datasets and the results prove the effectiveness and efficiency of our framework. This paper a promising attempt in the problem of text integration in BIS. In the future, we can integrate more different types of texts via HINs or other approaches to get more comprehensive information.

References

[1] F.-Y. Wang, K. M. Carley, D. Zeng, and W. Mao, "Social computing: From social informatics to social intelligence," *IEEE Intell. Syst.*, vol. 22, no. 2, pp. 79–83, Mar./Apr. 2007.

[2] L. Duan and L. Da Xu, "Business intelligence for enterprise systems: A survey," *IEEE Trans. Ind. Informat.*, vol. 8, no. 3, pp. 679–687, Aug. 2012.

[3] S. Wang *et al.*, "Estimating urban traffic congestions with multi-sourced data," in *Proc. 17th IEEE Int. Conf. Mobile Data Manag.*, 2016, pp. 82–91.

[4] A. Popovič, R. Hackney, P. S. Coelho, and J. Jaklič, "Towards business intelligence systems success: Effects of maturity and culture on analytical decision making," *Decis. Support Syst.*, vol. 54, no. 1, pp. 729–739, 2012.

[5] H.-K. Oh, S.-W. Kim, S. Park, and M. Zhou, "Can you trust online ratings? A mutual reinforcement model for trustworthy online rating systems," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 12, pp. 1564–1576, Dec. 2015.

[6] F.-Y. Wang, X. Wang, L. Li, and L. Li, "Steps toward parallel intelligence," *IEEE/CAA J. Autom. Sinica*, vol. 3, no. 4, pp. 345–348, Oct. 2016.

[7] C. Zhai, A. Velivelli, and B. Yu, "A cross-collection mixture model for comparative text mining," in *Proc. SIGKDD*, Seattle, WA, USA, 2004, pp. 743–748.

[8] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Amsterdam, The Netherlands: Elsevier, 2011.

[9] S. Agreste, P. De Meo, E. Ferrara, S. Piccolo, and A. Provetti, "Analysis of a heterogeneous social network of humans and cultural objects," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 4, pp. 559–570, Apr. 2015.

[10] J. Peng, D. D. Zeng, H. Zhao, and F.-Y. Wang, "Collaborative filtering in social tagging systems based on joint item-tag recommendations," in *Proc. ACM Conf. Inf. Knowl. Manag.*, Toronto, ON, Canada, 2010, pp. 809–818.

[11] Q. Wang *et al.*, "cluTM: Content and link integrated topic model on heterogeneous information networks," in *Proc. WAIM*, Qingdao, China, 2015, pp. 207–218.

[12] Y. Lv, Y. Chen, X. Zhang, Y. Duan, and N. L. Li, "Social media based transportation research: The state of the work and the networking," *IEEE/CAA J. Autom. Sinica*, vol. 4, no. 1, pp. 19–26, Jan. 2017.

[13] D. Zeng, F.-Y. Wang, and M. Liu, "Efficient Web content delivery using proxy caching techniques," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 34, no. 3, pp. 270–280, Aug. 2004.

[14] X. Kong, J. Zhang, and P. S. Yu, "Inferring anchor links across multiple heterogeneous social networks," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manag.*, San Francisco, CA, USA, 2013, pp. 179–188.

[15] M. Bayati, M. Gerritsen, D. F. Gleich, A. Saberi, and Y. Wang, "Algorithms for large, sparse network alignment problems," in *Proc. IEEE 9th Int. Conf. Data Min. (ICDM)*, 2009, pp. 705–710.

[16] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513–523, 1988.

[17] X. Wang, L. Li, Y. Yuan, P. Ye, and F.-Y. Wang, "ACP-based social computing and parallel intelligence: Societies 5.0 and beyond," *CAAI Trans. Intell. Technol.*, vol. 1, no. 4, pp. 377–393, 2016.

[18] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Commun. ACM*, vol. 39, no. 11, pp. 27–34, 1996.

[19] M. Kay and M. Röscheisen, "Text-translation alignment," *Comput. Linguist.*, vol. 19, no. 1, pp. 121–142, 1993.

[20] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Computer-based authorship attribution without lexical measures," *Comput. Humanities*, vol. 35, no. 2, pp. 193–214, 2001.

[21] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," in *Proc. LREC*, 2012, pp. 2214–2218.

[22] H. Gu, H. Hang, Q. Lv, and D. Grunwald, "Fusing text and frienships for location inference in online social networks," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol. (WI-IAT)*, vol. 1. 2012, pp. 158–165.

[23] V. Keselj, F. Peng, N. Cercone, and C. Thomas, "N-gram-based author profiles for authorship attribution," in *Proc. Conf. Pac. Assoc. Comput. Linguist. (PACLING)*, vol. 3. 2003, pp. 255–264.

[24] D. Mladenic and M. Grobelnik, "Word sequences as features in text-learning," in *Proc. 17th Electrotech. Comput. Sci. Conf. (ERK)*, 1998, pp. 145–148.

[25] R. Yan, M. Lapata, and X. Li, "Tweet recommendation with graph co-ranking," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguist. Long Papers*, vol. 1. Jeju-do, South Korea, 2012, pp. 516–525.

[26] T. Hua, F. Chen, L. Zhao, C.-T. Lu, and N. Ramakrishnan, "STED: Semi-supervised targeted-interest event detectionin in twitter," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, Chicago, IL, USA, 2013, pp. 1466–1469.

[27] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 58, no. 7, pp. 1019–1031, 2007.

[28] M. Al Hasan, V. Chaoji, S. Salem, and M. Zaki, "Link prediction using supervised learning," in *Proc. Workshop Link Anal. Counter Terrorism Security (SDM)*, 2006, pp. 798–805.

[29] C. Wang, V. Satuluri, and S. Parthasarathy, "Local probabilistic models for link prediction," in *Proc. 7th IEEE Int. Conf. Data Min. (ICDM)*, Omaha, NE, USA, 2007, pp. 322–331.

[30] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths, "Probabilistic author-topic models for information discovery," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, Seattle, WA, USA, 2004, pp. 306–315.

[31] H. Deng, J. Han, B. Zhao, Y. Yu, and C. X. Lin, "Probabilistic topic models with biased propagation on heterogeneous information networks," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, San Diego, CA, USA, 2011, pp. 1271–1279.

[32] S. Wang, X. Hu, P. S. Yu, and Z. Li, "MMRate: Inferring multi-aspect diffusion networks with multi-pattern cascades," in *Proc. KDD*, New York, NY, USA, 2014, pp. 1246–1255.

[33] S. Wang, Z. Yan, X. Hu, P. S. Yu, and Z. Li, "Burst time prediction in cascades," in *Proc. AAAI*, Austin, TX, USA, 2015, pp. 325–331.

[34] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang, "Mining topic-level influence in heterogeneous networks," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manag.*, Toronto, ON, Canada, 2010, pp. 199–208.

[35] T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection: A discriminative approach," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, Paris, France, 2009, pp. 927–936.

[36] L. Tang and H. Liu, "Relational learning via latent social dimensions," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, Paris, France, 2009, pp. 817–826.

[37] E. Zheleva, H. Sharara, and L. Getoor, "Co-evolution of social and affiliation networks," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, Paris, France, 2009, pp. 1007–1016.

[38] G. Kossinets, J. Kleinberg, and D. Watts, "The structure of information pathways in a social communication network," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, Las Vegas, NV, USA, 2008, pp. 435–443.

[39] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *Proc. 13th Int. Conf. World Wide Web*, New York, NY, USA, 2004, pp. 491–501.

[40] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 64, no. 2, 2001, Art. no. 025102.

[41] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, "Microscopic evolution of social networks," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, Las Vegas, NV, USA, 2008, pp. 462–470.

[42] S. Melnik, H. Garcia-Molina, and E. Rahm, “Similarity flooding: A versatile graph matching algorithm and its application to schema matching,” in *Proc. IEEE 18th Int. Conf. Data Eng.*, San Jose, CA, USA, 2002, pp. 117–128.

[43] C. D. Manning *et al.*, “The stanford coreNLP natural language processing toolkit,” in *Proc. ACL Syst. Demonstrations*, Baltimore, MD, USA, 2014, pp. 55–60.

[44] B. Zadrozny and C. Elkan, “Transforming classifier scores into accurate multiclass probability estimates,” in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, Edmonton, AB, Canada, 2002, pp. 694–699.

[45] C. Wang *et al.*, “Incorporating world knowledge to document clustering via heterogeneous information networks,” in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, Sydney, NSW, Australia, 2015, pp. 1215–1224.

[46] I. S. Dhillon, S. Mallela, and D. S. Modha, “Information-theoretic co-clustering,” in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, Washington, DC, USA, 2003, pp. 89–98.

[47] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, “PathSim: Meta path-based top-k similarity search in heterogeneous information networks,” *Proc. VLDB Endowment*, vol. 4, no. 11, pp. 992–1003, 2011.

[48] L. E. Dubins and D. A. Freedman, “Machiavelli and the Gale-Shapley algorithm,” *Amer. Math. Monthly*, vol. 88, no. 7, pp. 485–494, 1981.

[49] J. Chang, J. Boyd-Graber, and D. M. Blei, “Connections between the lines: Augmenting social networks with text,” in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, Paris, France, 2009, pp. 169–178.

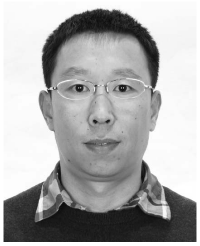
[50] W. Guo, H. Li, H. Ji, and M. Diab, “Linking tweets to news: A framework to enrich short text data in social media,” in *Proc. ACL*, vol. 1. Sofia, Bulgaria, 2013, pp. 239–249.

[51] Wiki. *Anchor Text*. Accessed: Apr. 24, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Anchor_text

[52] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, “New perspectives and methods in link prediction,” in *Proc. SIGKDD*, Washington, DC, USA, 2010, pp. 243–252.

Jianping Cao received the Ph.D. degree from the National University of Defense Technology, Changsha, China, in 2016.

He is currently an Engineer. His research interests include social computing, text analysis, and parallel management theory.



Senzhang Wang received the Ph.D. degree from Beihang Univeristy, Beijing, China, in 2016.

He is currently an Assistant Professor with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. He has published nearly 40 paper in premier conferences and journals in computer science including Special Interest Group (SIG) on Knowledge Discovery and Data Mining, the Association for the Advance of Artificial Intelligence, Society for Industrial and Applied Mathematics International Conference on Data Mining, Association for Computing Machinery (ACM) SIG SPATIAL International Conference on Advances in Geographic Information Systems, The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, *ACM Transactions on Intelligent Systems and Technology*, *Knowledge and Information Systems*, the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, and the IEEE TRANSACTIONS ON MULTIMEDIA. His current research interests include data mining, social computing, big data, and urban computing.

Benxian Li received the Ph.D. degree from the National University of Defense Technology, Changsha, China, in 2013.

He is currently an Associate Professor with the Armed Officer College, CAPF, Chengdu, China. His research interests include social network analysis and data mining.

Xiao Wang (M’16) received the bachelor’s degree in network engineering from the Dalian University of Technology, Dalian, China, in 2011, and the Ph.D. degree in social computing from CASIA, Beijing, China, in 2016.

She is currently a Research Assistant with CASIA, State Key Laboratory of Management and Control for Complex Systems. Her research interests include artificial intelligence, social transportation, cyber movement organizations, and social network analysis.

Zhaoyun Ding received the Ph.D. degree from the National University of Defense Technology (NUDT), Changsha, China, in 2012.

He is currently a Lecturer with NUDT. His research interests include data mining and influence analysis.

Fei-Yue Wang (S’87–M’89–SM’94–F’03) received the Ph.D. degree in computer and systems engineering from Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990.

He joined the University of Arizona, Tucson, AZ, USA, in 1990, and became a Professor and the Director of the Robotics and Automation Laboratory and Program in Advanced Research for Complex Systems. In 1999, he founded the Intelligent Control and Systems Engineering Center, Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China, under the support of the Outstanding Overseas Chinese Talents Program from the State Planning Council and “100 Talent Program” from CAS, and in 2002, where he was appointed as the Director of the Key Laboratory of Complex Systems and Intelligence Science, CAS. From 2006 to 2010, he was the Vice President for Research, Education, and Academic Exchanges, Institute of Automation, CAS. In 2011, he became the State Specially Appointed Expert and the Director of the State Key Laboratory of Management and Control for Complex Systems. His current research interests include methods and applications for parallel systems, social computing, and knowledge automation.

Dr. Wang was a recipient of the National Prize in Natural Sciences of China, the Outstanding Scientist by Association for Computing Machinery (ACM) for his research contributions in intelligent control and social computing in 2007, the IEEE ITS Outstanding Application and Research Awards in 2009, 2011, and 2015, and the IEEE SMC Norbert Wiener Award in 2014. He was the Founding Editor-in-Chief of the *International Journal of Intelligent Control and Systems* from 1995 to 2000, the IEEE INTELLIGENT TRANSPORTATION SYSTEMS MAGAZINE from 2006 to 2007, the Editor-in-Chief of the IEEE INTELLIGENT SYSTEMS from 2009 to 2012 and the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS from 2009 to 2016. He is currently the Editor-in-Chief of the IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, the Founding Editor-in-Chief of the IEEE/CAA JOURNAL OF AUTOMATICA SINICA, and the *Chinese Journal of Command and Control*. Since 1997, he has been served as a General or a Program Chair of over 20 IEEE, INFORMS, ACM, and American Society of Mechanical Engineers (ASME) conferences. He was the President of the IEEE ITS Society from 2005 to 2007, the Chinese Association for Science and Technology, West Windsor, NJ, USA, in 2005, the American Zhu Kezhen Education Foundation, Hangzhou, China, from 2007 to 2008, and the Vice President of the ACM China Council, from 2010 to 2011. Since 2008, he has been the Vice President and the Secretary General of Chinese Association of Automation. He has been elected as a Fellow of International Council on Systems Engineering, International Federation of Automatic Control, ASME, and American Association for the Advancement of Science.