Leadership in Sustainable Infrastructure

Leadership en Infrastructures Durables

Vancouver, Canada

May 31 – June 3, 2017/ *Mai 31 – Juin 3, 2017*

ONTOLOGY-BASED DATA INTEGRATION FOR SUPPORTING BIG BRIDGE DATA ANALYTICS

Liu, Kaijian[1] and El-Gohary, Nora[1,2]

[1] Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, United States

[2] gohary@illinois.edu

Abstract: The deterioration of bridges is dependent on complex interactions of multiple factors. Existing research efforts have focused on predicting bridge deterioration using indicators, which are limited in capturing the many deterioration factors and the interactions between them. On the other hand, a large amount of bridge data is being generated, which opens opportunities to big bridge data analytics for improved bridge deterioration prediction. Such bridge data include: (1) National Bridge Inventory (NBI) and National Bridge Elements (NBE) data, (2) traffic, weather, climate, and natural hazard data, and (3) data from bridge inspection reports. There is, thus, a need for data integration methods that are able to integrate bridge data from multiple sources and in heterogeneous formats. To address this need, this paper proposes an ontology-based data integration methodology. Ontology aims to facilitate the integration based on content and domain-specific meaning. The proposed methodology includes two primary components: (1) ontology-based data linking: identifying the links among data from different sources, and (2) ontology-based data fusion: resolving conflicts between the linked data and then fusing the conflict-resolved linked data. This paper focuses on presenting the proposed ontology-based data linking methodology and its experimental results. Data linking is defined as a multi-class classification problem – classifying data links into multiple types, including "is-type-of", "is-supertype-of", "is-part-of", "is-parent-of", "is-related-to", "is-equivalent-to", and "has-no-match". In developing the methodology, several comparison functions (for comparing the similarities between attribute values) and machine learning algorithms (for the classification of data links) were implemented and evaluated based on accuracy. The experimental results show that the proposed data linking methodology achieved an accuracy of 98.7%.

1 INTRODUCTION

The success of bridge maintenance, repair, and rehabilitation (MR&R) decisions for maintaining and improving bridge conditions largely depends on the ability of predicting future bridge deterioration (Huang 2010, Liu and Madanat 2014). Predicting bridge deterioration, however, is challenging. According to the Federal Highway Administration (FHWA), the deterioration of a bridge is dependent on complex interactions of multiple factors, such as: (1) the original design and the geometrical parameters of the bridge, (2) the previous extents and severities of the deterioration conditions of the bridge, such as cracking and corrosion, (3) the previous MR&R actions, (4) the environmental conditions of climate and natural hazards, and (5) the traffic volumes, frequencies, and types, etc. (FHW 2013).

Existing research efforts (e.g., Bu et al. 2014, Huang 2010) for bridge deterioration prediction, despite their importance, are still limited, because they have focused on predicting bridge deterioration using indicators that are insufficient in capturing the many deterioration factors and the interactions between them. More specifically, such efforts are limited in one or more of the following four ways. First, some efforts developed

prediction models using either only one source of data or one source partially, without taking advantage of data from multiple sources. For example, Bu et al. (2014) developed a backward prediction model for predicting the condition states of bridge elements using condition rating data collected from the Queensland Department of Transportation and Main Roads (QTMR). Second, others focused on predicting the deterioration of a single bridge element, without considering the deterioration of different elements and how they affect the bridge as a whole. For example, Huang (2010) developed an artificial neural network (ANN) model for predicting the condition states of concrete decks using concrete deck condition rating data from the Wisconsin Pontis bridge management system (BMS). Third, and most importantly, none of the existing efforts used any of the data that are buried in inspection reports. These reports contain very rich data about bridge deficiencies (e.g., types, onset times, and severities, etc.) and maintenance actions (e.g., methods, materials, etc.), which are important to utilize when predicting deterioration. Fourth, no deterioration models/methods can integrate (i.e., link and fuse) data of heterogeneous types to help predict deterioration.

On the other hand, a large amount of bridge data is being generated, which opens opportunities to big bridge data analytics for improved bridge deterioration prediction. Such bridge data include: (1) National Bridge Inventory (NBI) and National Bridge Elements (NBE) data, (2) traffic, weather, climate, and natural hazard data, and (3) data from bridge inspection reports. There is, thus, a need for data integration methods that are able to integrate bridge data from distributed sources and in heterogeneous formats. To address this need, the authors propose an ontology-based data integration methodology, which includes two primary components: (1) ontology-based data linking: identifying the links among data from different sources, and (2) ontology-based data fusion: resolving conflicts between the linked data, and then fusing the conflict-resolved linked data. This paper focuses on presenting the proposed ontology-based data linking methodology and its experimental results. In this paper, data linking is defined as a multi-class classification problem – classifying data links into multiple types, including "is-type-of", "is-supertype-of", "is-part-of", "is-parent-of", "is-related-to", "is-equivalent-to", and "has-no-match". In developing the methodology, several comparison functions (for comparing the similarities between attribute values) and machine learning (ML) algorithms (for classifying data links into the data link types) were implemented and evaluated based on accuracy. In the following sections, this paper presents the proposed ontology-based data integration approach, and then discusses the proposed ontology-based data linking methodology and its experimental results in more detail.

2 BACKGROUND

Data integration aims to link and fuse data residing at different data sources, and provide a reconciled, integrated, yet concise view of these data (Lenzerini 2002). Generally, a data integration system encompasses three primary components: schema mapping, data linking, and data fusion (Naumann et al. 2006, Bleihoder and Naumann 2009). Schema mapping aims to conduct the mapping and resolve inconsistencies at the schema level by identifying the corresponding classes and properties that are used in the different schemas that are being mapped. Data linking and data fusion, together, aim to conduct the mapping and resolve inconsistences at the tuple and value level. Data linking aims to link the data entries that follow certain link types (e.g., "is-equivalent-to"), while data fusion aims to identify and resolve the conflicts between the data values of the linked data (Naumann et al. 2006).

Data linking [also known as entity resolution, record linkage, de-duplication, data association (Singla and Domingos 2006)] is a critical step in data integration because it directly affects the performance of the subsequent data fusion step and the entire data integration system. For example, linking two data entries that should not be linked will likely result in subsequent errors in data fusion and, thus, affect the performance of the entire data integration system. Data linking has been previously studied in the computer science domain. For example, Cochinwala et al. (2001) proposed a data linking model using a decision tree algorithm. Bilenko and Mooney (2003) applied support vector machines (SVM) for linking data. Christen (2008) compared an SVM algorithm with a nearest neighbors (NN) algorithm for supporting data linking.

Despite the achievements of the efforts, three main gaps of knowledge in the area of data linking have not been well addressed. First, there is a lack of studies that compare the performance of different ML algorithms for supporting data linking, especially for supporting bridge data linking. Second, there is a lack

of efforts that identify the semantic types of the data links (e.g., the link between data describing a “wearing surface” and data describing “deck” is an “is-part-of” link). For example, the above-mentioned data linking models treated data linking as a binary classification problem that aims to classify data links into “match” and “non-match” only. Identifying the semantic types of the data links is essential to further determine how the linked data will be fused. Third, and most importantly, there is a lack of studies that use ontology for facilitating data linking. Ontology has been used in many ML tasks and has been proven to be able to improve the performance of ML. However, using ontology for supporting data linking, especially for supporting domain-specific data linking tasks, has not been well studied. To address these gaps, this paper proposes an ontology-based data integration methodology. In the following sections, the authors first present the proposed ontology-based data integration approach, and then focus on discussing the proposed ontology-based data linking methodology and its experimental results in more detail.

3 PROPOSED ONTOLOGY-BASED DATA INTEGRATION APPROACH

The proposed ontology-based data integration approach is illustrated in Figure 1. The proposed approach relies on three primary components, which are introduced in each of the following subsections: automated information extraction, global schema, and ontology-based data linking and fusion.

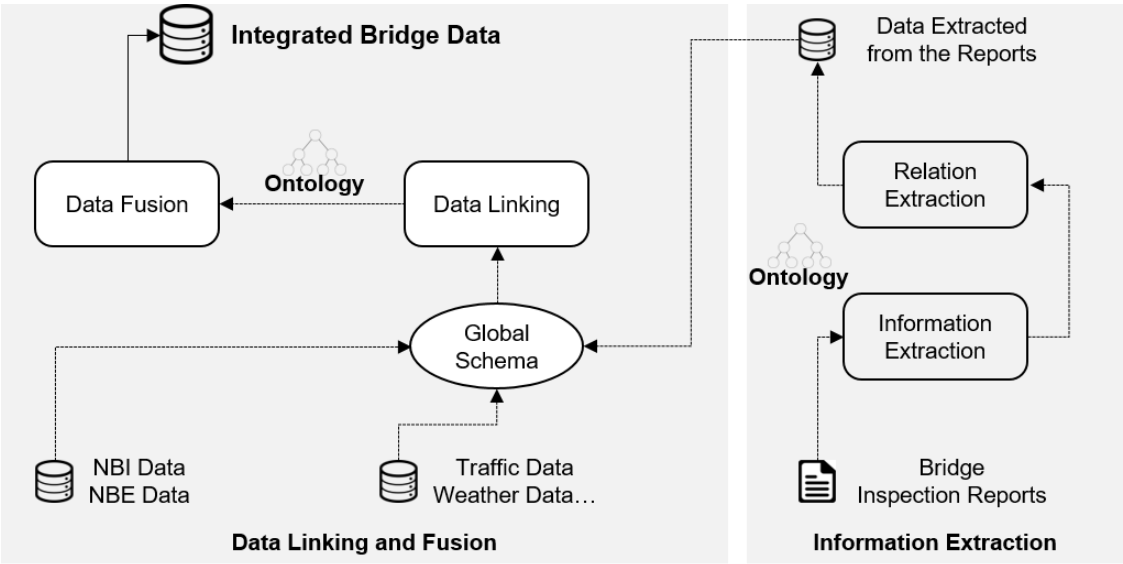

Figure 1: Proposed ontology-based data integration methodology

3.1 Automated Information Extraction

Automated information extraction (IE) aims to extract information about bridge conditions and maintenance actions from bridge inspection reports, and represent the extracted information in a structured way. The unstructured textual data that are buried in the reports provide detailed descriptions about bridge conditions and maintenance history. These data are, thus, expected to improve bridge deterioration prediction. However, because of the unstructured nature of the textual data, IE methods are needed to extract the target information from the reports, and represent the extracted information in a structured format. Subsequently, the extracted information can then be integrated with the other structured data/information (e.g., NBI and NBE data) for supporting improved bridge deterioration prediction. In the proposed approach, automated IE from bridge inspection reports includes two main steps. First, the following information entities are recognized and extracted using an ontology-based semi-supervised conditional random fields (CRF) IE methodology (Liu and El-Gohary 2017a). The information entities that need to be extracted include: <bridge element>, <deficiency>, <deficiency cause>, <maintenance action>, <maintenance material>, <numerical measure>, <numerical measure unit>, <categorical quantity measure>, <categorical severity measure>, and <date>. Second, the dependency relations between the extracted information entities are

analyzed using an ontology-based, similarity-based dependency parsing (DP) methodology (Liu and El-Gohary 2017b). The DP methodology aims to link the extracted, yet isolated, information entities into concepts, and represent the concepts into structured semantic information sets (SISs). The extracted, structured data will be linked to other structured bridge data (e.g., NBI and NBE data) under a global schema.

3.2 Global Schema

A data schema defines how data are organized and structured. Each type of bridge data – from a different source – is organized by a different local schema, depending on the data source. For example, the NBI data schema uses 140 attributes (e.g., condition ratings of deck, superstructure, and substructure) to organize bridge-level inspection data. The schema of the extracted data from the reports, as introduced above, uses 10 attributes to organize element-level inspection data. A global schema that provides a reconciled, integrated, yet concise view of the bridge data is needed for supporting the integration of the data. More importantly, the global schema should also be able to capture the semantics of the data links. To this end, the authors propose a heterogeneous information network (HIN)-based global schema. An HIN-based global schema is proposed because it allows to capture the rich, interrelated semantic information instances and their semantic links (Han et al. 2012). As shown in Figure 2, the proposed HIN-based global schema consists of nodes for representing attributes from different local schemas and edges for representing the relationships (i.e., semantic links) between nodes.

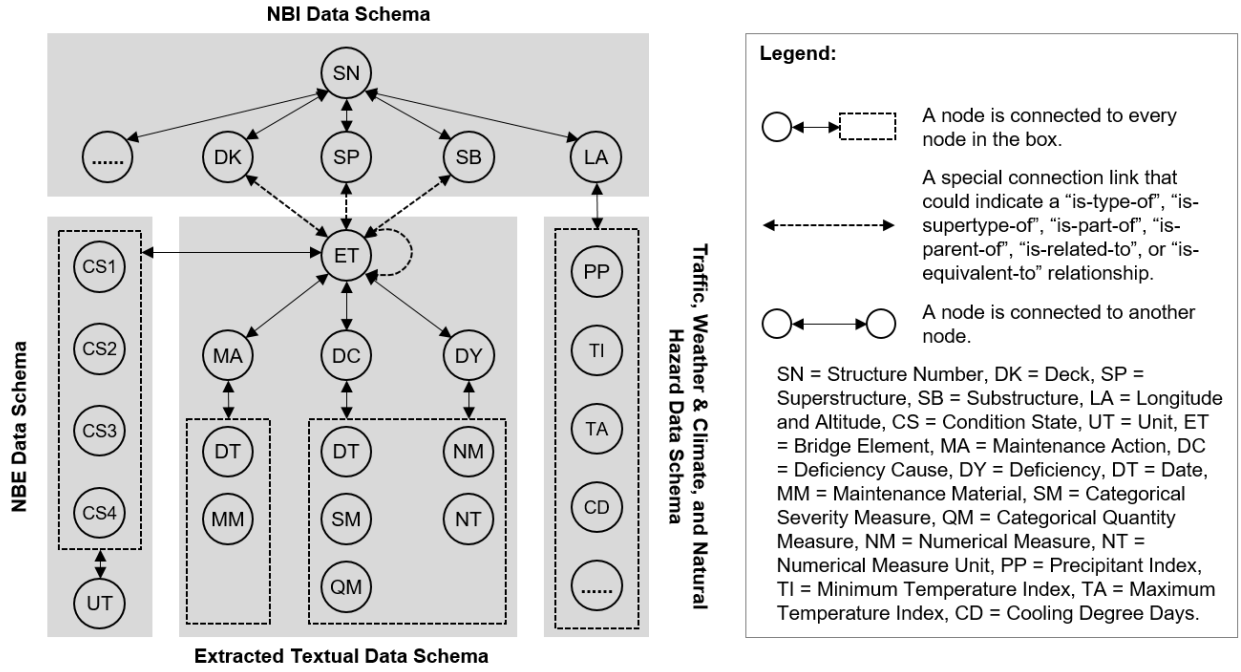

Figure 2: Proposed global schema

3.3 Ontology-Based Data Linking and Fusion

Ontology-based data linking and data fusion are at the cornerstone of the proposed data integration approach. Ontology aims to facilitate the integration based on content and domain-specific meaning. Ontology-based data linking aims to identify the links among data from different sources. Ontology-based data fusion aims to resolve conflicts between the linked data and then fuse conflict-resolved linked data. In the following sections, this paper focuses on presenting the proposed ontology-based data linking methodology and its experimental results.

4 PROPOSED ONTOLOGY-BASED DATA LINKING METHODOLOGY

The proposed ontology-based data linking methodology aims to identify the links among data from different sources. For a pair of data to be linked, the proposed methodology aims to classify the link of the data pair into one of the following types, including: (1) *is-type-of* (or *is-supertype-of*) link, where the subject datum is a subtype (or supertype) of the datum in comparison, (2) *is-part-of* (or *is-parent-of*) link, where the subject datum is a part (or whole) of the datum in comparison, (3) *is-related-to* link, where the subject datum has a non-hierarchical relation with the datum in comparison, (4) *is-equivalent-to* link, where the subject datum is same/equivalent as the datum in comparison, and (5) *has-no-match* link, where none of the above-mentioned links apply (i.e., a match was not found). To achieve this goal, the proposed ontology-based data linking methodology is composed of two main components: (1) ontology-based feature representation, and (2) ML-based link classification. In the following subsections, each component is introduced in more detail.

4.1 Ontology-Based Feature Representation

As shown in Figure 3, the authors propose a new ontology-based feature representation for representing data that need to be linked. The proposed representation works as follows. First, for a data pair to be linked, each pair of attribute values (e.g., $A1_A$ and $A1_B$ for data entry A and B in Figure 3) is compared using a comparison function (CF). Second, the CF returns a similarity value (e.g., S1 in Figure 3) indicating to what extent the attribute values are similar. Third, the previous two steps are repeated until all corresponding attribute value pairs are compared. Finally, based on a bridge deterioration knowledge ontology (Liu and El-Gohary 2016), the ontology-based (i.e., semantic) features of the data pair are generated (i.e., O1-5 in Figure 3). In the proposed representation, O1 to O5 are binary (with 1 being true and 0 being false) ontology-based features of data pair A and B, representing is-type-of, is-supertype-of, is-part-of, is-parent-of, and is-equivalent-to relationships between data entries A and B, respectively. The proposed representation, thus, includes a set of similarity values (measured by a CF) and ontology-based binary features. In selecting the CF, a set of potential CFs were tested and evaluated based on accuracy. The details about the tested CFs are presented in Table 1, and the evaluation results are presented and discussed in Section 5.

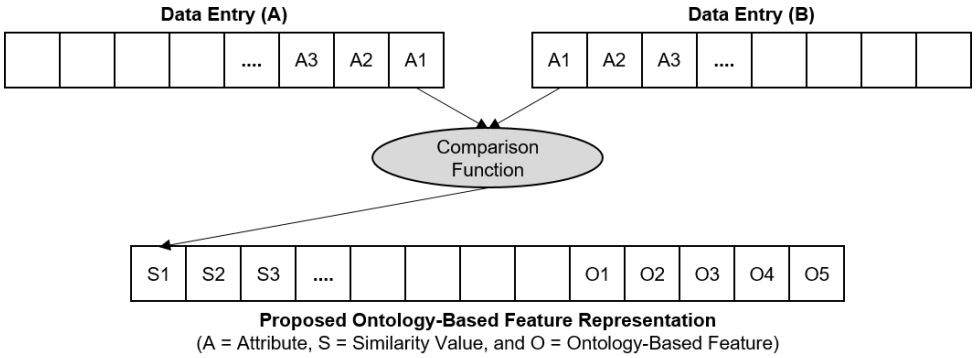

Figure 3: Proposed ontology-based feature representation

4.2 Machine Learning-Based Data Linking

In order to select which ML algorithm to use for classifying the data links into semantic types, five commonly-used ML algorithms were implemented and tested, including support vector machines (SVM), logistic regression (LR), Naïve Bayes (NB), decision tree, and nearest neighbors. SVM is a large margin theory-based ML algorithm that maps input data to a high-dimensional feature space using kernel functions and constructs a linear decision hyper plane in the feature space to classify data (even those non-linearly separable ones in original feature space) into different categories (Cortes and Vapnik 1995). In this paper, a set of SVM algorithms [with different kernel functions, including Polynomial, Radial Basis Function (RBF), and Sigmoid kernels] were implemented and tested. LR and NB are probabilistic models for ML. LR aims to learn a discriminative classifier that directly models the conditional probability of a label (i.e., a predicted class) given input data, using a logistic function (Ng and Jordan 2002). Similar to many other ML-based

classification algorithms, the performance of LR can be improved by using regularization, either with L1 or L2 norm (Ng and Jordan 2002). In this paper, LR algorithms with L1 and L2 regularizations were implemented and tested. NB, on the other hand, aims to learn a generative classifier that models the joint probability of a label and inputs and assumes independence between features of inputs to indirectly compute the probability of a label given inputs (Murphy 2006). Decision tree is a ML algorithm that aims to find a set of decision rules in a decision tree that can recursively split independent features into homogeneous zones and eventually reach a final decision that predicts a label for an input (Pradhan 2013). Nearest neighbors is an instance-based ML algorithm that predicts a label for an input based on the majority labels of its k-nearest neighbors (Weinberger and Saul 2009).

The performance of the above-mentioned ML algorithms in data linking was evaluated using accuracy. Accuracy, here, is the percentage of the number of correctly-classified links out of the total number of links. The accuracy was calculated by comparing the algorithm-predicted link types with the link types in a manually-developed gold standard. The evaluation results are discussed in Section 5.

Table 1: Selected comparison functions [a]

No.	Comparison function [b]	Description
CF1	Extract string comparison	It returns 1 if the two attribute values are exactly same; otherwise, returns 0.
CF2	Extract string comparison (beginning)	It returns 1 if the first 3 characters of the two attribute values are exactly same; otherwise, returns 0.
CF3	Extract string comparison (end)	It returns 1 if the last 3 characters of the two attribute values are exactly same; otherwise, returns 0.
CF4	Levenshtein edit distance string comparison	It measures the smallest number of operations (insertion, deletion, and substitution) needed to convert an attribute value to another. The number is normalized by the maximum string length of the two attribute values.
CF5	Smith-Waterman edit distance string comparison	Similar to CF4, it defines the costs for five operations (exact match, approximate match, mismatch, gap start penalty, and gap continuation penalty) for two characters. The total cost is normalized by the average length of the two attribute values.
CF6	Bigram-based string comparison (overlap coefficient)	It measures the number of common bigram characters in the two attribute values. The number is normalized by the minimum number of bigram characters of the two attribute values.
CF7	Bigram-based string comparison (Jaccard coefficient)	It measures the number of common bigram characters in the two attribute values. The number is normalized by the total number of bigram characters of the two attribute values minus the number of common bigrams.
CF8	Bigram-based string comparison (dice coefficient)	It measures the number of common bigram characters in the two attribute values. The number is normalized by the total number of bigram characters of the two attribute values.
CF9	Jaro string comparison	It measures the number of common bigrams and the number of transpositions in common bigrams of the two attribute values.
CF10	Winkler string comparison	As a modified version of CF9, it also considers the number of same characters at the beginning of the two attribute values.

[a] CFs are used for measuring the similarities between attribute values whose type is string. For the attribute values whose type is numeric (e.g., numerical measure and date in the proposed schema), their similarities are measured by exact comparison, with “1” being same and “0” being different.
[b] CFs are selected based on the review of Bleihoder and Naumann (2009) and Christen (2012). For detailed explanations of the CFs, the readers are referred to these references.

5 EXPERIMENTAL RESULTS AND ANALYSIS

In this paper, the authors created a dataset to evaluate the performance of the proposed ontology-based data linking methodology. The dataset contains the 2006 NBI data and the data from the 2006 bridge inspection report of the I-35W Mississippi River Bridge. In creating the dataset, first, a total of 400 data instances were extracted [using an ontology-based semi-supervised CRF-based IE methodology (Liu and El-Gohary 2017a)] and were represented in a structured way [using an ontology-based similarity-based DP methodology (Liu and El-Gohary 2017b)]. Then, the 2006 NBI data of the bridge were collected from the FHWA. The gold standard links between the collected data were, then, manually annotated. Examples of the links between the collected data are shown in Table 2. The performance results of the proposed data linking methodology are summarized in Tables 3 and 4. The comparison of the performances of the nonontology-based and ontology-based data linking approaches is shown in Tables 5 and 6.

Table 2: Examples of the links between the collected data [a]

No.	Data extracted from bridge inspection report (partial) [b]							
	ET	DY	NM	NU	SM	QM	DT	…
D1	Concrete overlay	Transverse crack	3,000	LF	N/A	N/A	2006	
D2	Overlay	Patched area	N/A	N/A	Minor	Some	2006	
D3	Deck	Transverse leaching crack	N/A	N/A	Moderate	N/A	N/A	…
D4	Curb	Crack	N/A	N/A	Moderate	N/A	N/A	
D5	Overlay	Patched area	N/A	N/A	Minor	Some	2006	

[a] D1 is linked to D2 through an "*is-type-of*" link. D2 and D4 are linked to D3 through "*is-part-of*" links. D2 is also linked to D4 through an "*is-related-to*" link because they are parts of D3. D5 is linked to D2 through an "*is-equivalent-to*" link because they are same.
[b] The indexes follow those defined in Figure 2.

Table 3 summarizes the performance of the proposed nonontology-based data linking methodology, using different ML algorithms and CFs. As shown in Table 3, the bigram-based string CFs (i.e., CF6 to CF8), compared to the other CFs, achieved the highest accuracy (i.e., 89.3%). This indicates that the bigram-based string CFs are more effective in capturing the similarities between the attribute values for classifying data links into the defined link types. Table 3 also shows that the SVM with RBF kernel and the LR with L1 regularization, compared to the other ML algorithms, achieved the highest accuracy (i.e., 89.3%). This indicates that SVM and LR are better at capturing the distributions of the input data for data linking.

Table 3: Accuracy of the nonontology-based data linking methodology using different machine learning algorithms and comparison functions

CF [a]	Machine learning algorithms							
	SVM			LR		DT	NN	NB
	Polynomial	RBF	Sigmoid	L1	L2			
CF1	82.7%	87.3%	87.3%	88.0%	87.3%	82.0%	87.3%	87.3%
CF2	82.7%	88.3%	88.3%	89.0%	87.3%	83.7%	87.3%	87.0%
CF3	82.7%	87.3%	87.3%	88.0%	87.3%	82.0%	87.3%	87.3%
CF4	82.7%	82.7%	82.7%	85.3%	78.7%	67.0%	79.3%	82.7%
CF5	82.7%	84.3%	84.3%	84.3%	84.3%	70.0%	83.7%	79.0%
CF6	82.7%	**89.3%**	87.3%	**89.3%**	87.3%	77.7%	84.0%	82.7%
CF7	82.7%	**89.3%**	87.3%	**89.3%**	87.3%	76.7%	84.0%	83.3%
CF8	82.7%	**89.3%**	87.3%	**89.3%**	87.3%	76.7%	84.0%	82.7%
CF9	82.7%	86.3%	87.3%	85.3%	87.3%	78.3%	86.0%	82.7%
CF10	82.7%	86.3%	87.3%	85.3%	87.3%	77.3%	76.0%	82.7%

[a] CF = comparison function, and the CF index follows that in Table 1.

Table 4 shows the performance of the proposed ontology-based data linking methodology, using different ML algorithms and CFs. As seen, CF4, CF9, and CF10 achieved the highest performance (i.e., an accuracy of 98.7%) compared to the other CFs. The performance result of CF4 is somewhat inconsistent with the results obtained when only using nonontology-based features (shown in Table 3). In the nonontology-based case, CF4 did not achieve an accuracy comparable to the highest. This could be attributed to the introduction of ontology-based features; the introduced features changed the entire feature space, resulting in CF4 becoming more informative. On the other hand, the performance results of CF9 and CF10 (which are bigram-based CFs) are consistent with the results in Table 3, where the bigram-based CFs outperformed the others. Also, similar to the nonontology-based case, LR with L1 and L2 regularizations achieved the highest accuracy for the ontology-based methodology. This further indicates the better capability of LR in supporting bridge data linking. However, although SVM did not achieve the best performance in this case, it only performed slightly worse than LR with an accuracy of 98.0%.

Table 4: Accuracy of the ontology-based data linking methodology using different machine learning algorithms and comparison functions

CF [a]	Machine learning algorithms							
	SVM			LR		DT	NN	NB
	Polynomial	RBF	Sigmoid	L1	L2			
CF1	82.7%	96.0%	91.3%	98.0%	96.3%	87.7%	87.3%	90.3%
CF2	82.7%	97.3%	91.0%	96.7%	98.0%	89.0%	88.0%	90.3%
CF3	82.7%	96.0%	89.7%	94.0%	98.0%	87.7%	87.3%	90.7%
CF4	82.7%	98.3%	89.7%	**98.7%**	**98.7%**	88.0%	88.0%	90.3%
CF5	82.7%	89.3%	84.3%	90.3%	89.3%	84.3%	84.0%	89.3%
CF6	82.7%	98.0%	91.0%	98.0%	98.0%	87.7%	87.3%	90.0%
CF7	82.7%	98.0%	91.0%	95.3%	98.0%	87.7%	87.3%	90.3%
CF8	82.7%	98.0%	91.0%	96.3%	98.0%	87.7%	87.3%	90.0%
CF9	82.7%	97.3%	91.0%	96.7%	**98.7%**	87.3%	88.3%	89.7%
CF10	82.7%	97.3%	91.0%	93.3%	**98.7%**	87.3%	88.3%	89.7%

[a] CF = comparison function, and the CF index follows that in Table 1.

Table 5: Accuracy of the nonontology-based and ontology-based data linking methodologies across the comparison functions, using the eight tested machine learning algorithms

CF [a]	Nonontology-based [b]		Ontology-based [b]	
	Mean	Standard deviation	Mean	Standard deviation
CF1	86.2%	0.024	91.2%	0.053
CF2	86.7%	0.023	91.6%	0.054
CF3	86.2%	0.024	90.8%	0.050
CF4	80.1%	0.057	91.8%	0.061
CF5	81.6%	0.050	86.7%	0.031
CF6	85.0%	0.040	91.6%	0.059
CF7	85.0%	0.043	91.3%	0.055
CF8	84.9%	0.043	91.4%	0.056
CF9	84.5%	0.031	91.5%	0.056
CF10	83.1%	0.044	91.0%	0.053
Average	**84.3%**	**0.038**	**90.9%**	**0.053**

[a] CF = comparison function, and the CF index follows that in Table 1.
[b] Mean is the average of the accuracies using the eight machine learning algorithms in Table 3.

Table 5 compares the performance of the nonontology-based and ontology-based data linking methodologies across the comparison functions, using the eight tested machine learning algorithms; and Table 6 shows the comparison across the machine learning algorithms, using the ten tested comparison functions. As seen from both tables, the nonontology-based method achieved an average accuracy of 84.3% and the ontology-based one achieved an average accuracy of 90.9%. This result, along with the rest of the results presented in Tables 5 and 6, show that the ontology-based data linking methodology outperforms the nonontology-based one. Table 6 also indicates that the use of ontology-based features makes the ML algorithms more stable across different CFs, because the average standard deviation (SD) across the CFs when considering ontology-based features (i.e., SD = 0.016) is smaller than that when not considering such features (i.e., SD = 0.025). However, Table 5 suggests that the use of ontology-based features makes the CFs less stable across different ML algorithms, because the average SD across the ML algorithms when considering ontology-based features (i.e., SD = 0.053) is larger than that when not considering such features (i.e., SD = 0.038).

Table 6: Accuracy of the nonontology-based and ontology-based data linking methodologies across the machine learning algorithms, using the ten tested comparison functions

ML [a]	Nonontology-based [b]		Ontology-based [b]	
	Mean	Standard deviation	Mean	Standard deviation
ML1	82.7%	0.000	82.7%	0.000
ML2	87.1%	0.022	96.6%	0.027
ML3	86.7%	0.017	90.1%	0.021
ML4	87.3%	0.020	95.7%	0.026
ML5	86.2%	0.028	97.2%	0.028
ML6	77.1%	0.052	87.4%	0.012
ML7	83.9%	0.037	87.3%	0.013
ML8	83.7%	0.027	90.1%	0.004
Average	**84.3%**	**0.025**	**90.9%**	**0.016**

[a] ML index follows the ML algorithm sequence in Table 3.
[b] Mean is the average of the accuracies using the ten comparison functions in Table 1.

6 CONCLUSION AND FUTURE WORK

In this paper, the authors proposed an ontology-based data integration methodology to integrate bridge data from multiple sources and in heterogeneous formats for supporting improved bridge deterioration prediction. Such bridge data include: (1) National Bridge Inventory (NBI) and National Bridge Elements (NBE) data, (2) traffic, weather, climate, and natural hazard data, and (3) data from bridge inspection reports. The proposed ontology-based data integration methodology includes three components: (1) ontology-based information extraction, (2) heterogeneous information network-based global schema, and (3) ontology-based data linking and fusion. Ontology-based data linking and fusion are at the cornerstone of the proposed methodology, which aim to identify the links among data from different sources, resolve conflicts between the linked data, and then fuse the conflict-resolved linked data.

This paper focused on presenting the proposed ontology-based data linking methodology. The proposed methodology achieved a data linking accuracy of 98.7%. In developing the methodology, a set of comparison functions (CFs) (for measuring the similarities between attribute values) and machine learning algorithms (for classifying data into different link types) were implemented and tested. The following conclusions were drawn from the experimental results: (1) bigram-based string CFs (e.g., bigram-based string comparison using overlap, Jaccard, and dice coefficients, Jaro string comparison, and Winkler string comparison) perform better in capturing the similarities between attribute values, (2) logistic regressions with L1 and L2 regularizations and support vector machines with RBF kernel function perform better in classifying data links into the defined types, and (3) most importantly, ontology-based features can facilitate

data linking. Compared to the nonontology-based data linking approach, the ontology-based approach improves data linking accuracy by 6.6% on average.

In their future work, the authors will test the proposed ontology-based data linking methodology using a larger dataset, and will further improve the methodology based on the testing results. Also, an ontology-based data fusion methodology will be developed to resolve conflicts between the linked data. These efforts will eventually enable the utilization of integrated bridge data for better supporting improved bridge deterioration prediction.

Acknowledgements

This material is based upon work supported by the Strategic Research Initiatives (SRI) Program by the College of Engineering at the University of Illinois at Urbana-Champaign.

References

Bilenko, M. and Mooney, R.J. 2003. Adaptive Duplicate Detection Using Learnable String Similarity Measures. *9th Intl. conf. on knowledge discovery and data mining*, Washington, DC, 39-48.

Bleiholder, J. and Naumann, F. 2009. Data Fusion. *ACM Computing Surveys*, **41**(1): 1-41.

Bu, G., Lee, J., Guan, H., Blumenstein, M. and Loo, Y. 2014. Development of an Integrated Method for Probabilistic Bridge-Deterioration Modeling. *J. of Performance of Constructed Facilities*, **28**(2): 330-340.

Christen, P. 2008. Automatic Record Linkage Using Seeded Nearest Neighbor and Support Vector Machine Classification. *14th Intl. conf. on knowledge discovery and data mining*, Las Vegas, NV, 151-159.

Christen, P. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer Science & Business Media, Berlin, Germany.

Cochinwala, M., Kurien, V., Lalk, G. and Shasha, D. 2001. Efficient Data Reconciliation. *Information Sciences*, **137**(1): 1-15.

Cortes, C. and Vapnik, V. 1995. Support-Vector Networks. *Machine Learning*, **20**(3): 273-297.

FHWA (Federal Highway Administration). 2013. LTBP Bridge Performance Primer. *No. FHWA-HRT-13-051*, Washington, DC.

Han, J., Sun, Y., Yan, X. and Philip, S.Y. 2012. Mining Knowledge from Data: An Information Network Analysis Approach. *28th Intl. conf. on data engineering*, Arlington, VA, 1214-1217.

Huang, Y. 2010. Artificial Neural Network Model of Bridge Deterioration. *J. of Performance of Constructed Facilities*, **24**(6): 597-602.

Lenzerini, M. 2002. Data Integration: A Theoretical Perspective. *21st ACM symposium on principles of database systems*, Madison, WI, 233-246.

Liu, H. and Madanat, S. 2015. Adaptive Optimisation Methods in System-Level Bridge Management. *Structure and Infrastructure Engineering*, **11**(7): 884-896.

Liu, K. and El-Gohary, N. 2016. Semantic Modeling of Bridge Deterioration Knowledge for Supporting Big Bridge Data Analytics. *2016 Construction research congress*, San Juan, Puerto Rico, 930-939.

Liu, K. and El-Gohary, N. 2017. Ontology-Based Semi-Supervised Conditional Random Fields for Automated Information Extraction from Bridge Inspection Reports. *Automation in Construction*, In press.

Liu, K. and El-Gohary, N. 2017. Similarity-Based Dependency Parsing for Extracting Dependency Relations from Bridge Inspection Reports, *2017 Intl. workshop on computing in civil engineering*, Seattle, WA.

Murphy, K.P. 2006. Naive Bayes Classifiers. University of British Columbia.

Naumann, F., Bilke, A., Bleiholder, J. and Weis, M. 2006. Data Fusion in Three Steps: Resolving Schema, Tuple, and Value Inconsistencies. *IEEE Data Engineering Bulletin*, **29**(2): 21-31.

Ng, A.Y. and Jordan, M.I. 2002. On Discriminative Vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes. *Advances in Neural Information Processing Systems*, **2**(2002): 841-848.

Pradhan, B. 2013. A Comparative Study on the Predictive Ability of the Decision Tree, Support Vector Machine and Neuro-Fuzzy Models in Landslide Susceptibility Mapping Using GIS. *Computers & Geosciences*, **51**(2013): 350-365.

Singla, P. and Domingos, P. 2006. Entity Resolution with Markov Logic. *6th Intl. conf. on data mining*, Hong Kong, China, 572-589.

Weinberger, K.Q. and Saul, L.K. 2009. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *J.of Machine Learning Research*, **10**(2): 207-244.