Journal of Data Intelligence, Vol. 1, No. 3 (2020) 333–350

KNOWLEDGE GRAPH COMPLETION TO SOLVE UNIVERSITY CAMPUS ISSUES

YUTO TSUKAGOSHI
Graduate School of Informatics and Engineering,
The University of Electro-Communications,
1-5-1 Chofugaoka, Chofu, Tokyo, Japan
tsukagoshi.yuto@ohsuga.lab.uec.ac.jp

TAKAHIRO KAWAMURA
National Agriculture and Food Research Organization,
3-1-1 Kannondai, Tsukuba, Ibaraki
takahiro.kawamura@affrc.go.jp

YUICHI SEI
Graduate School of Informatics and Engineering,
The University of Electro-Communications,
1-5-1 Chofugaoka, Chofu, Tokyo, Japan
seiuny@uec.ac.jp

YASUYUKI TAHARA
Graduate School of Informatics and Engineering,
The University of Electro-Communications,
1-5-1 Chofugaoka, Chofu, Tokyo, Japan
tahara@uec.ac.jp

AKIHIKO OHSUGA
Graduate School of Informatics and Engineering,
The University of Electro-Communications,
1-5-1 Chofugaoka, Chofu, Tokyo, Japan
ohsuga@uec.ac.jp

A number of urban challenges are encountered by modern societies. Governments, businesses and public bodies need to make statistical data widely available in order to tackle these challenges. Nonetheless, current literature and data are problematic; they have inaccuracies which lead to less effective methods of resolving these issues. This research aims to solve this challenge by thinking of a university campus as a microcosm of society, implementing a data integration schema, and combining data into a knowledge graph. Existing completion methods will then be applied and updated. Especially in regards to bicycle environment, our knowledge graph was tailored and evaluated in line with conventional methods, and secondly with our proposed derivative methods. Roughly 650 pieces of parking data, with various dates and times, was contrasted with each time's mean absolute error. Our approach accurately projected 54.5 more bicycles than the conventional method.

Keywords: Knowledge graph, Linked open data, Knowledge graph completion, Translation-based model

1. Introduction

1.1. *Background*

Modern societies are challenged with a number of urban problems such as littering, vandalism, and illegally parked bicycles. Governments, businesses and public bodies need to make statistical data widely available in order to tackle these challenges. More precisely, these bodies should publish information revealing the level of damage these issues have on each area, how capable they are of tackling these issues, and figures recording the number of trash cans and street lights in the region, which may show how remedies are being applied. Data-driven countermeasures have become increasingly popular, and administrations and governments have allowed masses of helpful open data to become accessible online [7].

The U.S. Congress passed the Open Government Data Act on December 22, 2018. This Act instructs all government bodies to publish non-confidential information online in a universal machine-readable format. Other countries such as France and Germany are also attempting to move their government's open data online. The French government publishes open data such as budgets, spending, statistics, environmental and geographical information on .gouv.fr. The "Open Data Watch" has revealed that lately, countries are more devoted to ensuring that their data is more accessible (Fig 1).

The suggested format for the amalgamation of data, or to make open data more beneficial, is Linked Open Data (LOD). Projected by Tim Berners-Lee, father of the World Wide Web and creator of the 5-star rating method, LOD is the chief 5-star open data plan. LOD also uses an impressive mixture of linked data and open data; data is connected and uses open sources [2]. A 'triple' is a semantic connection used by Linked data. The triple is made up of a subject, predicate, and object and is organized based on the Resource Description Framework (RDF). The RDF is an all-purpose approach for demonstrating metadata and has been used in a prototype for RSS, which supplies websites with updated information. Generally, linked data is also known as a knowledge graph; a data management system that demonstrates the links between information and is created using the same technology. Lately, there has been a significant rise in the production and availability of knowledge graphs, such as DBpedia, Freebase and YAGO [6]. Adapting linked data into open data in this way means that these knowledge links can be published on the Web and modified into a secondary practical form, allowing anybody to use the data for additional purposes [16].

Globally, LOD is an increasingly popular choice for open data. A number of countries collect data using LOD and publish it onto the knowledge base, called LOD Cloud [3]. Figures in March 2019 shows that the LOD Cloud has 1,239 data sets with 16,147 links.

Moreover, by using the semantic assembly of data, a characteristic of knowledge graphs, new approaches for predicting absent data in graphs have been established. There are a variety of methods; however, H. Mousselly-Sergieh et al. [4] claims that translation-based tactics are particularly effective in the present day. H. Yoon et al. [5] explains that inserting graphs into a low-dimensional continuous vector space is the most successful approach for to finalizing knowledge graphs. In addition, one of the most recognized translation-based methods for these challenges is TransE [19].

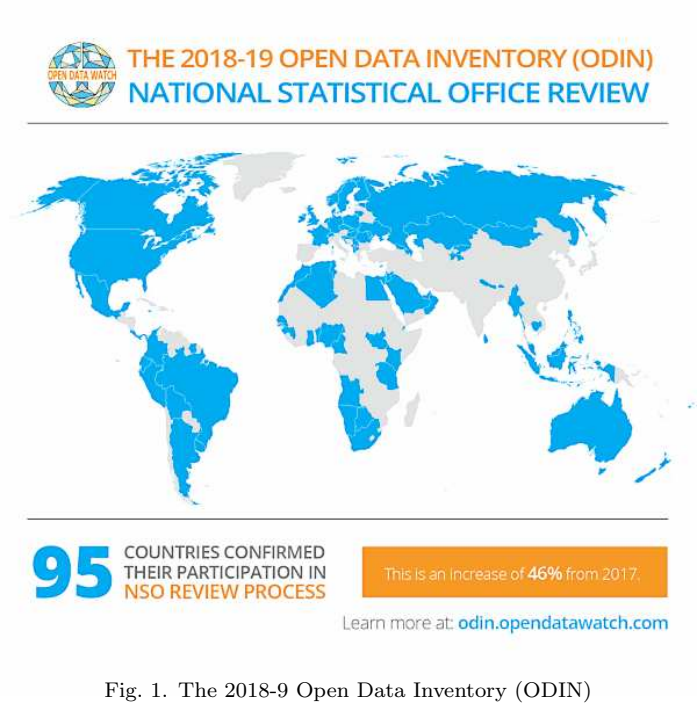

Fig. 1. The 2018-9 Open Data Inventory (ODIN)
Open Data Watch[15]

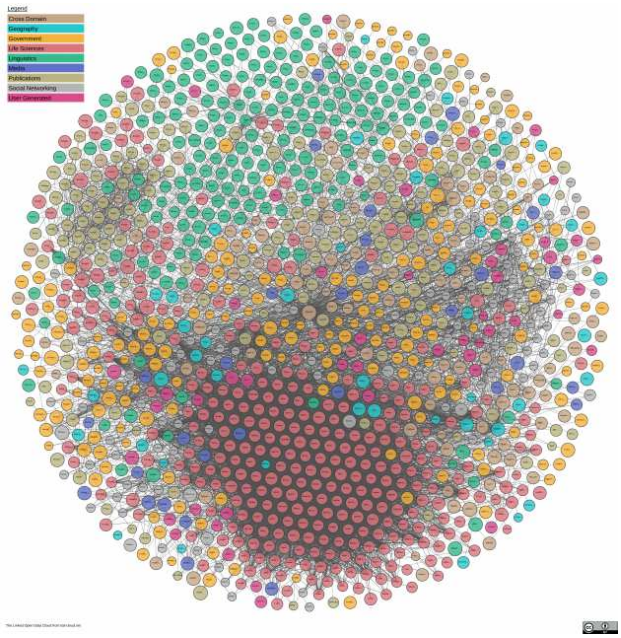

Fig. 2. The Linked Open Data Cloud
Insight Centre for Data Analytics[3]

Table 1. Format of data collected or published by the university

Source of Data	Data	Type of Format
SA	The number of bicycles	xls
SA	Parking areas	PDF
University	Time table (course titles, and classrooms)	PDF
University	Seating capacity	html
University	Name of rooms in every building	PDF
University	Event titles and venues	html

1.2. *Issues and Purpose of This Research*

An indefinite amount of people make used of Japan's University of Electro-Communications campus each day. The campus can therefore be thought of as a microcosm of society. A number of social challenges, especially bicycle use, are faced by the university and the neighboring community. Around 60% of students commute on a bike to the densely populated campus, which has over 30 bicycle parking stations, with a usage rate of over 130% at peak times, and over 1,200 parking spaces. The etiquette of individuals in the parking stations is complained about by over 40% of students [18]. Therefore, there are some significant issues that need to be addressed. Another issue involves the main parking station, as it is the station with excessive crowding at particular times and days. Due to this, the use of bikes on campus poses a hazard for the evacuation route in event of an emergency, and has an effect on the campus landscape. The student organization, Student Assistants (SA), observe the bicycle parking lot and record the amount of vehicles present, to tackle these issues. Unfortunately, they have time and staff limitations, so they are unable to conduct surveys on the university's bicycle parking stations (Fig 3). Their data therefore omits some of the parking stations and has some insufficient data. The general understanding of the parking lot use is therefore limited, and any countermeasures are hindered. It is anticipated that the numerous data published by the university will guide resolution of these issues. Nonetheless, the data format is not clear or uniform, and is challenging to use (Table 1).

By constructing a knowledge graph where data is uniform, this research has created an environment that facilitates us to solve these issues in various ways. The graph shows the number of bicycles in each parking area alongside additional relevant data of the campus. The research also links data to the bicycle numbers previously recorded in a LOD format, which is part of the aforementioned open data strategy and facilitates everybody in the creation of ideas to tackle the issues on campus. In addition, we have introduced an amended approach, founded on current completion methods such as TransE [19], to create knowledge graphs and predict the lacking data and defect values of the SA regarding the number of bicycles parked. Lastly, the knowledge graph will show, compare and assess the accuracy of these predicted values and the methods used. This in turn, will allow a more beneficial graph to be constructed for the campus, which can be used to create countermeasures for the parking station issues.

Section 2 presents a selection of relevant research and previously used completion approaches. A knowledge graph is then created based on the amount of bicycles and additional

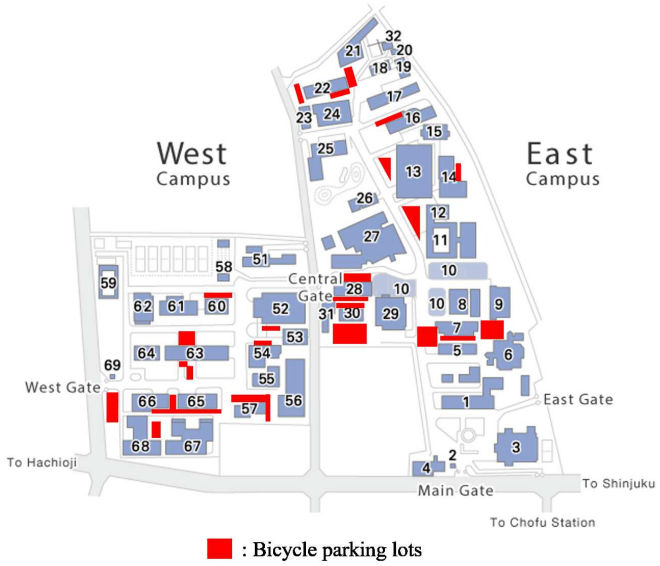

Fig. 3. Bicycle parking lots in our University

data in Section 3.1. Section 3 then debates the data on the knowledge graph, completion techniques and how they can be adapted for predicting the deficit numbers. These methods are then assessed and contrasted with current approaches in Section 4. Lastly, in Section 5, the research is concluded and ideas for forthcoming research are discussed.

2. Related Work

2.1. *Attempts to Adress Issues Using Knowledge Graph*

Japan's Ministry of Internal Affairs and Communications promoted a plan to use such open data. In December 2018, it established the Open Data Training Portal which includes e-learning and web exercise environment. It was built that local public organizations can acquire necessary knowledge and techniques to solve social problem using data[1].

There are several studies to build linked data based on city data. Lopez et al.[8] has developed a platform for publishing sensor data as linked data. The platform collects sensor data and publishes RDF in real time. The biggest advantage of this research is that it is possible to integrate and process multiple data of city by using linked data. However, in terms of cost, this approach is not suitable for our environment where an unspecified number of people access.

Bischof et al. [9] argued that a large number of data about city is still collected manually and can not be linked to each other. Hence, they proposed a method to collect city data, convert it into linked data, and combine standard regression analysis and principal component analysis(PCA) to complete linked data, In order to propose better solutions for cities. However, no corresponding dataset could be found and there are other completion methods

more suitable for knowledge graph, thus the same approach could not be applied to our study.

Zhao et al. [10] have proposed a core conceptual model of a smart city in which multiple domains and cities are related. It incorporates a core ontology for describing social organizations and physical entities. This is to define a standard for concept understanding and data exchange before building a domain-specific knowledge model. Abstract concepts are valuable, but more specific concept designs are needed to apply them to specific domains.

Bellini et al. [11] has built a smart city ontology Km4City, which has a large knowledge base that can integrate and operate various formats of data that is difficult to interoperate. Both static and dynamic data related to smart cities are captured and verified, and it is effective for many general data related to cities. However, in organizations such as universities that have more specific data, more detailed domain-specific classification and interconnection are required. In this research, we first reorganize the concept of data, such as the lecture, movement of mobility, facilities, etc. which is specific to the target campus, and reconstruct it as a knowledge graph.

Furthermore, Egami et al.[12] suggested that, the development of data utilization services to contribute to solving social problems has not progressed sufficiently and that the included data are not reusable. In addition, the author mentioned the usefulness of open data. Egami et al. presented a methodology for designing a unified LOD schema for social issues on the basis of abandoned bicycles, and collected actual data from Social Networking Service (SNS) and administrative websites. They then estimated missing values and proposed a framework for integrating and visualizing these values as LOD. In addition, they realized smooth and dynamic visualization by converting sparse data into dense data by completing missing values. They also proposed and introduced a system for sustainable development of LOD and solving social issues by raising awareness of problems. In particular, the authors proposed a methodology for constructing LOD from Web data, constructed illegally parked bicycle LOD using SNS and administrative data and disclosed the data. Furthermore, they completed missing attribute values using the Bayesian network[17], developed a visualization application for neglected bicycle problem and evaluated the application. Also, they extended the schema based on QB4OLAP in their subsequent study[13]. They made it possible to enrich instances according to the schema semi-automatically using natural language processing and crowdsourcing. Moreover, in other study[14], they applied LOD to large social activity for social awareness improvement in cooperation with the Office for Youth Affairs and Public Safety of the Tokyo Metropolitan Government (Tokyo Bureau). In our research, although we follow this previous schema design methodology and ontology, we estimate the number of bicycles using a specialized internal data knowledge graph and knowledge graph completion method that can easily handle sparse data.

2.2. *Knowledge Graph Completion Methods*

As mentioned above, Bordes et al.[19] proposed the TransE method, which embeds entities and the relationships that connect these entities into a multidimensional vector space and estimates their similarities using an energy-based model to knowledge graph, such as, Wordnet[20] or Freebase.[a] An overview of this method is shown in Fig. 4(a).

Specifically, given the triple (h, r, t), the semantics of which are stored in the subject,

[a]freebase.com

predicate, and object relationship, the triple's score is equal to the dissimilarity distance $d(h+r,t)$ based on the energy-based model shown in Formula1, where entities $h,t \in E$ (the set of entities) and relations $r \in R$ (the set of relations).

$$d(h+r,t) = ||h+r-t||_{l_{1/2}} \tag{1}$$

Here, $l_{1/2}$ describes either the L^1 or the L^2-norm. In addition, to perform vector embedding so that the score is small when the triple (h,r,t) is a correct combination, and the score is otherwise large, we use the loss function $\mathcal{L}$ as follows:

$$\mathcal{L} = \sum_{(h,r,t)\in S} \sum_{(h',r,t')\in S'_{(h,r,t)}} [\gamma + d(h+r,t) - d(h'+r,t')]_+ \tag{2}$$

$[x]_+$ represents the positive component of x and $\gamma > 0$ is a margin hyper parameter. S is the training set of triple (h,r,t), while $S'_{(h,r,t)}$ is the negative triple set where the head or tail (but not both at the same time) is replaced by a possible entity and denoted in formula3 below with $h' \neq h$, $t' \neq t$.

$$S'_{(h,r,t)} = \{(h',r,t)|h' \in E\} \cup \{(h,r,t')|t' \in E\} \tag{3}$$

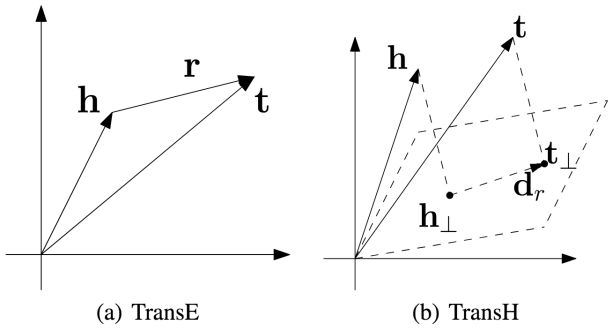

Fig. 4. Overview of TransE[19] and TransH[21]

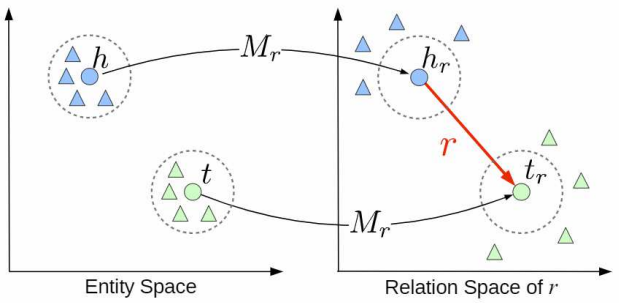

Fig. 5. Overview of TransR[22]

The vector which that makes Equation2 minimum is learned using the stochastic gradient descent to generate the model, so the positive side $d(h + r, t)$ is small and the negative side $d(h' + r, t')$ is large as mentioned above. For missing triples where one entity is missing, if we replace the missing entity with all entities, and score the triple using the model learned above, we find that the smaller the value is, the closer the estimate is to the correct answer. This makes it possible to compensate for defects in the knowledge graph.

Wang et al.[21] proposed the TransH method as shown in Fig. 4(b) as a further extension of the expressiveness of TransE. This method extends TransE's vector embedding and maps entities on relation-specific hyperplanes to amplify the expressiveness of entities with different relations and improve accuracy. Here, TransE entities h and t are mapped on different hyperplanes for each relationship in Equation4 and become $h_\perp$ and $t_\perp$.

$$h_\perp = h - w_r^{\mathrm{T}} h w_r, \quad t_\perp = t - w_r^{\mathrm{T}} t w_r \tag{4}$$

The scoring function is following like TransE using relationship d_r on hyperplane.

$$d(h_\perp, r, t_\perp) = ||(h - w_r^{\mathrm{T}} h w_r) + d_r - (t - w_r^{\mathrm{T}} t w_r)||_2^2 \tag{5}$$

In a further extension of TransE, Lin et al.[22] introduced TransR, which maps entities in different spaces for each relationship, as shown in Fig. 5, and Guoliang et al.[23] introduced TransD, which separates the mapping vector in the triple's subject and the object parts. However, these studies' estimations were based on a large-scale knowledge graph that stores information across various fields and problems rather than addressing familiar topics, such as number estimation in a small domain such as a University campus. Therefore, in this paper, we apply this method to the university campus schema we designed and modify the method to aim at better number estimation accuracy for a small knowledge graph.

3. Building and Completing the Campus Issues Knowledge Graph

3.1. *Building the Knowledge Graph*

3.1.1. *Data set*

We previously constructed the knowledge graph as LOD regarding University campus issues[24]. In this paper, to extend this previous knowledge graph, we retrieved the following information, including bicycle parking status at the main parking lots for the 2017 fiscal year. These data collected from SA, the university website, Google Maps, and the Japan Meteorological Agency.

- Time, e.g., fiscal year, semester, month, day, date, time zone
- Parking areas and the number of bicycles
- Course titles, and classrooms
- Seating capacity in every room
- Names of rooms in every building
- Event titles and venues
- Temperature and precipitation

- Latitude and longitude of every place

3.1.2. *Schema design*

We then revised the RDF schema designed in our previous work[24] emphasizing secondary availability, and defined the following RDF schema, as shown in Fig. 6. This schema can create unique entities with mutual connections via specific relationships. We worked to improve the convenience of search by connecting everything around the bicycle and reducing the number of hops as much as possible.

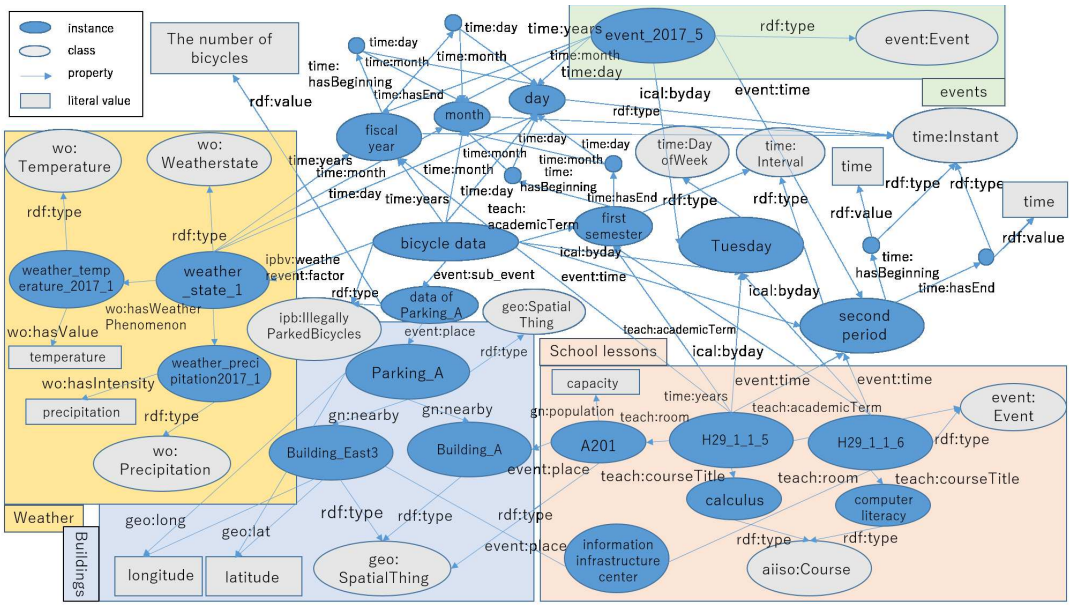

Fig. 6. The schema for campus issues

In addition, we built the schema using various existing ontologies to improve data reusability. Specifically, we used 12 ontologies, 25 properties, and 10 classes, including aiiso, event, geo, gn, owl, rdf, rdfs, time, teach, wo, and the IPBLOD [12]. The relationships between prefix and URI for each name space is as shown in Table 2.

Because the collected data were published on the Web in various formats, after converting the data into CSV format, we assigned uniform resource identifier (URIs), a format used to uniquely identify each resource, to all entities and relations. We then converted the collected data set to RDF data with the defined schema by assigning URIs to unique entities. Currently, the graph contains 20,889 triples, which is data connections with semantics stored in relation to subject, predicate, and object.

3.1.3. *Extraction of results with SPARQL*

We used Open Link Virtuoso 7[25] as an RDF database to store the constructed data, and

Table 2. Prefix and URI for each name space

Prefix	URI
aiiso:	`http://purl.org/vocab/aiiso/schema\#`
event:	`http://purl.org/NET/c4dm/event.owl\#`
geo:	`http://www.w3.org/2003/01/geo/wgs84pos\#`
gn:	`http://www.geonames.org/ontology\#`
ipb:	`http://www.ohsuga.is.uec.ac.jp/bicycle/resource`
ipbv:	`http://www.ohsuga.is.uec.ac.jp/ipblod/vocabulary\#`
owl:	`http://www.w3.org/2002/07/owl\#`
rdf:	`http://www.w3.org/1999/02/22-rdf-syntax-ns\#`
rdfs:	`http://www.w3.org/2000/01/rdf-schema\#`
time:	`http://www.w3.org/2006/time\#`
teach:	`http://linkedscience.org/teach/ns\#`
wo:	`http://www.auto.tuwien.ac.at/downloads/thinkhome/ontology/WeatherOntology.owl`

Table 3. Estimation accuracy of existing method with campus knowledge graph[24]

	MeanRank	Hits@10	Hits@3	Hits@1
TransE	89.747	0.351	**0.284**	0.208
TransH	**70.479**	**0.390**	0.274	**0.219**
TransR	289.414	0.205	0.086	0.023
TransD	148.131	0.288	0.208	0.051
Random	1298	0.385	0.116	0.039

made it possible to resolve URIs reference[26], through the SPARQL endpoint.[b] We published the resulting knowledge graph on the Web[c], and we made it possible for anyone to extract the necessary information through the SPARQL endpoint. As a specific example of data extraction, when we enter a query that containsBicycle data with any year, date, semester, day of the week, time interval, and place and acquire returns such asthe number of bicycles and courses held at that time at facilities near the bicycle parking lot, it is possible to extract the results shown in Fig. 7. Here, each column represents interval, place, the number of bicycles, and the number of courses held, from the left respectively. On the other hand, we can get detailed information about courses such as the course title, the room, etc. as shown in Fig. 8 with a slightly different query. When the output is limited to three in order of course title, it can be seen that three courses with the same title from different teachers are held at the same time. In addition, data in both figures can be acquired simultaneously depending on the style of query.

interval	place	value	callret-3
http://www.ohsuga.lab.uec.ac.jp/interval_1	http://www.ohsuga.lab.uec.ac.jp/pA	179	7

Fig. 7. Example of data extracted using SPARQL query that requires interval, place, the number of bicycles, and the number of courses held at the same time, same place

[b] `http://www.ohsuga.lab.uec.ac.jp/sparql`
[c] `http://www.ohsuga.lab.uec.ac.jp/campus_2017`

class	title	room
http://www.ohsuga.lab.uec.ac.jp/UEC_timetable_id_H29_1_1_130	http://www.ohsuga.lab.uec.ac.jp/UEC_courseTitle_Academic_written_English1	http://www.ohsuga.lab.uec.ac.jp/UEC_room_newC-103
http://www.ohsuga.lab.uec.ac.jp/UEC_timetable_id_H29_1_1_129	http://www.ohsuga.lab.uec.ac.jp/UEC_courseTitle_Academic_written_English1	http://www.ohsuga.lab.uec.ac.jp/UEC_room_newC-303
http://www.ohsuga.lab.uec.ac.jp/UEC_timetable_id_H29_1_1_128	http://www.ohsuga.lab.uec.ac.jp/UEC_courseTitle_Academic_written_English1	http://www.ohsuga.lab.uec.ac.jp/UEC_room_A402

Fig. 8. Example of data extracted using SPARQL query that requires detailed information about courses

3.2. *Knowledge Graph Completion*

In each piece of data collected for knowledge graph generation, defects exist because some information is unobserved. Especially, many of these exist in the bicycle number data set. This is because the bicycle parking lots in the entire area cannot be monitored due to the lack of SA and time, as we mentioned in section . The knowledge graph that aggregates these data also contains many missing triples such as ¡parking lots, rdf:value(property), [m]¿, where [m] is defect value. Thus, we estimated the number of bicycles using the connections of the entire knowledge graph.

3.2.1. *Completion by existing methods*

First, we implemented the previous four methods (TransE[19], TransH[21], TransR[22], and TransD[23]) and estimated the number of bicycles in our previous work[24]. These techniques extract all triples, or data connections with semantics, from the knowledge graph and estimate the defects by embedding the entities and relations that comprise the triples in the vector space. Generally, existing vector embedding and defect estimation methods including the above four, are translation-based models. In these models, for the purpose of evaluation, a new defect is forcibly generated and complemented. Therefore, in this study, we used 20,517 triples excluding 372 triples that contained defects before the experiment.

In this experiment, 1,000 epoch sessions and model evaluation were performed 5 times each with the following hyperparameters: [learning rate, hidden layers, mini-batch size, margin] = [0.001,
100,100,1.0]. Because we aimed to estimate the deficit of the number of bicycles at the university and complete the knowledge graph, only 649 triples that included bicycle number data were used for model evaluation. Table 3 shows the results of this evaluation. MeanRank is the average of the scores for each triple calculated using energy-based distance calculation $d(h, r, t)$ in the translation-based model. The lower the value, the higher the accuracy. However, Hits@n(n=10,3,1) represents the ratio of the correct entities predicted for the experiment's top n, for all the triples. In this case, the higher the value, the higher the accuracy. Here, the bottom row is the numerical value when the number of bicycles is estimated at random. With the exception of TransR, the existing techniques outperform the random value, especially in Hits@3 and Hits@1. Therefore, the knowledge connections in the graph have a certain effect on the number estimation. Among the four evaluated techniques, TransH has the highest accuracy for almost all indexes because TransH can presume triples that have a 1-to-many or many-to-1 relationships, predominantly in the subject and object parts. In number estimation, the number triple is composed of three nodes such as ¡parking lots, rdf:value (property), number [literal value]¿ , and the relationship between the subject and object parts

is many-to-1 in almost all properties. Therefore, TransH results had the highest accuracy.

3.2.2. *Issues with existing methods*

From Table 3 and section , it is clear that the previous techniques are effective for estimating the number of bicycles, and TransH is the most ideal method among them. However, in Hits@10 for TransH on Table 3, the probability that the correct answer is estimated within top 10 is quite low at 0.390%. This is not sufficient to supplement SA's parking status monitoring or to improve student's bicycle parking manners and the parking environment.

3.2.3. *Our method*

We considered the number of bicycles and the population around the bicycle parking area are functionally dependent and proposed a method for more strongly relating the data of the surrounding population in the knowledge graph to estimation. We assumed that bicycle parking lot users park their bicycles during the first period and often do not move their bicycles between classes. We then presumptively calculated the population around the parking lot during the first period of the day and the places in which the bicycles were observed. We aimed for more accurate presumption by reflecting this number in the vector mapping and scoring in TransH. We calculated the population using the class curriculum that seemed most likely to affect population change at the university and the number of seats[d] in the classrooms where the classes were taught, targeted classrooms within a radius of 50-m radius of the parking lot.

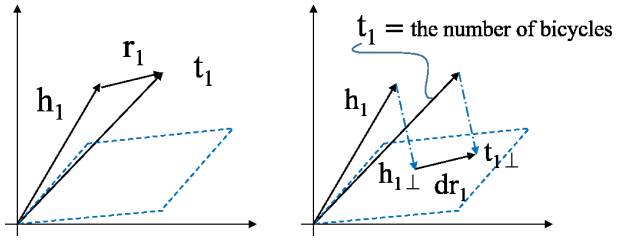

Fig. 9. Translating the relationships of the number of bicycles on hyperplane

In our method, as shown in Fig. 9 and Fig. 10, we extracted triples representing the connection with the surrounding population and triples representing the connection with the number of bicycles for the bicycle parking lot entity h_1 with arbitrary date and time. We then generated a hyperplane for each relation in Equation 4 in TransH for each triple with the same subject part, and we mapped the entities. Furthermore, as shown in Fig. 11, by calculating the difference between entities after mapping and adding this difference to the scoring function denoted in Equation 5, we calculated the by associating the number entity with the surrounding population entities. We used the same principle as Equation 4 in the existing technique, to map triples (h_1, r_1, t_1) of the number of bicycles and triples (h_1, r_2, t_2)

[d]obtained from register of the University of Electro-Communications(`http://kyoumu.office.uec.ac.jp/gakunai/gakubu/room.html`,) for unlisted classrooms, calculated in reference to the floor space obtained from the facilities section(`http://shisetsu.office.uec.ac.jp/gakunai/tatemonogaiyou/gakunaijyoujou.html`)

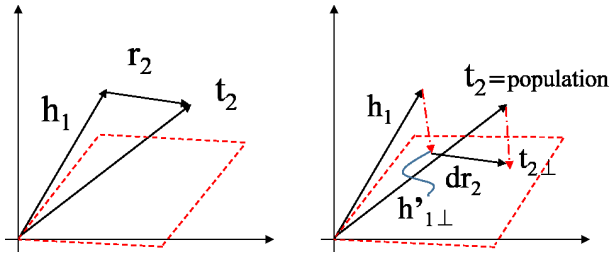

Fig. 10. Translating the relationships of the population on hyperplane

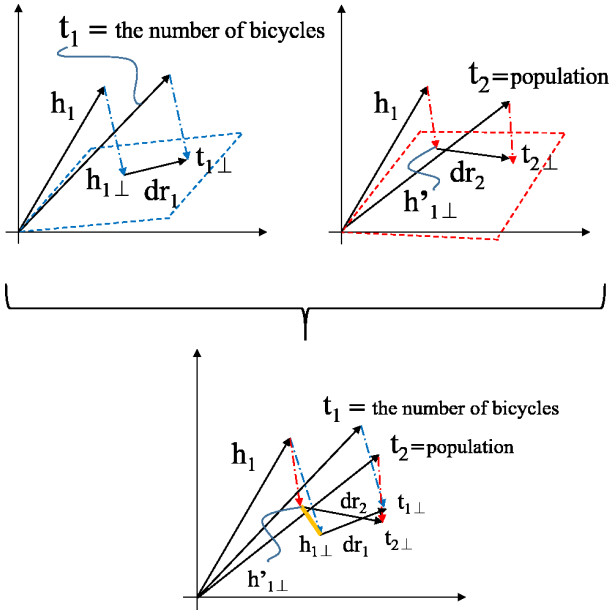

Fig. 11. Overview of the proposed method

of the surrounding population as shown in Equation6 and Equation7, respectively.

$$h_{1\perp} = h_1 - w_{r_1}^{\mathrm{T}} h_1 w_{r_1}, \quad t_{1\perp} = t_1 - w_{r_1}^{\mathrm{T}} t_1 w_{r_1}, \tag{6}$$

$$h'_{1\perp} = h_1 - w_{r_2}^{\mathrm{T}} h_1 w_{r_2}, \quad t_{2\perp} = t_2 - w_{r_2}^{\mathrm{T}} t_2 w_{r_2}, \tag{7}$$

The scoring function is described in Equation 8 in comparison with Equation 5 in the previous method by reflecting the subject part of the mapping Equation 7 of the surrounding population triples in the prediction for triples related to the number of bicycles. Here, d_{r1} is the relationship on the hyperplane after mapping the number triples.

$$\begin{aligned} d \quad & (h_{1\perp}, r_1, t_{1\perp}) \\ = \quad & ||(h_1 - w_{r_1}^{\mathrm{T}} h_1 w_{r_1}) + d_{r1} - (t_1 - w_{r_1}^{\mathrm{T}} t_1 w_{r_1})||_2^2 \\ & + ||(h_{1\perp} - h'_{1\perp})||_2^2 \end{aligned} \tag{8}$$

4. Evaluation

4.1. *Modification of evaluation method*

Considering the above-mentioned evaluation index used for each technique, especially Hits@n, the evaluation method that considers only the upper layer of ranks, not the overall. This means that these methods can not express overall results properly. This is because, as described in Section , the knowledge base used in the previous studies is a large-scale knowledge graph, and various entities are selected as options for inference results from a variety of widely-spread fields. Therefore, the types of entities replaced by the scoring function are also diverse. In addition, the pilot study mainly evaluated node estimation instead of numerical estimation. Based on the above, it is difficult to evaluate the generated model's accuracy.

However, in the estimation performed in this paper, only the bicycle number triples were used as test data. Therefore, all the entities replaced in the scoring function had numerical values, so we could estimate the model's average estimated values. We calculated the weighted average from the generated model's entire prediction order by training of 1,000 epochs, and we used the absolute value error between this and the correct answer, as the model's evaluation metric. Here, the weighted average is an average that uses the weight of each variable and is obtained as shown in Equation 9. The weighted average $\overline{x}$ is as follows:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i w_i^{\alpha}}{\sum_{i=1}^{n} w_i^{\alpha}} \tag{9}$$

Here, variables : weights of variables :$w_1, w_2, w_3 \cdots, w_n$ for each variables : $x_1, x_2, x_3 \cdots, x_n$. The parameter α adjusts the degree of importance placed on the higher ranks.

This estimation of the number of bicycles contains no clear weighting criteria. Hence, we used 649, the total of triples that describ number, as the weight of the number estimated to be most probable, then reduced the weight by 1 when the rank dropped by 1, We assigned a weight of 1 to lowest estimated number. We evaluated the generated model by calculating the absolute value error between this weighted average and the correct answer. When we used the

Table 4. Number estimation accuracy of the existing method and the proposed method

Method	Sum of the absolute value error					
	Minimum	**Maximum**	**5-times average**	**10-times average**	**15-times average**	**20-times average**
Existing method ($\alpha = 0.5$)	44227	46490	44951.6	45495.6	45366.67	45358.35
Proposed method ($\alpha = 0.5$)	45573	47545	46530.4	46690.7	56676.93	46545.5
Existing method ($\alpha = 1.0$)	**37237**	37430	37269.8	37279	37289.6	37302.2
Proposed method ($\alpha = 1.0$)	37239	**37285**	**37249.8**	**37245.7**	**37247.33**	**37247.7**
Existing method ($\alpha = 2.0$)	48621	53768	51194.5	51659.14	51713.63	51713.63
Proposed method ($\alpha = 2.0$)	48670	52916	50503.2	51092.2	50942	50891.45
Evaluation value difference ($\alpha = 0.5$)	-1346	-1055	-1578.8	-1195.1	-1310.26	-1187.15
Evaluation value difference ($\alpha = 1.0$)	-2	145	20	33.3	42.27	54.5
Evaluation value difference ($\alpha = 2.0$)	-49	852	691.3	566.94	771.63	822.18

The difference between the evaluation values is positive when the proposed method outperforms the existing one.

evaluation method, we found that higher model accuracies indicated smaller absolute value errors. In other words, the smaller the evaluation value, the higher the accuracy.

4.2. *Comparison of the Existing Method and the Proposed Method*

In the implementation of the previous method and the proposed method, the linear congruential generator, which is a pseudorandom number generator, is used to select training triples. The congruential generator generates an random integer between 0 and $M = 20,889$, which is the total number of training triples. The same value is not output twice in the maximum period M. Hence, we performed one training for 1,000 epochs. In other words, the number of triples learned per training was 1,000 triples, and repeating this 20 times results in 20,000 triples, which is quite close to the total number of triples learned during training. After that, we confirmed the accuracy based on the evaluation method proposed in section . Table 4 shows the results of applying the proposed evaluation method when we repeated the specified number of trainings and presumptions with TransH and our method. In this paper, we compare the accuracy of the existing method and the proposed method for the case involving the highest accuracy among multiple experiments. As the Table 4 shows, the best results were obtained when $\alpha = 1.0$ for both existing method and proposed method.

4.3. *Consideration*

In our method, the maximum value was greatly reduced compared to the pilot study, and difference between the evaluation values was such that we correctly estimated 54.5 units in an average of 20 experiments. Considering that our method handled 649 bicycle parking lot data, the estimated accuracy per one parking lot at any given time and place is about 0.083 more accurate, on average, than the existing method. In particular, when the existing method and generated models are compared and the following three points are considered, the deviation in the generated model's accuracy is smaller than that of the existing method. First, only the maximum decreased. Second, the difference between the minimum and the maximum was reduced by nearly 150 from 193 to 46. Third, the fluctuation in the average of the sum of absolute value errors was reduced from 32.4 to 2. This made it possible to surpass the conventional method's accuracy. We also conducted a preliminary experiment in which only the first period of the observation date was reflected in the parking data of all time zones. We based this experiment on the assumption that bicycle users park when the first class begins

and do not move their bicycles between classes. Although our method estimated the number of bicycles more accurately than the existing method, the result was inferior to the method shown in Table 4. Therefore, the number of classes offered in the area around the observed parking lot and the population that attended these classes influenced the number of bicycles in the parking lot. In addition, the trial of Equation 8 was useful in the proposed method, which reflects this fact in its predictions.

5. Conclusion and Future Works

5.1. *Conclusion*

When tackling social issues, the assembly and application of disclosed data can be imperative. Making links between various data can allow new outlooks to develop to resolve these issues, as shown by the recent strategies of Japan's Ministry of Internal Affairs and Communications, "Open Data Watch" and the other institutions who are dedicated to developing open data. Our research regarded the local university environment as a microcosm of society and concentrated on the issues faced by bicycle parking stations on campus. It has also conducted data collection and defect approximation using a knowledge graph. We ensured data was reusable and subsequently created an original schema for internal data and gathered data from the Web in a variety of formats. This data was then combined with the information in the knowledge graph, following the schema. Moreover, the research used existing approaches for predicting the deficit triples in the knowledge graph, established the accuracy of number prediction and enhanced the accuracy by translating the data to an approach created specifically for our knowledge graph.

5.2. *Future Works*

Future research will concentrate on three particular aspects. Firstly, further stored data will be aggregated and used to enhance the knowledge graph. This research was limited in the range and timescale of data that could be collected because university data was gathered manually and implemented into the graph. It is therefore noteworthy that the conclusions and predictions may hold some bias or inadequacies. Going forward, computerized data assemblage and storage will be deliberated. Secondly, an improved algorithm will be projected. Section discusses our approaches as superior to existing methods regarding the predictions of bicycle numbers. However, it is uncertain whether the present algorithm is beneficial in accuracy. Consequently, and enhanced algorithm is needed. The third aspect involves the introduction of amenities such as congestion visualization. It is believed that the profit returns for producing facilities for familiar resources is substantial. At present, building services are time consuming; however, estimation accuracy may see the development of congestion status visualizations, bicycle parking recommendations, and evacuation guidance. Using these ideas, our goal is to circulate common databases, including knowledge graphs, concerning the challenges faced by university campuses, and to make use of cyclical data, including developing services.

Acknowledgments

We are grateful to Shusaku Egami, Ph.D. for helpful discussions. This work was supported by JSPS KAKENHI Grant Numbers JP16K00419, JP16K12411, JP17H04705, JP18H03229, JP18H03340, JP18K19835, and NAKAJIMA FOUNDATION.

References

1. Ministry of Internal Affairs and Communications, Japan*Establishment of "Open Data Training Portal"*, `http://www.soumu.go.jp/english/`.
2. Ontotext, *What are Linked Data and Linked Open Data?*, `https://www.ontotext.com/knowledgehub/fundamentals/linked-data-linked-open-data/`.
3. J. P. McCrae for the Insight Centre for Data Analytics, *The Linked Open Data Cloud*, `https://lod-cloud.net/`.
4. H. Mousselly-Sergieh, T. Botschen, I. Gurevych, and S. Roth (2018), *A Multimodal Translation-Based Approach for Knowledge Graph Representation Learning*, Joint Conference on Lexical and Computational Semantics, pp. 225-234.
5. H.Yoon, H.Song, and S.Park (2016), *A Translation-based Knowledge Graph Embedding Preserving Logical Property of Relations*, Proc. of NAACL-HLT, pp. 907-916.
6. G. Costa and J. Oliveira (2018), *Linguistic Frames as Support for Entity Alignment in Knowledge Graphs*, proc. of the 20th International Conference on Information Integration and Web-based Applications & Services (iiWAS).
7. T.-D. Trinh, P. Wetz, B.-L. Do, A. Anjomshoaa, E. Kiesling, and A. M. Tjoa (2014), *A web-based platform for dynamic integration of heterogeneous data*, proc. of the 16th International Conference on Information Integration and Web-based Applications & Services (iiWAS), pp.253-261.
8. V. Lopez, S. Kotoulas, M. L. Sbodio, M. Stephenson, A. Gkoulalas-Divanis, and P. M. Aonghusa (2012), *QuerioCity: A Linked Data Platform for Urban Information Man-agement*, proc. of the 11th International Semantic Web Conference (ISWC), pp.148-163.
9. S. Bischof, C. Martin, A. Polleres, and P. Schneider (2015), *Collecting, Integrating, Enrichingand Republishing Open City Data as Linked Data*, proc. of the 14th InternationalSemantic Web Conference (ISWC), pp.57-75.
10. J. Zhao and Y. Wang (2015), *Toward domain knowledge model for smart city: The core conceptual model*, Smart Cities Conference (ISC2), IEEE First International, pp. 15.
11. P. Bellini, M. Benigni, R. Billero, P. Nesi, and N. Rauch. (2014), *Km4City ontology building vs data harvesting and cleaning for smart-city services. Journal of Visual Languages and Computing*, Vol.25(6), pp. 827-839.
12. S. Egami, T. Kawamura, and A. Ohsuga (2016), *Building Urban LOD for Solving Illegally Parked-Bicycles in Tokyo*, proc. of the 15th International Semantic Web Conference (ISWC), pp.291-307.
13. S. Egami, T. Kawamura, K. Kouji, and A. Ohsuga (2017), *Linked Urban Open Data Including Social Problems' Causality and Their Costs*, proc. of the 7th Joint International Semantic Technology Conference (JIST), pp.334-349.
14. S. Egami, T. Kawamura, and A. Ohsuga (2018), *Temporal and Spatial Expansion of Urban LOD for SolvingIllegally Parked Bicycles in Tokyo*, IEICE Trans. on Information and Systems, Vol.E101.D, No.1, pp.116-129.
15. Open Data Watch (2019)*Country Engagement Reaches New High*, `https://opendatawatch.com/blog/country-engagement-reaches-new-high/`.
16. FUJITSU JOURNAL (2019), *Next-generation Internet: LOD Represents the Future of Data*, `https://journal.jp.fujitsu.com/en/2014/01/10/01/`.
17. R.G. Cowell, A.P.Dawid, S.L. Lauritzen, and D.J. Spiegelhalter (1999), *Probabilistic Networks and Expert Systems*, Springer-Verlag.
18. S. Kuwahara and H. Kanai (2015), *Norm Attitude Survey on Suppression of Unoccupied Bicycles by Notification of Monitoring Information*, International Processing Society of Japan (IPSJ).

19. A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and L. Yakhnenko (2013), *Translating Embeddings for Modeling Multi-relational Data*, proc. of NIPS, pp.2787-2795.
20. G. Miller (1995), *WordNet: a Lexical Database for English*, Communications of the ACM, Vol.38(11), pp.39-41.
21. Z. Wang, J. Zhang, J. Feng, and Z. chen (2014), *Knowledge Graph Embedding by Translting on Hyperplanes*, proc. of AAAI, pp.1112-1119.
22. Y. Lin, J. Zhang, Z. Liu, M. Sun, Y. Liu, and X. Zhu (2015), *Learning Entity and Relation Embeddings for Knowledge Graph Completion*, proc. of AAAI, pp.2181-2187.
23. J. Guoliang, H. Shizhu, X. Liheng, L. Kang, and Z. Jun (2015), *Knowledge Graph Embedding via Dynamic Mapping Matrix*, proc. of ACL, pp.687-696.
24. Y. Tsukagoshi, T. Kawamura, and A. Ohsuga (2018), *Knowledge Graph on University Campus Issues*, proc. of the 8th International Sematntic Technology Conference (JIST), pp. 118-121.
25. Open Link Software, *About OpenLink Virtuoso*, `https://virtuoso.openlinksw.com`.
26. T. Berners-Lee, R. T. Fielding, and L. Masinter (2015), *Uniform Resource Identifier (URI): Generic Syntax, IETF Request for Comments 3986*, `https://tools.ietf.org/html/rfc3986#page-28`.