JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice
eISSN : 2287-4577 pISSN : 2287-9099

Research Paper
J Inf Sci Theory Pract 6(3): 06-15, 2018
https://doi.org/10.1633/JISTaP.2018.6.3.1

A Combinational Method to Determining Identical Entities from Heterogeneous Knowledge Graphs

Haklae Kim*

Korea Institute of Science and Technology Information,
Daejeon, Korea
E-mail: haklaekim@gmail.com

ABSTRACT

With the increasing demand for intelligent services, knowledge graph technologies have attracted much attention. Various application-specific knowledge bases have been developed in industry and academia. In particular, open knowledge bases play an important role for constructing a new knowledge base by serving as a reference data source. However, identifying the same entities among heterogeneous knowledge sources is not trivial. This study focuses on extracting and determining exact and precise entities, which is essential for merging and fusing various knowledge sources. To achieve this, several algorithms for extracting the same entities are proposed and then their performance is evaluated using real-world knowledge sources.

Keywords: entity consolidation, knowledge extraction, knowledge graph, knowledge creation, knowledge interlinking

Open Access

Accepted date: July 09, 2018
Received date: December 07, 2017

***Corresponding Author:** Haklae Kim
Senior Researcher
Korea Institute of Science and Technology Information, 245 Daehak-ro, Yuseong-gu, Daejeon, 34141, Korea
E-mail: haklaekim@gmail.com

1. INTRODUCTION

With the increasing demand for intelligent services, knowledge graph technologies have attracted much attention for applications, ranging from question-answer systems to enterprise data integration (Gabrilovich & Usunier, 2016). A number of research efforts have already developed open knowledge bases such as DBpedia (Lehmann et al., 2009), Wikidata (Vrandecic, 2012), YAGO (Suchanek, Kasneci, & Weikum, 2007), and Freebase (Bollacker, Evans, Paritosh, Sturge, & Taylor, 2008). Most open knowledge bases heavily use Linked Data technologies for constructing, publishing, and accessing knowledge sources. Linked data is one of the core concepts of the Semantic Web, also called the Web of Data (Bizer, Cyganiak, & Heath, 2007; Gottron & Staab, 2014). It involves making relationships such as links between datasets understandable to both humans and machines. Technically, it is essentially a set of design principles for sharing machine-readable interlinked data on the Web (Berners-Lee, 2009). According to LODstats,[1] 149B triples from 2,973 datasets have been published in public, and 1,799,869 identical entity relations have already been made from 251 datasets. The standard method for stating a set of the same entities is to use the *owl:same* property. This property is used to describe homogeneous instances that refer to the same object in the real world. It aims to indicate that two uniform resource identifier (URI) references actually refer to the same thing (Berners-Lee, 2009).

Existing knowledge bases can be used to construct new ones to meet certain objectives, since constructing a new knowledge base from scratch is not easy. However, various issues arise when creating a new knowledge base by integrating multiple knowledge sources. One issue is whether the relationships in the existing knowledge base are always reliable. All individual instances of given knowledge sources should be identified and linked to these sources before integrating knowledge sources (Halpin, Hayes, McCusker, McGuinness, & Thompson, 2010). The problem of discovering the same entities in various data sources has been studied extensively; it is variously referred to as entity reconciliation (Enríquez, Mayo, Cuaresma, Ross, & Staples, 2017), entity resolution (Stefanidis, Efthymiou, Herschel, & Christophides, 2014), entity consolidation (Hogan, Zimmermann, Umbrich, Polleres, & Decker, 2012), and instance matching (Castano, Ferrara, Montanelli, & Lorusso, 2008). All of these approaches are very important for identifying the same relationships to extract and generate knowledge from different data sets. Entity consolidation for data integration at the instance level has attracted interest in the semantic web and linked data communities. It refers to the process of identifying same entities across heterogeneous data sources (Hogan et al., 2012). A problem can be simplified such that different identifiers are used for identical entities scattered across different datasets in a web of data. Because redundancy causes an increase in noisy or unnecessary information across a distributed web of data, identifying the same items can be advantageous in that multiple descriptions of the same entity can mutually complete and complement each other (Enríquez et al., 2017).

This study proposes a combinational approach for extracting and determining same entities from heterogeneous knowledge sources. It focuses on extracting exact and precise entity linkages, which is the key to merging and fusing various knowledge sources into new knowledge. The remainder of this paper is organized as follows. Section 2 presents a literature review of related works. Section 3 introduces research methods and basic principles of defining an entity pair from multiple knowledge bases. Section 4 introduces a formal model for entity consolidation and presents several strategies for extracting and identifying same entities. Section 5 introduces implementations of proposed strategies with some examples. Section 6 addresses and discusses findings from the evaluation using real-world knowledge bases. Section 7 concludes this study and discusses future work.

2. RELATED WORK

A number of open knowledge bases already exist such as DBpedia, Freebase, Wikidata, and YAGO (Paulheim, 2017). Wikidata (Vrandecic, 2012) is a knowledge base about the world that can be read and edited by humans and machines with the Creative Commons Zero license (CC-0).[2] Information from Wikidata is called items, which are comprised of labels, descriptions, and aliases in all languages of Wikipedia. Wikidata does not aim to offer a single truth about things; instead, it provides statements given in a particular context. DBpedia (Lehmann et al., 2009) is a structured, multilingual knowledge set from Wikipedia and is made freely available on the Web using semantic web and linked data technologies. It has developed into the central

[1] http://lodstats.aksw.org/stats

[2] https://creativecommons.org/choose

interlinking hub in the Web of linked data, because it covers a wide variety of topics and sets resource data framework (RDF) links pointing to various external data sources. Freebase (Bollacker, Evans, Paritosh, Sturge, & Taylor, 2008) was a large collaborative and structured knowledge base harvested from diverse data sources. It aimed to create a global resource graph that allowed human and machines to access common knowledge more effectively. Google developed a Knowledge Graph using Freebase. On the other hand, Knowledge Vault is developed by Google to extract facts, in the form of disambiguated triples, from the entire web (Dong et al., 2014). The main difference from other works is that it fuses together facts extracted from text with prior knowledge derived from the Freebase graph. YAGO (Suchanek et al., 2007) fuses multilingual knowledge with English WordNet to build a coherent knowledge base from Wikipedia in multiple languages.

Färber, Ell, Menne, and Rettinger (2015) analyses existing knowledge graphs based on 35 characteristics, including general information (e.g., version, languages, or covered domains), format and representation (e.g., dataset formats, dynamicity, or query languages), genesis and usage (e.g., provenance of facts, influence on other linked open data [LOD] datasets), entities (e.g., entity reference, LOD registration and linkage), relations (e.g., reference, relevance, or description of relations), and schema (e.g., restrictions, constraints, network of relations). According to the comparison of entities, most knowledge graphs provide human-readable identifiers, however, Wikidata provides entity identifiers, which consists of "Q" followed by a specific number (Wang, Mao, Wang, & Guo, 2017). Most knowledge graphs are published in RDF and link their entities to entities of other datasets in LOD cloud.[3] In particular, DBpedia and Freebase have a high degree of connectivity with other LOD datasets.

Note that Google recently announced that it transferred data from Freebase to Wikidata, and it launched a new API for entity search powered by Google's Knowledge Graph. Mapping tools[4] have been provided to increase the transparency of the publication process of Freebase content to integrate into Wikidata. Tanon, Vrandecic, Schaffert, Steiner, and Pintscher (2016) provided a method for migrating from Freebase to Wikidata with some limitations, including entity linking and schema mapping. This study provides comprehensive entity extraction techniques for interlinking from two knowledge sources. However, identifying same entities from knowledge sources is not enough to integrating two knowledge bases. Various studies have investigated pragmatic issues of *owl:sameAs* in the context of the Web of Data (Halpin et al., 2010; Ding, Shinavier, Shangguan, & McGuinness, 2010; Hogan et al., 2012; Idrissou, Hoekstra, van Harmelen, Khalili, & den Besselaar, 2017). In particular, Hogan et al. (2012) discuss scalable and distributed methods for entity consolidation to locate and process names that signify the same entity. They calculate weighted concurrence measures between entities in the Linked Data corpus based on shared inlinks/outlinks and attribute values using statistical analyses. This paper proposes a combinational approach to extract identical entity pairs from heterogeneous knowledge sources.

[3] http://lod-cloud.net/

[4] https://github.com/google/freebase-wikidata-converter

3. METHODOLOGY

3.1. Research Approach

This study proposes a method for extracting a set of identical entities from heterogeneous knowledge sources. An identical relationship of entities is based on calculating the properties and its values of the entities. The analysis is performed through a combination of several methods called 'strategy.' In this paper, five strategies are introduced and are combined for extracting and verifying identical relationships of entities. Each strategy has its own advantages and disadvantages. For example, a consistency strategy is a simple method for extracting entities, but it returns high ambiguities as noise to some extent, whereas a max confidence strategy delivers reduced ambiguities by calculating a confidence score of entity pairs. Although the max confidence method would be useful for extracting entity pairs compared to the consistency method, the max confidence strategy is based on the entity pairs extracted by the consistency one. Therefore, each strategy can be used for individual purposes, and also can be applied to determine a high quality of identical entity pairs by combining several strategies.

3.2. A Formal Model of an Entity Pair

Let knowledge bases K_1 and K_2 contain a set of entities and properties, respectively. The set of entities is $K_i^E = \{K_i^{e_1}, \ldots, K_i^{e_n}\}$ and the set of properties in K_i is $K_i^P = \{K_i^{P_1}, \ldots, K_i^{P_n}\}$. In addition, let $K_i^O = \{K_i^C, \ldots, K_i^P\}$ be the ontology schema of K_i, where K_i^C is the set of classes and K_i^P is the set of properties. Thus, entity pairs $\boldsymbol{EP}_{(K_1,K_2)}$ as a set of identical entities for given knowledge bases K_1 and K_2 are denoted as follows:

$$\boldsymbol{EP}_{(K_1,K_2)} = \{(K_1^{e_1}, K_1^{e_j}), \ldots, (K_1^{e_s}, K_2^{e_t})\}$$

where $K_i^{e_1}$ is identical to $K_2^{e_j}$. On the other hand, the schema alignment K^O is aligned to its schemas:

$$K^O = K_1^O \overset{\text{align}}{\Longleftrightarrow} K_2^O$$

where $K_1^C \overset{\text{align}}{\Longleftrightarrow} K_2^C$ is the class alignment and $K_1^P \overset{\text{align}}{\Longleftrightarrow} K_2^P$ is the property alignment for K_1 and K_2. In this sense, $K_1^{C_i} \overset{align}{\Longleftrightarrow} K_2^{C_j}$ means that $K_1^{C_i}$ is identical to $K_2^{C_j}$, and $K_1^{P_i} \overset{align}{\Longleftrightarrow} K_2^{P_j}$ means that the value of P_i in K_1 corresponds to that of P_j in K_2. Thus, according to K^0, a set of property mappings to the matching keys is defined as follows:

$$\boldsymbol{MK}_{(K_1,K_2)} = \{(K_1^{P_i}, K_2^{P_j}), \ldots, (K_1^{P_s}, K_2^{P_t})\}$$

4. STRATEGIES FOR ENTITY CONSOLIDATION

A number of approaches is available for identifying the same entities from heterogeneous knowledge bases (Hors & Speicher, 2014; Nguyen & Ichise, 2016; Moaawad, Mokhtar, & al Feel, 2017). This section addresses some methods to determine identical relationships from the extracted entities. Note that formal models of four strategies are introduced and their characteristics are also discussed.

4.1. Consistency Strategy

This strategy aims to extract a set of precise entities by mapping property values on specific knowledge bases. That is, to determine the consistency of $K_1^{e_i}$ and $K_2^{e_j}$ based on matching keys $\boldsymbol{MK}$, two strategies, S_I and S_U, are defined:

Strategy S_I: For $K_1^{e_m}$ and $K_2^{e_n}$ from K_1 and K_2, $\forall(K_1^{P_i}, K_2^{P_j}) \in \boldsymbol{MK}_{(K_1,K_2)}$, the $K_1^{P_i}$ value of $K_1^{e_m}$ is exactly equal to the $K_2^{P_j}$ value of $K_2^{P_j}$. Then, $(K_1^{e_m}, K_2^{e_n})$ is an identical entity pair, and the consistency determination is of the intersection strategy S_I.

Strategy S_U: For $K_1^{e_m}$ and $K_2^{e_n}$ from K_1 and K_2, $\exists(K_1^{P_i}, K_2^{P_j}) \in \boldsymbol{MK}_{(K_1,K_2)}$, the $K_1^{P_i}$ value of $K_1^{e_m}$ is exactly equal to the $K_2^{P_j}$ value of $K_2^{e_n}$. Then, $(K_1^{e_m}, K_2^{e_n})$ is an identical entity pair, and the consistency determination is of the union strategy S_U.

This strategy is based on the assumption that all knowledge sources are trustworthy: The knowledge in K_i is precise and without defect. The identical relations $\boldsymbol{EP}_{(K_1,K_2)}$ extracted by this strategy are considered precise because the mapping of the property values is exact without bias. On the contrary, most open knowledge bases contain some defects which may be caused by false recognition, inaccurate source, or knowledge redundancy. Note that one entity can be interlinked to multiple entities of different knowledge sources (e.g. $\exists(K_1^{e_i}, K_2^{e_j})$, $(K_1^{e_s}, K_2^{e_t}) \in \boldsymbol{EP}_{(K_1,K_2)}$, and $K_1^{e_i} = K_1^{e_s}$ and $K_2^{e_j} \neq K_2^{e_t}$). This ambiguous pair might arise from a defect in the knowledge base K_i. For establishing high-quality linkages across heterogeneous knowledge sources, it is essential to extract confident $\boldsymbol{EP}_{(K_1,K_2)}$ by eliminating ambiguities to the greatest extent possible. Therefore, alternative strategies are proposed.

4.2. Max Confidence Strategy

This strategy calculates a confidence score for the entity pairs extracted by the consistency strategy to reduce the noise caused by defects and determines precise and confident entity pairs. The formal notation of this strategy is defined as follows:

Given matching keys $\boldsymbol{MK}_{(K_1,K_2)} = \{(K_1^{P_i}, K_2^{P_j}), \ldots, (K_1^{P_s}, K_2^{P_t})\}$, for $K_1^{e_m} \in K_1^E$ and $K_2^{e_n} \in K_2^E$, let $\boldsymbol{MK}_m = \{(K_1^{P_{im}}, K_2^{P_{jm}}), \ldots, (K_1^{P_{sm}}, K_2^{P_{tm}})\}$ be the matched $\boldsymbol{MK}_{(K_1,K_2)}$, where $(K_1^{P_{im}}, K_2^{P_{jm}})$ indicates that the $K_1^{P_{im}}$ value of $K_1^{e_m}$ is exactly equal to the $K_2^{P_{jm}}$ value of $K_2^{e_n}$. Based on this, $\boldsymbol{MK}_m$ and $\boldsymbol{MK}_{(K_1,K_2)}$ can be defined as $\boldsymbol{MK}_m \subseteq \boldsymbol{MK}_{(K_1,K_2)}$, then a confidence score of $(K_1^{e_m}, K_2^{e_n})$ is calculated by the following equation:

$$\boldsymbol{conf}(K_1^{e_m}, K_2^{e_n}) = ||\boldsymbol{MK}_m|| \,/\, ||\boldsymbol{MK}_{(K_1,K_2)}||$$

where $||\cdot||$ is the cardinality. Therefore, a confidence score is assigned to each entity pair, and for $(K_1^{e_i}, K_2^{e_j}), (K_1^{e_s}, K_2^{e_t}) \in \boldsymbol{EP}_{(K_1,K_2)}$. Therefore, $(K_1^{e_i}, K_2^{e_j})$ is the confident identical entity pair where $conf(K_1^{e_i}, K_2^{e_j}) > conf(K_1^{e_s}, K_2^{e_t})$.

4.3. Threshold Filtering Strategy

The Max Confidence allows to filter out ambiguous same entity pairs; nonetheless, some of entity pairs may have relatively high scores with low confidence levels. To solve this issue, a threshold is added to the extraction process: If an entity pair has determined with the highest confidence and it has a low score compared to other scores, it can be removed from a set of candidates. The threshold filtering strategy aims to improve a confidence level of extracted entity pairs by using a threshold score. Given a threshold θ for $(K_1^{e_i}, K_2^{e_j}), (K_1^{e_i}, K_2^{e_t}) \in \boldsymbol{EP}_{(K_1,K_2)}$, where $conf(K_1^{e_i}, K_2^{e_j}) > conf(K_1^{e_i}, K_2^{e_t})$ and $conf(K_1^{e_i}, K_2^{e_j}) \geq \theta$, $(K_1^{e_i}, K_2^{e_j})$ is selected as the confident same entity pair.

4.4. One-to-One Mapping Strategy

This strategy extracts simply 1-1 entity pairs from heterogeneous knowledge sources by ignoring multiple

relations in which one identifier is matched to multiple identifiers of different sources. Formally, it is represented as $\forall(K_1^{e_i}, K_2^{e_j}) \in \boldsymbol{EP}_{(K_1,K_2)}, \nexists(K_1^{e_s}, K_2^{e_t}) \in \boldsymbol{EP}_{(K_1,K_2)}$ where $i = s$ or $j = t$. By applying one-to-one mapping, identical entity pairs $\boldsymbol{EP}_{(K_1,K_2)}$ have no ambiguous relations.

4.5. Belief-based Strategy

The four strategies introduced so far focus on inter-relations between entity pairs by comparing properties of knowledge bases, whereas they do not consider intra-relations in a certain pair. In other words, property values of entities in a certain pair should be checked for determining identical relations. The belief-based strategy aims to analyse property values of extracted entity pairs that is based on the Dempster-Shafer theory (Yager, 1987), also called the theory of evidence.

Given a set of same entity pairs $\boldsymbol{EP}$, let X_{EP} denote the set representing all possible states of an entity pair. Here, two cases are possible: The two entities are linked (L) or the two entities are not linked (U). Note that $X_{EP} = \{L, U\}$. Then, $\Omega_{X_{EP}} = \{\Phi, L, U, \{L, U\}\}$, where Φ indicates the empty set, and $\{L, U\}$ indicates that it is uncertain whether they are linked. Therefore, a belief degree is assigned to each element of $\Omega_{X_{EP}}$:

$$m: \Omega_{X_{EP}} \rightarrow [0,1] \tag{1}$$

where m is the degree of same belief, which is the basic belief assignment in the Dempster-Shafer theory. Then, each pair of knowledge sources has four hypotheses, and the formal model is represented as follows:

$$m(\Phi) = 0 \tag{2}$$

Given $\boldsymbol{MK}_{(K_1,K_2)} = \{(K_1^{P_i}, K_2^{P_j}), \ldots, (K_1^{P_s}, K_2^{P_t})\}$ and $\boldsymbol{MK}_{(K_1,K_2)} = \{(K_1^{P_{im}}, K_2^{P_{jm}}), \ldots, (K_1^{P_{sm}}, K_2^{P_{tm}})\}$, m is assigned as follows:

$$m(\{L\}) = \|\boldsymbol{MK}_m\| \,/\, \|\boldsymbol{MK}_{(K_1,K_2)}\| \tag{3}$$

$$K_1^{e_m} \|\boldsymbol{MK}_{um}\| \,/\, \|\boldsymbol{MK}_{(K_1,K_2)}\| \tag{4}$$

where $\boldsymbol{MK}_{um}$ represents the unmatched $\boldsymbol{MK}_{(K_1,K_2)}$, that is, the $K_1^{P_i}$ value of $K_1^{e_m}$ is not equal to the $K_2^{P_j}$ value of $K_2^{e_n}$. And uncertain pairs of knowledge sources are calculated by the following model:

$$m(\{L, U\}) = 1 - m(\{L\}) - m(\{U\}) \tag{5}$$

According to the theory of evidence, the basic belief assignment $m(A)$, $A \in \Omega$, expresses the proportion of all relevant and available evidence that supports the claim that the actual state belongs to A. In this sense, a degree of belief is represented as a belief function rather than a Bayesian probability distribution.

5. IMPLEMENTATION OF THE BELIEF-BASED STRATEGY

The proposed strategies are developed in the entity extraction framework (Kim, Liang, & Ying, 2014), which is to extract identical entities among heterogeneous knowledge sources. In particular, entity matching is carried out by configured property values for each entity pair. As illustrated in Fig. 1, it is comprised of several components: *Preprocessor* for normalising entities and properties and to extract a set of URI from knowledge sources, Matching for extracted entities and properties based on exact and similarity measure, *Optimization* for better extracting a set of same entity pairs using several strategies, and *Knowledge Base Management* that aims to create and interlink a knowledge base for the consolidation results.

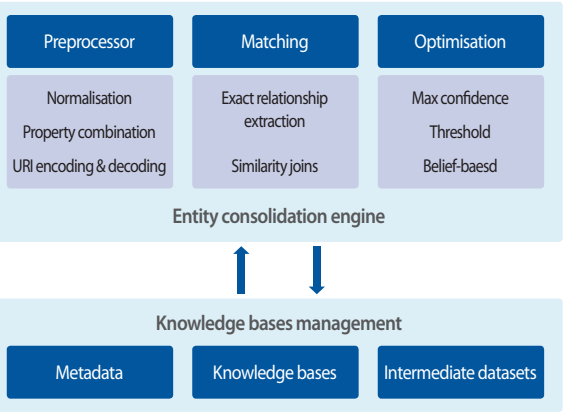

Fig. 1. Entity extraction framework. URI, uniform resource identifier.

Currently, this framework is being used for extracting relations from both Wikidata and Freebase. To identify the same entities from both knowledge sources, Wikipedia is the primary data source used to detect relations between Freebase and Wikidata. Therefore, for detecting source errors and identifying exact identical relationships, four strategies are implemented. In particular, those strategies are fully implemented in this framework: for example, the workflow of entity consolidation based on the Max Confidence as shown in Fig. 2. It is designed to compute the Max Confidence for entity consolidation to reduce the noise

$EP \leftarrow$ the set of entity pairs of data sources (i.e. F and W)
$EP \leftarrow \emptyset$
foreach $K_F^{e_i}$ in K_F **do**
  $K_W^{e_{ep}} \leftarrow$ the match entity of $K_F^{e_i}$ in K_F,
  $K_W^{e_{ep}} \leftarrow \emptyset$
  $EP_{K_F}^{e_{ep}} \leftarrow$ the set of all candidates in K_W for $K_F^{e_i}$
  foreach *each pair* $\left(K_F^{p_s}, K_W^{p_t}\right)$ in MK **do**
    Find all the potential same entity candidates $K_W^{e_j}$
    $EP_{K_F}^{e_i} \leftarrow append\ K_W^{e_j}$
  end
  $Max_{conf} \leftarrow \emptyset$
  foreach $K_W^{e_j}$ in $EP_{K_F}^{e_i}$ **do**
    $Compute\ Conf\left(K_F^{e_i}, K_W^{e_j}\right)$
    if $Conf\left(K_F^{e_i}, K_W^{e_j}\right) > Max_{conf}$ **then**
      $Max_{conf} \leftarrow Conf\left(K_F^{e_i}, K_W^{e_j}\right)$
      $K_W^{e_{ep}} \leftarrow K_W^{e_j}$
    end
  end
  if $K_W^{e_{ep}} \neq \emptyset$ **then**
    $EP \leftarrow$ append $\left(K_F^{e_i}, K_W^{e_{ep}}\right)$

Fig. 2. The algorithm of the Max Confidence strategy.

caused by defects and to obtain precise and confident same entity pairs.

For the threshold strategy, a threshold score is set as 0.5 by default. After eliminating a set of pairs under the threshold score, the Max Confidence approach is applied. Furthermore, the belief-based approach is developed and applied by using the same datasets. As shown in Table 1, for the Persian soldier *Pharnabazus II* (https://en.wikipedia.org/wiki/Pharnabazus_II), Freebase (http://rdf.freebase.com/ns/m.01d89y) has 8 Wikipedia links whereas Wikidata (https://www.wikidata.org/wiki/Q458256) has 20 Wikipedia links in Table 2. Note that the belief-based approach for the case shown in Tables 1 and 2 can be calculated as follows:

$$mass(\{\Phi\}) = 0$$
$$mass(\{\text{Link}\}) = \textit{Matched Wikipedia Link Number} / \textit{Total Wikipedia Link Number}$$
$$mass(\{\text{Unlink}\}) = \textit{Unmatched Wikipedia Link Number} / \textit{Total Wikipedia Link Number}$$
$$mass(\{\text{Link, Unlink}\}) = 1 - mass(\{\Phi\}) - mass(\{\text{Link}\}) - mass(\{\text{Unlink}\})$$

There are matched and unmatched links compared to the given identifiers based on a Wikipedia link. On the other hand, both Wikidata and Freebase do not have the corresponding links. In this case, the status is uncertain. Therefore, the belief-based approach for the given example is calculated:

$$mass(\{\text{Link}\}) = 3/8 = 0.375$$
$$mass(\{\text{Unlink}\}) = 5/8 = 0.625$$
$$mass(\{\text{Link, Unlink}\}) = 0$$

As a result, for entity '*m.01d89y*,' the belief degree for unlinking with entity '*Q458256*' is much greater than the belief degree for linking. Therefore, we consider that '*m.01d89y*' is different from '*Q458256*.'

Table 1. An example of Freebase entity

Identifier	Wikipedia language	Wikipedia link	Matched
m.01d89y	en	http://en.wikipedia.org/wiki/Pharnabazos_II,_Satrap_of_Phrygia	Unmatched
	es	http://es.wikipedia.org/wiki/Farnabazo_I	Unmatched
	it	http://it.wikipedia.org/wiki/Farnabazo_II	Matched (1)
	ja	http://ja.wikipedia.org/wiki/%E3%83%95%E3%82%A1%E3%83%AB%E3%83%8A%E3%83%90%E3%82%BE%E3%82%B9_%28%E3%82%A2%E3%83%AB%E3%82%BF%E3%83%90%E3%82%BE%E3%82%B9%E3%81%AE%E5%AD%90%29	Unmatched
	ca	http://ca.wikipedia.org/wiki/Farnabazos_I	Unmatched
	he	http://he.wikipedia.org/wiki/%D7%A4%D7%A8%D7%A0%D7%91%D7%96%D7%95%D7%A1_%D7%94%D7%A9%D7%A0%D7%99	Matched (2)
	hr	http://hr.wikipedia.org/wiki/Farnabaz_I.	Unmatched
	el	http://el.wikipedia.org/wiki/%CE%A6%CE%B1%CF%81%CE%BD%CE%AC%CE%B2%CE%B1%CE%B6%CE%BF%CF%82_%CE%92%CE%84	Matched (3)

The full uniform resource identifier of Freebase entity has 'http://rdf.freebase.com/ns/' with identifier, i.e., http://rdf.freebase.com/ns/m.01d89y.

Table 2. An example of Wikidata

Identifier	Wikipedia language	Wikipedia link	Matched
Q458256	be_x_old	http://be-x-old.wikipedia.org/wiki/%D0%A4%D0%B0%D1%80%D0%BD%D0%B0%D0%B1%D0%B0%D0%B7_II	Uncertain
	be	http://be.wikipedia.org/wiki/%D0%A4%D0%B0%D1%80%D0%BD%D0%B0%D0%B1%D0%B0%D0%B7_II	Uncertain
	bg	http://bg.wikipedia.org/wiki/%D0%A4%D0%B0%D1%80%D0%BD%D0%B0%D0%B1%D0%B0%D0%B7_II	Uncertain
	ca	http://ca.wikipedia.org/wiki/Farnabazos_II	Unmatched
	de	http://de.wikipedia.org/wiki/Pharnabazos_II.	Uncertain
	it	http://it.wikipedia.org/wiki/Farnabazo_II	Matched (1)
	en	http://en.wikipedia.org/wiki/Pharnabazus_II	Uncertain
	es	http://es.wikipedia.org/wiki/Farnabazo_II	Unmatched
	fr	http://fr.wikipedia.org/wiki/Pharnabaze	Uncertain
	he	http://he.wikipedia.org/wiki/%D7%A4%D7%A8%D7%A0%D7%91%D7%96%D7%95%D7%A1_%D7%94%D7%A9%D7%A0%D7%99	Matched (2)
	hr	http://hr.wikipedia.org/wiki/Farnabaz_II.	Unmatched
	el	http://el.wikipedia.org/wiki/%CE%A6%CE%B1%CF%81%CE%BD%CE%AC%CE%B2%CE%B1%CE%B6%CE%BF%CF%82_%CE%92%CE%84	Matched (3)
	ja	http://ja.wikipedia.org/wiki/%E3%83%95%E3%82%A1%E3%83%AB%E3%83%8A%E3%83%90%E3%82%BE%E3%82%B9_(%E3%83%95%E3%82%A1%E3%83%AB%E3%83%8A%E3%82%B1%E3%82%B9%E3%81%AE%E5%AD%90)	Unmatched
	nl	http://nl.wikipedia.org/wiki/Pharnabazus	Uncertain
	no	http://no.wikipedia.org/wiki/Farnabazos	Uncertain
	pl	http://pl.wikipedia.org/wiki/Farnabazos_II	Uncertain
	ru	http://ru.wikipedia.org/wiki/%D0%A4%D0%B0%D1%80%D0%BD%D0%B0%D0%B1%D0%B0%D0%B7	Uncertain
	sh	http://sh.wikipedia.org/wiki/Farnabaz_II	Uncertain
	sv	http://sv.wikipedia.org/wiki/Farnabazos	Uncertain
	uk	http://uk.wikipedia.org/wiki/%D0%A4%D0%B0%D1%80%D0%BD%D0%B0%D0%B1%D0%B0%D0%B7	Uncertain

The full uniform resource identifier of Freebase entity has 'https://www.wikidata.org/wiki/' with identifier, i.e., https://www.wikidata.org/wiki/Q458256.

6. EVALUATION

6.1. Data Collection

Two knowledge bases (i.e., Wikidata and Freebase) are selected to demonstrate the proposed strategies. Wikidata and Freebase are receiving great attention from academia and industry for constructing their own knowledge bases, and there are realistic issues for data integration between two knowledge sources. It is essential to derive homogeneous entities for knowledge integration, since Wikidata and Freebase have been developed independently. A set of same entities between Freebase (2015-02-10)[5] and Wikidata (2015-02-07) is extracted via their own Wikipedia reference links (i.e., wiki-keys of Freebase and Wikipedia URLs of Wikidata). After pre-processing the collected datasets, 4,446,380 entities from Freebase and 15,403,618 entities from Wikidata are extracted with Wikipedia links. By using the consistency strategy (i.e., S_1), 4,400,955 pairs are obtained from both knowledge sources.

6.2. Results

The aim of applying different approaches for same extraction is to generate links with the highest confidence between Freebase and Wikidata entities. The results differed slightly with the given datasets. Fig. 3. The result of extracting same entities between Freebase and Wikidata.3 illustrates the results obtained using different mapping styles with the proposed strategies. Note that the consistency strategy obtains the largest number of entity pairs. Nonetheless, there are a number of 1-multiple/multiple-1/

[5] https://developers.google.com/freebase/

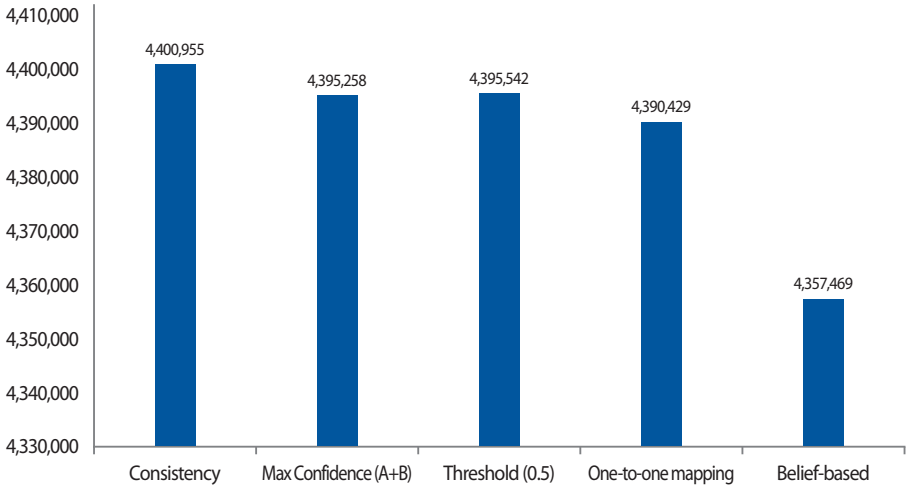

Fig. 3. The result of extracting same entities between Freebase and Wikidata.

multiple-multiple links which cause ambiguities as shown in Table 3. Without applying any approaches, the consistency strategy possesses the largest ambiguity (0.37%). The one-to-one mapping obviously holds the full confident same entity mapping pairs. The Max Confidence, the Threshold Filtering (0.5 threshold), and the Belief-based strategies show great effect on elimination of ambiguity. The number of mapping pairs based on belief degree is approximated to that of Max Confidence. The belief degree greatly influences the reduction in ambiguity in the multiple Freebase case but not in the multiple Wikidata case.

As shown in Table 4, the precision and F1 score are 100 percent for all strategies, because the set of matching pairs is extracted by using the Strategy S_U, whereas both the precision and F1 score are greater than 98.1371 percent, and precision scores are slightly differed among these strategies. Based on this result, a combination of each strategy can reduce some ambiguities that are not removed using a single approach. On the other hand, both the precision and F1 score of the belief-based strategy are 99.1165 and 99.5563, respectively. This demonstrates that the belief-based strategy provides an extremely high matching quality.

Note that Google has also constructed a mapping between Freebase and Wikidata that was published in October 2013. They detected 2,099,582 entity pairs with 2,096,745 Freebase entities and 2,099,582 Wikidata entities. Fig. 4 illustrates the result of identical entity pairs using the same datasets from Freebase and Wikidata. The entity pairs from all proposed strategies have some differences compared to the Google result. Although they did not explicitly

Table 3. Composition of mapping results based on different strategies

	Consistency	Max Confidence	Threshold Filtering	One-to-one mapping	Belief-based
1 Freebase and 1 Wikidata	4,384,747	4,390,685	4,390,423	4,390,423	4,352,022
1 Freebase and multiple Wikidata	14,586	4,400	4,814	0	4,704
Multiple Freebase and 1 Wikidata	957	143	262	6	632
Multiple Freebase and multiple Wikidata	665	30	43	0	111
Total	4,400,955	4,395,258	4,395,542	4,390,429	4,357,469

Table 4. Matching quality of proposed strategies

	Consistency	Max Confidence	Threshold Filtering	One-to-one mapping	Belief-based
Recall (%)	100	100	100	100	100
Precision (%)	98.1371	98.2643	98.2580	98.3724	99.1165
F1 score (%)	99.0598	99.1246	99.1213	99.1795	99.5563

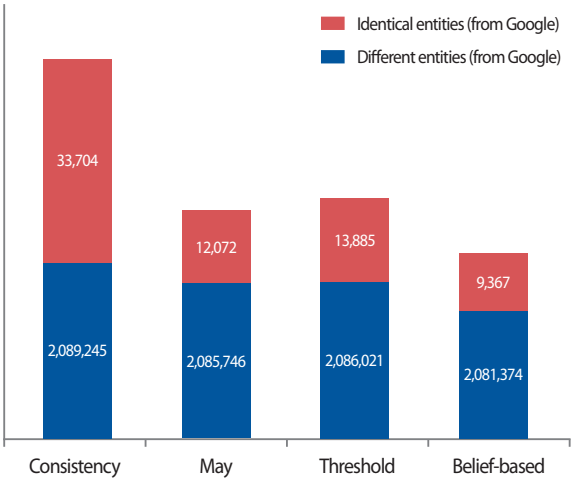

Fig. 4. A comparison of the Google result.

announce how they extracted this result, it might use an exact matching of Wikipedia URL. Applying the proposed strategies to the Google results, the identical mapping pairs are more than 99.51%. However, they include ambiguous results according to individual strategies. For example, the consistency strategy has the highest different entities (1.59%), whereas the belief-based strategy is the smallest (0.45%). In summary, the belief-based strategy can be considered as an effective approach to reduce ambiguity for entity extraction. Note that matching performance of the Google result is not conducted, because they provided this dataset only once, and did not update related data sources.

7. CONCLUSIONS

This study proposed several approaches for identifying the same entities from heterogeneous knowledge sources and evaluated these approaches by using Wikidata and Freebase. According to the evaluation results, the belief-based approach is most effective for reducing the ambiguous relations between the given datasets. Although the consistency strategy returned the largest number of pairs of the same relation, it also had the highest number of errors. Entity resolution is a popular topic in industry and academia. Currently, common and popular approaches for entity resolution focus on similarity-join techniques, but few studies have focused on belief-based approaches. The proposed belief-based same extraction approach can be a new technique for measuring the matching degree of entity pairs.

Although this paper conducted an entity extraction using large-scale real-world datasets, there are more experiments for integrating heterogeneous knowledge sources. Future work may explore the alternative expanding algorithms for handling different property values and evaluating the impact of optimised approaches. Another potential area of research is to integrate heterogeneous knowledge into existing knowledge sources by instance matching techniques.

REFERENCES

Berners-Lee, T. (2009). *The semantic web: linked data*. Retrieved Jun 10, 2018 from https://www.w3.org/DesignIssues/LinkedData.html.

Bizer, C., Cyganiak, R., & Heath, T. (2007). *How to publish linked data on the web*. Retrieved Jun 10, 2018 from http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/LinkedDataTutorial/.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: A collaboratively created graph database for structuring human knowledge. In *SIGMOD '08: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data* (pp. 1247-1250). New York, NY: ACM.

Castano, S., Ferrara, A., Montanelli, S., & Lorusso, D. (2008). Instance matching for ontology population. In S. Gaglio, I. Infantino, & D. Saccà (Eds.), *Proceedings of the Sixteenth Italian Symposium on Advanced Database Systems* (pp. 121-132). Mondello, Italy: SEBD.

Ding, L., Shinavier, J., Shangguan, Z., & McGuinness, D. L. (2010). SameAs networks and beyond: Analyzing deployment status and implications of owl: sameAs in linked data. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan,…B. Glimm (Eds.), *International Semantic Web Conference* (pp. 145-160). Berlin: Springer.

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K.,…Zhang, W. (2014). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 601-610).

Enríquez, J. G., Mayo, F. J. D., Cuaresma, M. J. E., Ross, M., & Staples, G. (2017). Entity reconciliation in big data sources: A systematic mapping study. *Expert Systems with Applications*, 80, 14-27.

Färber, M., Ell, B., Menne, C., & Rettinger, A. (2015). A comparative survey of DBpedia, Freebase, OpenCyc,

Wikidata, and YAGO. *Semantic Web Journal*, 1, 1-5.

Gabrilovich, E., & Usunier, N. (2016). Constructing and mining web-scale knowledge graphs. In R. Perego, F. Sebastiani, J. A. Aslam, I. Ruthven, & J. Zobel (Eds.), *SIGIR '16 Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 1195-1197). New York, NY: ACM.

Gottron, T., & Staab, S. (2014). Linked open data. In *Encyclopedia of social network analysis and mining* (pp. 811-813). New York, NY: Springer.

Halpin, H., Hayes, P., McCusker, J. P., McGuinness, D., & Thompson, H. S. (2010). When owl:sameAs isn't the same: An analysis of identity in linked data. In *Proceedings of the 9th International Semantic Web Conference (ISWC)* (pp. 53-59). Berlin, Heidelberg: IOS Press.

Hogan, A., Zimmermann, A., Umbrich, J., Polleres, A., & Decker, S. (2012). Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora. *Journal of Web Semantics*, 10, 76-110.

Hors, A. L., & Speicher, S. (2014). Using read-write linked data for application integration. In A. Harth, K. Hose, & R. Schenkel (Eds.), *Linked data management* (pp. 459-483). Lyon, France: Chapman and Hall/CRC.

Idrissou, A. K., Hoekstra, R., van Harmelen, F., Khalili, A., & den Besselaar, P. V. (2017). Is my sameAs the same as your sameAs? Lenticular lenses for context-specific identity. In Ó. Corcho, K. Janowicz, G. Rizzo, I. Tiddi, & D. Garijo (Eds.), *K-CAP* (pp. 23:1-23:8). New York, NY: ACM.

Kim, H., Liang, H., & Ying, D. (2014). Knowledge extraction framework for building a largescale knowledge base. *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, 16(7), 1-8.

Lehmann, J., Bizer, C., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia: A crystallization point for the Web of Data. *Journal of Web Semantics*, 7, 154-165.

Moaawad, M. R., Mokhtar, H. M. O., & al Feel, H. T. (2017). On-the-fly academic linked data integration. In *ICCDA '17 Proceedings of the International Conference on Compute and Data Analysis* (pp. 114-122). New York, NY: ACM.

Nguyen, K., & Ichise, R. (2016). Linked data entity resolution system enhanced by configuration learning algorithm. *IEICE Transactions*, 99-D, 1521-1530.

Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 8, 489-508.

Stefanidis, K., Efthymiou, V., Herschel, M., & Christophides, V. (2014). Entity resolution in the web of data. In *Proceedings of the 23rd International Conference on World Wide Web* (WWW '14 Companion) (pp. 203-204). New York, NY: ACM.

Suchanek, F., Kasneci, G., & Weikum, G. (2007). YAGO-A core of semantic knowledge. In *Proceedings of International Conference on World Wide Web* (pp. 697-706). New York, NY: ACM.

Tanon, T. P., Vrandecic, D., Schaffert, S., Steiner, T., & Pintscher, L. (2016). From Freebase to Wikidata: The great migration. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)* (pp. 1419-1428). Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Vrandecic, D. (2012). Wikidata: A new platform for collaborative data collection. In A. Mille, F. L. Gandon, J. Misselis, M. Rabinovich, & S. Staab (Eds.), *WWW (Companion Volume)* (pp. 1063-1064). New York, NY: ACM.

Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29, 2724-2743.

Yager, R. R. (1987). On the Dempster-Shafer framework and new combination rules. *Information Sciences*, 41(2), 93-137.