Detecting Vicious Cycles in Urban Problem Knowledge Graph using Inference Rules (Postprint)

Authors: Shusaku, Egami, Takahiro, Kawamura, Kouji, Kozaki, Akihiko, Ohsuga, Shusaku, Egami

Date: 2022-11-28T00:00:00+00:00

Abstract

Urban areas have many problems, including homelessness, graffiti, and littering. These problems are influenced by various factors and are linked to each other; thus, an understanding of the problem structure is required in order to detect and solve the root problems that generate vicious cycles. Moreover, before implementing action plans to solve these problems, local governments need to estimate cost-effectiveness when the plans are carried out. Therefore, this paper proposed constructing an urban problem knowledge graph that would include urban problems' causality and the related cost information in budget sheets. In addition, this paper proposed a method for detecting vicious cycles of urban problems using SPARQL queries with inference rules from the knowledge graph. Finally, several root problems that led to vicious cycles were detected. Urban-problem experts evaluated the extracted causal relations.

Full Text

Preamble

Detecting Vicious Cycles in Urban Problem Knowledge Graph using Inference Rules

Shusaku Egami[1,2]†, Takahiro Kawamura[1,3], Kouji Kozaki[1,4] & Akihiko Ohsuga[2]

[1]National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan
[2]Graduate School of Informatics and Engineering, The University of Electro-Communications, Tokyo 182-8585, Japan
[3]National Agriculture and Food Research Organization, Ibaraki 305-8517,

Japan
Faculty of Information and Communication Engineering, Osaka Electro-Communication University, Osaka 572-8530, Japan

Keywords: Linked data; Social problem; Causality extraction; Crowdsourcing; Text mining; Open city data; Semantic inference; SPARQL; Semantic Web Rule Language (SWRL)

Citation: Egami, S., et al.: Detecting vicious cycles in urban problem knowledge graph using inference rules. Data Intelligence 4(1), 88-111 (2022). doi: 10.1162/dint_a_{00113}
Received: August 11, 2021; **Revised:** October 29, 2021; **Accepted:** November 9, 2021

ABSTRACT

Urban areas face numerous problems, including homelessness, graffiti, and littering. These problems are influenced by various factors and are interconnected, requiring an understanding of their underlying structure to detect and solve the root problems that generate vicious cycles. Moreover, before implementing action plans, local governments need to estimate cost-effectiveness. This paper proposes constructing an urban problem knowledge graph that incorporates both the causality of urban problems and related cost information from budget sheets. Additionally, we propose a method for detecting vicious cycles of urban problems using SPARQL queries with inference rules derived from the knowledge graph. Several root problems leading to vicious cycles were identified, and urban-problem experts evaluated the extracted causal relations.

1. INTRODUCTION

Local governments must address numerous urban problems, including suburban crimes, dead shopping streets, and littering. However, because various factors are socially intertwined, these problems are difficult to solve without understanding the causal relationships among them. Structural management of data on both urban problems and their causality is therefore required for visualization and problem-solving. In this context, "causality" describes cause-and-effect relationships between two or more factors. To implement action plans, local governments must also grasp cost-effectiveness, as most are highly cost-sensitive and cannot establish new projects without clear estimates of their effects (e.g., cost reduction). This concern was explicitly raised by an official from the Yokohama Policy Bureau in Kanagawa Prefecture, Japan.

Our first objective is to construct a knowledge graph (KG) that includes the causal relations of urban problems and related cost information from budget sheets. This KG can predict the impact of urban problems by tracing both causality and hierarchical links in background ontologies. Additionally, we aim to detect vicious cycles among urban problems, identify their root problems, and search for related budget information using the constructed KG. This can help local governments consider solutions to intertwined urban problems in terms of cost-effectiveness.

We first designed an ontology representing urban problem causality using Web Ontology Language (OWL), then extended the vocabularies to include budget information based on QB4OLAP [1], an extension of the RDF Data Cube vocabulary. Next, we semi-automatically constructed the KG based on this ontology.

For constructing a KG of urban problems, it is necessary to extract as many words as possible related to causal relations while removing unrelated words. To address this challenge, we: (1) collected various data sources from the Web, including government web pages, open data, news articles, PDF documents from academic literature, and citizen voices on blogs and SNS; (2) filtered the target to locate sentences containing "factor," "affect," and their synonyms; (3) extracted candidate causal relations using dependency structure analysis; (4) extracted causal word candidates with a certain level of agreement through crowdsourcing using word clouds; and (5) defined 11 patterns of causal relations to be considered, proposing a method for complementing them using inference rules written in Semantic Web Rule Language (SWRL).

Furthermore, we detected vicious cycles and root problems using SPARQL, then evaluated them based on comments from urban-problem experts. We also identified root problems that lead to multiple vicious cycles and searched for related budget information. However, since no absolutely correct dataset exists for urban-problem causality, evaluating the detected vicious cycles and root problems estimated through SPARQL queries is challenging. Therefore, we conducted these evaluations in cooperation with the Osaka City Citizens Bureau and report the results of two case studies. Our contributions are as follows:

1. Constructing an urban problem knowledge graph that includes causality and budget information
2. Semi-automatically extracting urban problem causality using NLP and crowdsourcing
3. Detecting vicious cycles and root problems using SPARQL with inference rules
4. Evaluating those vicious cycles and root problems based on comments from urban-problem experts

The remaining sections are organized as follows. Section 2 provides an overview of knowledge graphs relating to urban problems, city data, and crowdsourcing. Section 3 describes the schema design. Section 4 outlines the method for constructing KGs related to urban problems and evaluates its effectiveness. Section 5 describes the detection of vicious cycles and root problems and discusses expert evaluation in cooperation with the Osaka City Citizens Bureau. Finally, Section 6 concludes with future extensions.

2.1 Using Knowledge Graphs to Solve Social Problems

Some studies have proposed using linked data to solve social issues. Szekely et al. [2] built a knowledge graph from crawled websites to combat human trafficking and develop a lost children search system that six law enforcement agencies and several NGOs have since deployed. While Szekely et al. built a knowledge graph related to a specific domain of social problems, we aim to build a KG related to multiple urban problems including their causality.

In our previous work [3], we built and visualized Linked Open Data (LOD) to solve the problem of illegally parked bicycles, an urgent urban problem in Japan. We also proposed a methodology for designing an LOD schema involving everyday urban problems. However, this methodology required all steps to be completed manually, relied on limited web documents, and depended on workers' knowledge, resulting in low coverage for extracting urban problem causality. Thus, we propose a method for semi-automatically extracting urban problem causality [4]. Moreover, the LOD constructed in our previous study [5] was based on a schema extended from Event Ontology, but it was difficult to search for urban-problem causality using OWL inference rules and did not contain budgetary information. Therefore, in this paper we define an ontology representing urban problem causality and budget information.

Shiramatsu et al. [6] proposed using LOD to share goals and solve social issues, resulting in goal-matching that facilitates civic technology—solving social issues through information technology and collaboration between citizens and local governments. This research led to a web application called GoalShare, used for domestic civic technology events. However, Shiramatsu' s LOD mainly describes public goals for solving social issues rather than causalities. Associating this LOD with ours would facilitate finding solutions to social problems.

In summary, while Szekely et al. [2] and our previous work [3] aimed to solve specific problems, this paper adopts a cross-domain perspective common to social problems. Our knowledge graph schema has been well-considered for future extendability. Moreover, Shiramatsu et al. [6] focus on goal-sharing and goal-matching, differing from our objectives.

2.2 Using Knowledge Graphs to Analyze City Indicators

Santos et al. [7] defined city knowledge graphs using OWL to analyze various city indicators. They proposed quality-of-experience ontological indicators—calculated numerical values that support convenient visualization—and developed a dashboard application that generates widgets to aid in visualizing knowledge graph data. Pileggi et al. [8] defined an ontological framework implemented using OWL-DL ontology to represent dynamic, fine-grained urban indicators. To simplify data structure understanding and facilitate usability, they partitioned the ontology into five sub-ontologies based on function and scope: indicator, data, profiling, computations, and geographic context.

LinkedSpending [9] represents linked data based on OpenSpending, an open platform for public financial information including budgets, spending, balance sheets, and procurement. As of May 2017, users had registered 1,104 datasets from 75 countries, with Japan having the most (415). The data is modeled on RDF Data Cube vocabulary, designed for multidimensional statistical data. However, LinkedSpending does not describe urban problems and cannot be directly used to solve them. Based on LinkedSpending, we used QB4OLAP, an extension of RDF Data Cube.

2.3 Crowdsourcing and Natural Language Processing for Linked Data

Demartini et al. [10] proposed an entity linking method using crowdsourcing to improve link quality and developed a probabilistic framework to integrate inconsistent results. Celino et al. [11] developed a mobile application to link point-of-interest data to pictures using crowdsourcing and introduced a game with a purpose (GWAP) [12] to incentivize users. However, no study has yet created LOD related to urban-problem causality using crowdsourcing.

Nguyen et al. [13] proposed constructing linked data concerning users' activities. They used conditional random fields to extract activities from Japanese weblogs and constructed triples of action, object, time, and location. These datasets can analyze user activities during earthquakes. LODifier [14] extracted entities from unstructured text using a named entity recognition (NER) system called Wikifier [15], then combined entities using DBpedia and WordNet.

Many studies use natural language processing (NLP) to construct linked datasets. Current state-of-the-art NER systems in English typically achieve 85-90% accuracy for news text (e.g., CoNLL03 shared task dataset) but perform poorly (30-50% accuracy) on short texts lacking linguistic formalism (e.g., punctuation, spelling, spacing, formatting, unorthodox capitalization, emoticons, abbreviations, or hashtags) [16]. Therefore, this paper combines NLP with crowdsourcing to extract urban-problem causality. While other studies refine linked data using crowdsourcing, ours differs in its combination with NLP.

2.4 Semantic Inference to Detect Relations

In the drug-drug interaction (DDI) field, many studies use inference rules to detect new relations in knowledge graphs, and we referenced their evaluation methods. Moitra et al. [17] modeled pharmacokinetic DDI using Semantic Application Design Language, then estimated interactions related to several enzymes using SWI-Prolog. Herrero-Zazo et al. [18] provided a comprehensive ontology for pharmacokinetic and pharmacodynamic interactions (DINTO), then estimated relations such as "may interact with" using DINTO and SWRL rules. While many studies address ontological reasoning, to our knowledge none have used inference rules to infer causal relations among urban problems.

3. DESIGNING AN ONTOLOGY OF PROBLEM CAUSALITY AND COSTS

Our KG is primarily intended for investigating solutions to urban problems. Specifically, local governments can query our KG to consider solution effects and budgets. Thus, we designed the ontology shown in Figure 1 [Figure 1: see original paper] to represent urban problem causality and local government budgets.

The upper half of Figure 1 defines vocabulary representing urban-problem causality. All resources are classified as `upv:CausalEntity`, with a subset as `upv:UrbanProblem`. There are two main causality properties: `upv:factor` and `upv:affect`, both subproperties of `upv:related`. Because urban problems are not events and lack temporal or spatial aspects, we did not reuse the `event:factor` property from Event Ontology. Sub-properties of `upv:factor` and `upv:affect` represent crowdsourcing agreement levels. Dividing causality properties into `upv:factor` and `upv:affect` enables forward or backward chaining with agreement levels that restrict domain or range. For example, when extracting strong causality, `upv:factor_{level4}` and `upv:affect_{level4}` properties can be used. The values of `upv:affect_{level4}` are words selected as factors by more than 35 crowdsourcing workers. When extracting causality regardless of agreement level, `upv:factor` and `upv:affect` properties can be used in SPARQL queries.

The lower half of Figure 1 defines vocabulary representing budget information. Because most local government budget information is published as tabular data (e.g., Microsoft Excel, PDF), we described it using QB4OLAP, an extension of RDF Data Cube vocabulary. QB4OLAP has been used in business-intelligence tool data models and includes `qb4o:LevelProperty` and `qb4o:AggregateFunction`, which support aggregation operations. Thus, users can query the total budget of each department and determine which urban problem requires the highest budget. The `upq:Project` class represents projects for solving social problems, with instances having at least one `dcterms:subject` property whose values are instances of `upv:CausalEntity`. Instances of the Observation class are cells (values) in tabular data of projects for solving social problems, having a `upq:budget` property whose value is the budget. A ward allocating a project' s budget is described as an instance of `upq:Ward` class, and a city as an instance of `upq:City`. The `rdfs:range` of `upq:ward` is `upq:Ward`. Lower-case letters denote properties. This schema design enables users to query project budgets using aggregate functions.

4.1 Extraction of Urban Problem Causality

We propose a method for semi-automatically extracting urban problem causality through: (1) collecting web documents using search engines; (2) extracting causality words using natural language processing; (3) generating word clouds from extracted words; and (4) filtering words using crowdsourcing.

Our goal is to aggregate qualitative causal knowledge of urban problems from diverse sources: government web pages, open data, news articles, academic literature, blogs, and social networking sites. We also believe human subjective choices must be considered. Our envisioned applications include discussion tools for experts solving urban problems and tools for explaining evidence of causal relationships. Thus, constructed causal relations must reflect common understanding to some extent, be explainable, and contain unexpected results. We believe crowdsourcing using word clouds—which presents cause-and-effect relationship candidates—is suitable for extracting meaningful causality words.

We collected documents from search engines using urban problem names and synonyms of "factor" as keywords. For example, the first keyword was "suburban crime" and the second was "factor" (with synonyms including "element," "origin," and "cause") obtained from Japanese WordNet. We used Google Custom Search API and Bing Web Search API to obtain document lists, collecting 50 HTML and 50 PDF files separately for each keyword set (unique combination of first and second keywords). This included reports from governments and citizens. However, many unrelated documents were also collected, so we excluded those containing few words related to urban problem names. The number of documents containing urban problem words was: ((kinds of urban problems $\times$ # of synonyms of "factor" $\times$ 50 HTMLs) + (kinds of urban problems $\times$ # of synonyms of "factor" $\times$ 50 PDFs) + (kinds of urban problems $\times$ # of synonyms of "affect" $\times$ 50 HTMLs) + (kinds of urban problems $\times$ # of synonyms of "affect" $\times$ 50 PDFs)) -# of noise documents = 3,903.

Next, we extracted noun words using morphological analysis. To facilitate crowdsourcing, we concatenated verbal nouns to construct noun phrases. For example, "preventing delinquency" was split into "preventing" and "delinquency" by morphological analysis, but we concatenated these into a noun phrase.

Using Japanese dependency analysis, we extracted noun phrases with causal relationships to synonyms of "factor" based on dependency relations in each sentence [19]. Likewise, we extracted affecting words using synonyms of "affect" such as "influence," "effect," and "evoke" as second keywords. We generated word clouds from extracted possible causality words and filtered them using crowdsourcing, assuming word clouds would increase word visibility and simplify extraction of important words. We received comments that this method was fun and game-like.

Spelling inconsistencies reduce word cloud visibility. We used Jaro-Winkler distance [20] to calculate word similarity, empirically setting the threshold to 0.8. When similar words were found, we integrated their occurrence counts into the longest word. Figure 2 [Figure 2: see original paper] shows a word cloud of suburban crime factors.

We conducted two crowdsourcing tasks: "select 10 words considered factors in suburban crime" and "select 10 words considered affected by suburban crime." We used the Lancers crowdsourcing service, setting the reward at 50 JPY and asking up to 50 people per problem. We gathered all words selected by more than 10% of workers (translated in Figure 2).

To enrich urban problem causality, we repeated our method using extracted causality words. This repetition increased intermediate nodes in our knowledge graph. However, not all words had causal relations, so we also extracted co-occurring words from the top 50 web documents related to causality words. If we found more than 336 co-occurring words (top 5%) such as "cause," "factor," "influence," "urban," and "city," we applied our extraction method to those causality words.

4.2 Building KG Based on the Extracted Causality Words

We built the KG based on our designed ontology and extracted words. Since Lancers exports results in CSV format, we used Apache Jena to convert CSV files to RDF files based on our schema. We created urban-problem resources as subclasses of `upv:UrbanProblem` and other resources as subclasses of `upv:CausalEntity`. We used both SKOS and OWL for usability: naive users can intuitively search the graph by tracing simple SKOS relations (broader and narrower) without violating QB4OLAP' s SKOS restrictions. OWL reasoning is useful for detecting vicious cycles and root problems. We formed causality links using `upv:factor` and `upv:affect` properties corresponding to the number of participants who agreed.

If resources had the same name as extracted noun words or matched `skos:altLabel` values in WikiData, we created alternative resources and alternative hyper resources. For example, the "temporary work" resource in WikiData has Japanese labels such as "パートタイマー" (part-time work) and a hyper class called "employment" with Japanese labels such as "雇用契約" (employment contract). Figure 3 [Figure 3: see original paper] shows generated KG fragments. If no matching resources were found in WikiData, we extracted noun words from class names using morphological analysis and created hyper classes based on those noun words using head and modifier matching methods [21].

4.3 Generating Instances Based on Budget Data of Local Government

Osaka is an ordinance-designated city in Japan and the capital of Osaka Prefecture. Osaka City has published various open datasets on its Osaka City Open Data Portal Site, most of which are CC-BY 4.0 licensed. We used Apache Jena and Apache POI to convert tabular data and PDF files from this site to RDF files based on our schema.

We then linked local government project resources to causality resources. Since budget sheets lacked detailed project descriptions, we linked project resources to causality resources using project names. However, determining relations such as between a "project for solving street smoking problems" and "cigarettes" was difficult. Therefore, we used Algorithm 1 to link project resources to causality resources.

The algorithm extracts all noun words except stop words from local project names, obtains synonyms corresponding to extracted noun words using Japanese WordNet as linking candidates, and extracts glosses (multiple short sentences describing word senses and usage). From these glosses, we also extract noun words as linking candidates. If candidate words match causality resources, we link project resources to causality resources using the `dcterms:subject` property.

Figure 4 [Figure 4: see original paper] shows part of the constructed KG, accessible from our website. The source code for data collection and KG building is available on GitHub. The ontology contains 70,076 triples. We validated our KG using RDFUnit [22], a test-driven data-debugging framework that automatically generated 68 test cases, all of which passed without timeouts, errors, or violations. Thus, we correctly reused existing vocabulary without violating domain or range restrictions. All resources are linked, and none are independent.

4.4 Result of NLP

Table 1 shows statistics for causality word extraction. Because "affect" has many synonyms, documents related to affecting words numbered 2,465—larger than those related to factors (1,438). We excluded some "factor" synonyms (e.g., "procatarxis") that are rarely used, as they returned many unrelated documents. Consequently, the number of affecting words was large, and selection agreement was lower than for factor words.

Missing words related to urban problems can lower agreement in causality-word extraction. In some cases, phrases extracted using our method did not match those describing causality. Complex sentences in government documents prevented causality word extraction in many cases, which we attempted to solve through text simplification methods. In other cases, extracted causality words were unrelated to urban problems. To exclude these errors, we extracted words appearing in multiple documents rather than those appearing many times in a single document.

Jaro-Winkler distance was used only to solve spelling inconsistencies displayed in word clouds. We also used WikiData to obtain `skos:altLabel` values for synonyms (Section 4.2). However, we must also consider unifying phrases with low string similarity but identical meaning. Word embedding techniques might solve this by calculating similarity after obtaining vector representations of causal word candidates. Since dependency structure analysis extracts noun phrases, we need noun phrase embeddings, requiring generation of embedding models based on our collected data rather than pre-trained models.

4.5 Crowdsourcing Results

To calculate agreement in causality-word selection through crowdsourcing, we used Fleiss' s kappa [23]. With 50 users, the average agreement for extracted factor words was 0.291 and for affecting words 0.212, with total average agreement of 0.256—indicating fair agreement according to benchmarks [24].

High agreement for traffic accident factors (0.443) resulted from extensive reporting by the Metropolitan Police Department, educational institutions, and news organizations, giving workers substantial background knowledge. High agreement for noise factors (0.468) occurred because workers had personal experiences of noise affecting them.

5.1 Complementing Missing Links Using Causal Inference Rules

Our KG contained many missing causal links, making most vicious cycles undetectable from direct causal links alone. Therefore, we defined causal inference rules to complement missing links. Figure 5 [Figure 5: see original paper] shows complementary missing links based on hyper classes and alternative classes, with numbers corresponding to inference rules stored in SWRL and converted to Stardog rules (Stardog recommends native syntax based on SPARQL rather than SWRL):

[1] (?x upv:affect ?y), (?y skos:broader ?yp) -> (?x upv:probablyAffect ?yp).
[2] (?x upv:affect ?y), (?x skos:broader ?xp) -> (?xp upv:probablyAffect ?y).
[3] (?x upv:affect ?y), (?x skos:broader ?xp), (?y skos:broader ?yp) -> (?xp upv:mayAffect ?yp).
[4] (?x upv:affect ?y), (?y prov:alternateOf ?yalt) -> (?x upv:likelyAffect ?yalt).
[5] (?x upv:affect ?y), (?y skos:broader ?yp), (?yp prov:alternateOf ?ypalt) -> (?x upv:mightAffect ?ypalt).
[6] (?x upv:affect ?y), (?x skos:broader ?xp), (?y prov:alternateOf ?yalt) -> (?xp upv:mightAffect ?yalt).
[7] (?x upv:affect ?y), (?x skos:broader ?xp), (?y skos:broader ?yp), (?yp prov:alternateOf ?ypalt) -> (?xp prov:possiblyAffect ?ypalt).
[8] (?z upv:affect ?yalt), (?yalt prov:alternateOf ?y), (?y skos:broader ?yp) -> (?z upv:mightAffect ?yp).
[9] (?z upv:affect ?yalt), (?yalt prov:alternateOf ?y), (?y skos:broader ?yp), (?yp prov:alternateOf ?ypalt) -> (?z upv:possiblyAffect ?ypalt).
[10] (?z upv:affect ?yalt), (?z skos:broader ?zp), (?yalt prov:alternateOf ?y), (?y skos:broader ?yp) -> (?zp upv:possiblyAffect ?yp).
[11] (?z upv:affect ?yalt), (?z skos:broader ?zp), (?yalt prov:alternateOf ?y), (?y skos:broader ?yp), (?yp prov:alternateOf ?ypalt) -> (?zp upv:possiblyAffect ?ypalt).

These rules created five causal relation properties: `probablyAffect`, `likelyAffect`, `mayAffect`, `mightAffect`, and `possiblyAffect`. We defined costs as: `upv:affect` = 1, `prov:alternateOf` = 0.75, and `skos:broader` = 0.5. Therefore, causality strength for inference properties follows total antecedent costs: `probablyAffect` > `likelyAffect` > `mayAffect` > `mightAffect` > `possiblyAffect`. These properties are subproperties of `upv:affect`. The experiment complemented 1,058 `probablyAffect`, 122 `likelyAffect`, 191 `mayAffect`, 333 `mightAffect`, and 179 `possiblyAffect` properties.

5.2 Detecting Vicious Cycles of Urban Problems

We define vicious cycles of urban problems as loops of three or more nodes using only sub-properties of `upv:affect` (Figure 6 [Figure 6: see original paper]). Each node corresponds to a subclass of either `upv:UrbanProblem` or `upv:CausalEntity`, with at least one node being an urban problem. We used SPARQL queries to extract cycles containing 3 to 6 nodes (the maximum cycle length can be incrementally increased in consultation with experts). Figure 7 [Figure 7: see original paper] shows an example SPARQL query for detecting 3-node vicious cycles. Since obtained cycles included duplicates (e.g., "Poverty → Truancy → Disease" and "Truancy → Disease → Poverty"), we deleted such duplicates.

Table 2 shows the number of detected vicious cycles: 951 through SPARQL queries and 1,904 through SPARQL queries with inference rules. The "Inference" column shows results after applying inference rules from Section 5.1. When searching for vicious cycles and root problems, we changed `upv:affect` from `owl:TransitiveProperty` to `owl:ObjectProperty`, preventing arbitrary long cycles from appearing in 3-node results. Duplicate 3-node results were removed from other results. Our ontology is based on OWL DL, ensuring sound reasoning.

5.3 Experiment for Detecting Root Problems Using SPARQL Patterns

We used SPARQL queries to detect root problems leading to multiple vicious cycles. Figure 8 [Figure 8: see original paper] shows the query for detecting root problems affecting two vicious cycles. Figure 9 [Figure 9: see original paper] shows the query for detecting root problems included in two vicious cycles. Figure 10 [Figure 10: see original paper] shows the graph patterns: the left side shows root problems from Figure 8' s query, the right side from Figure 9' s query. With vicious-cycle nodes set between 3 and 6, we obtained 144 graph patterns of root problems and detected 28 root problems.

5.4 Evaluation of the Detected Vicious Cycles

In our previous study [26], we defined vicious cycles as consisting only of correct causal relations, assuming relations in official government documents were correct. However, governments publish limited information, making the correct relations dataset incomplete. Therefore, we evaluated causal relations in cooperation with experts working on various urban problems, including homelessness and crime prevention specialists and Osaka Citizens Bureau representatives. We evaluated results related to homelessness and crime through questionnaires and interviews. Figure 11 [Figure 11: see original paper] depicts the interview conducted at Osaka City Citizens Bureau on January 25, 2018, with six experts from NPOs, companies, the Institute for Municipal Research, and the Osaka City Citizens Bureau.

Experts rated 194 causal relations related to homelessness and crime with four options: (1) The extracted causal relation is true; (2) The extracted causal relation might be true (including new knowledge); (3) The extracted causal relation is false; (4) Additional causal relations were not extracted but should be added. Experts chose Option (1) 21 times, Option (2) 154 times, Option (3) 5 times, and Option (4) 14 times (answers published on our website).

For example, experts classified "Multiple debt → Homeless" as Option 1 and "Homelessness → Environmental pollution" as Option 2, based on homeless people scattering plastic trash when collecting scraps. They classified "Population aging → Homelessness" as Option 3 because aging among homeless people is a problem but not a factor causing homelessness. As an Option 4 example, one expert commented that nuclear family structure is a crime factor, but we could not extract "nuclear family" as a crime factor. Option (2) "might be true" means "Experts knew it but could not affirm confidently" or "Experts did not know it but consider it new possibilities." Selecting this option indicates that causal reasoning offered new insights, allowing experts to consider problems based on hypotheses from causal relations—precisely our study' s purpose. Thus, the evaluation can be interpreted as successful.

We then evaluated vicious cycle accuracy based on these results, defining vicious cycles as consisting of causal relations from Options 1 or 2. Consequently, 196 detected vicious cycles related to homelessness and crime were correctly extracted. For example, "Poverty → Poverty business → Disease" was detected as a possible vicious cycle. The temporary staffing business of day labor is a poverty business exploiting vulnerable people [27]. Since day laborers cannot earn stable income, they might become homeless. Homelessness may affect Disease, which may affect Poverty again, as increasing hospitalization expenses might lead to poverty. However, long-term surveys are needed to determine if these vicious cycles occur in reality.

5.5 Evaluation of Detecting Root Problems

We found that illegally parked bicycles can affect traffic accidents and traffic jams, forming vicious cycles: "Traffic accident → Stress" and "Littering → Graffiti." This problem has been identified as a factor in traffic accidents, safety security, and many other urban problems by several city bureaus in Japan and possibly other Asian countries. Since illegally parked bicycles constitute a root problem, solving it might make a large positive impact.

We searched for budget information related to root problems leading to multiple vicious cycles. For example, the truancy problem could lead to vicious cycles "Poverty → Deteriorated security → Poverty business → Homelessness → Day labor" (consisting only of Options 1 and 2) and "Deterioration of security → Graffiti → Thief." Truancy may lead to future poverty and deteriorated security. Graffiti gives the public the impression that local government is not functioning, leading to crimes like theft—a phenomenon known as broken window theory [25]. Experts agreed that truancy and lack of educational opportunities are root problems. However, according to Osaka city manager budget data, the budget for solving truancy in Abeno ward was only 15,000 JPY (Figure 12 [Figure 12: see original paper] shows the SPARQL query). The average budget for solving truancy in other wards was 15,083 JPY. Therefore, increasing these budgets could reduce the risk of children entering vicious cycles.

Finally, experts and the Osaka City Citizens Bureau agreed that "This KG is useful for recognizing the overview of urban problems, as urban-problem experts sometimes have a certain mind-set"and "This KG is useful as a tool for improving discussions."

6. CONCLUSION

This paper described an ontology for urban-problem causality and budget examination, built a KG based on the ontology, and used it to detect vicious cycles through SPARQL and inference rules. We evaluated results with six urban-problem experts. To understand which problems should be resolved first, we proposed SPARQL patterns for detecting root problems and discussed detection results using budget information.

We constructed urban-problem causality based on government documents, sociology articles, and social opinions from the Web. Future work will consider adding probabilities of causal relations as numerical values.

ACKNOWLEDGEMENTS

This work was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI (No. 16K12411, No. 16K00419, No. 16K12533, No. 17H04705, and No. 18J13988).

AUTHOR CONTRIBUTIONS

S. Egami (s-egami@aist.go.jp), T. Kawamura (takahiro.kawamura@naro.go.jp), K. Kozaki (kozaki@osakac.ac.jp), and A. Ohsuga (ohsuga@is.uec.ac.jp) proposed research problems, performed research, designed the framework, collected and analyzed data, and wrote and revised the manuscript.

REFERENCES

[1] Etcheverry, L., Vaisman, A., Zimányi, E.: Modeling and querying data warehouses on the semantic Web using QB4OLAP. In: The 16th International Conference on Data Warehousing and Knowledge Discovery (DaWaK), pp. 45-56 (2014)

[2] Szekely, P., et al.: Building and using a knowledge graph to combat human trafficking. In: The 14th International Semantic Web Conference (ISWC), pp. 205-221 (2015)

[3] Egami, S., Kawamura, T., Ohsuga, A.: Building urban LOD for solving illegally parked bicycles in Tokyo. In: The 15th International Semantic Web Conference (ISWC), pp. 291-307 (2016)

[4] Egami, S., et al.: Construction of linked urban problem data with causal relations using crowdsourcing. In: The 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 814-819 (2017)

[5] Egami, S., et al.: Linked urban open data including social problems' causality and their cost. In: The 7th Joint International Semantic Technology Conference (JIST), pp. 334-349 (2017)

[6] Shiramatsu, S., et al.: Towards continuous collaboration on civic tech projects: Use cases of a goal sharing system based on linked open data. In: The 7th IFIP International Conference on Electronic Participation (ePart), pp. 81-92 (2015)

[7] Santos, H., et al.: From data to city indicators: A knowledge graph for supporting automatic generation of dashboards. In: The 14th Extended Semantic Web Conference (ESWC), pp. 94-108 (2017)

[8] Pileggi, S.F., Hunter, J.: An ontological approach to dynamic fine-grained urban indicators. In: The 17th International Conference on Computational Science (ICCS), pp. 2059-2068 (2017)

[9] Höffner, K., Martin, M., Lehmann, J.: Linkedspending: Openspending becomes linked open data. Semantic Web Journal 7(1), 95-104 (2016)

[10] Demartini, G., Difallah, D.E., Cudré-Mauroux, P.: Large-scale linked data integration using probabilistic reasoning and crowdsourcing. The International Journal on Very Large Data Bases 22(5), 665-687 (2013)

[11] Celino, I., et al.: Linking smart cities datasets with human computation—the case of urbanmatch. In: The 11th International Semantic Web Conference (ISWC), pp. 34-49 (2011)

[12] Ahn, L.V.: Games with a purpose. IEEE Computer 39(6), 92-94 (2006)

[13] Nguyen, T.M., et al.: Self-supervised capturing of users' activities from weblogs. International Journal of Intelligent Information and Database Systems 6(1), 61-76 (2012)

[14] Augenstein, I., Padó, S., Rudolph, S.: Lodifier: Generating linked data from unstructured text. In: The 9th Extended Semantic Web Conference (ESWC), pp. 210–224 (2012)

[15] Milne, D., Witten, I.H.: Learning to link with Wikipedia. In: The 17th ACM Conference on Information and Knowledge Management (CIKM), pp. 509–518 (2008)

[16] Strauss, B., et al.: Results of the WNUT16 named entity recognition shared task. In: The 2nd Workshop on Noisy User-generated Text (WNUT), pp. 138–144 (2016)

[17] Moitra, A., et al.: Semantic inference for pharmacokinetic drug-drug interactions. In: The 8th IEEE International Conference on Semantic Computing (ICSC), pp. 92–95 (2014)

[18] Herrero-Zazo, M., et al.: Dinto: Using OWL ontologies and SWRL rules to infer drug-drug interactions and their mechanisms. Journal of Chemical Information and Modeling 55(8), 1698–1707 (2015)

[19] Kudo, T., Matsumoto, Y.: Japanese dependency analysis using cascaded chunking. In: The 6th Conference on Natural Language Learning (CoNLL), pp. 1–7 (2002)

[20] Winkler, W.: The state of record linkage and current research problems. Technical report, Statistics of Income Division, Internal Revenue Service Publication (1999)

[21] Ponzetto, S.P., Strube, M.: Deriving a large scale taxonomy from Wikipedia. In: The 22nd AAAI Conference on Artificial Intelligence (AAAI), pp. 1440–1445 (2007)

[22] Kontokostas, D., et al.: Test-driven evaluation of linked data quality. In: The 23rd International Conference on World Wide Web (WWW), pp. 747–758 (2014)

[23] Fleiss, J.L., Cohen, J.: The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement 33(3), 613–619 (1973)

[24] Viera, A.J., Garrett, J.M.: Understanding interobserver agreement: The kappa statistic. Family Medicine 37(5), 360–363 (2005)

[25] Wilson, J.Q., George, L.K.: Broken windows. In: Dunham, R.G., Albert, G.P. (eds.) Critical Issues in Policing: Contemporary Readings, pp. 369–381 (1989)

[26] Egami, S., et al.: Urban problem LOD for understanding the problem structure and detecting vicious cycles. In: The 12th IEEE International Conference on Semantic Computing (ICSC), pp. 186–193 (2018)

[27] Sekine, Y.: The rise of poverty in Japan: The emergence of the working poor. Japan Labor Review 5(4), 49–66 (2008)

AUTHOR BIOGRAPHY

Shusaku Egami received his Ph.D. in Computer Science from the University of Electro-Communications, Tokyo, Japan, in 2019. He is currently a researcher at the National Institute of Advanced Industrial Science and Technology, Japan, a part-time lecturer at Hosei University, Tokyo, and a collaborative researcher at the University of Electro-Communications. His research interests include the semantic Web, knowledge graphs, and ontology. ORCID: 0000-0002-3821-6507

Takahiro Kawamura is an Administrator of Data Management at the National Agriculture and Food Research Organization, a member of the Evaluation Committee at Riken BioResource Research Center, a Visiting Researcher at the Artificial Intelligence Research Center at AIST, and a Part-time Lecturer in the Graduate School of Engineering at Osaka University. He received M.S. and Ph.D. degrees in electrical engineering from Waseda University and was a Visiting Scholar at Carnegie Mellon University' s Robotics Institute. His research interests include knowledge graph construction and analysis combining semantic technology and machine learning. He served on the board of the Japanese Society for Artificial Intelligence (JSAI) and as chair of its SIG on Semantic Web and Ontology, and was a W3C Advisory Committee Representative. ORCID: 0000-0002-2765-6232

Kouji Kozaki is a professor at Osaka Electro-Communication University (OECU), Osaka, Japan. His expertise is in Computer Science, particularly Ontology Engineering and Semantic Technologies. His research interests include ontology building, fundamental theory of ontological engineering (especially for role and identity), ontology development systems, and ontology-based applications. A major research theme is Hozo (http://www.hozo.jp/), an environment for building/using ontologies in domains such as clinical medicine, bioinformatics, and environmental engineering. He was co-Program Committee chair of JIST 2015 and co-General Chair of JIST2018. ORCID: 0000-0003-4578-4980

Akihiko Ohsuga received his Ph.D. in Computer Science from Waseda University in 1995. From 1981 to 2007 he was with Toshiba Corporation. He joined the University of Electro-Communications in 2007 and is currently a professor in the Graduate School of Informatics and Engineering and a visiting professor at the National Institute of Informatics. His research interests include agent technologies, Web intelligence, and software engineering. He is a member of IEEE CS, IPSJ, IEICE, JSAI, JSSST, and IEEJ, and an IPSJ fellow since 2017. He served as chair of IEEE CS Japan Chapter and on boards of JSAI and JSSST, receiving IPSJ Best Paper Awards in 1987 and 2017. ORCID: 0000-0001-6717-7028

Note: Figure translations are in progress. See original paper for figures.