Knowl Inf Syst (2019) 59:465–495
https://doi.org/10.1007/s10115-018-1192-z

REGULAR PAPER

A new truth discovery method for resolving object conflicts over Linked Data with scale-free property

Wenqiang Liu[1] · Jun Liu[1] · Bifan Wei[1] · Haimeng Duan[1] · Wei Hu[2]

Received: 10 October 2017 / Revised: 5 January 2018 / Accepted: 18 April 2018 /
Published online: 3 May 2018

Abstract Considerable effort has been exerted to increase the scale of Linked Data. However, an inevitable problem arises when dealing with data integration from multiple sources. Various sources often provide conflicting objects for a certain predicate of the same real-world entity, thereby causing the so-called *object conflict* problem. Existing truth discovery methods cannot be trivially extended to resolve object conflict problems because Linked Data has a scale-free property, i.e., most of the sources provide few objects, whereas only a few sources have numerous objects. In this study, we propose a novel approach called TruthDiscover to determine the most trustworthy object in Linked Data with a scale-free property. More specifically, TruthDiscover consists of two core components: Priori Belief Estimation for smoothing the trustworthiness of sources by leveraging the topological properties of the Source Belief Graph, and Truth Computation for inferencing the trustworthiness of source and trust value of an object. Experimental results conducted on six datasets show that TruthDiscover achieves higher accuracy than existing approaches, and it is robust and consistent in various domains.

The current work is the extension and continuation of our previous work that has been published in a conference paper of ESWC 2017 [23].

✉ Jun Liu
liukeen@xjtu.edu.cn

Wenqiang Liu
liuwenqiangcs@gmail.com

Bifan Wei
weibifan@xjtu.edu.cn

Haimeng Duan
duanhaimeng@gmail.com

Wei Hu
whu@nju.edu.cn

1 MOEKLINNS Lab, Xi'an Jiaotong University, Xi'an, China

2 State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

Keywords Linked Data · Linked Data quality · Truth discovery

1 Introduction

Considerable effort has been made to increase the scale of Linked Data. In particular, the number of available Linked Data sources in the Linking Open Data (LOD) project increased from 12 in 2007 to 1,146 in 2017.[1] In this paper, a Linked Data source refers to a dataset that has been published in the LOD project by individuals or organizations, such as YAGO. Linked Data resources are encoded in the form of ⟨*Subject*, *Predicate*, *Object*⟩ triples through the Resource Description Framework (RDF) format. The subject denotes the resource, and predicate is used to express a relationship between subject and object. Inevitably, errors occur during such creation process, given that many Linked Data sources on the web have been created from semi-structured datasets (e.g., Wikipedia) and unstructured ones through automatic or semiautomatic algorithms [8]. As a result, a predicate for the same real-world entity can have multiple inconsistent objects when dealing with data integration from multiple sources. For example, the objects of the *dbp:populationTotal* for *Beijing* in Freebase[2] and DBpedia[3] are "20,180,000" and "21,516,000," respectively. In this paper, this problem is called the *object conflict* problem. The concept of object conflicts can be defined as two objects being in conflict only when their similarity is less than the defined threshold. According to this definition, it is also likely to regard two objects expressed in terms of different measure units as conflicts. But, the purpose of our study is to rank the trust values of all objects and provide the most common ones for users, rather than remove some objects directly. Therefore, people who use our methods can still see all objects.

1.1 Problems of object conflicts in Linked Data

We constructed six massive real-world Linked Data datasets in this study to understand object conflicts of Linked Data. These datasets comprise six domains: *persons*, *locations*, *organizations*, *descriptors*, *films* and *music*. The first four datasets are constructed based on the OAEI2011 New York Times dataset,[4] which is a well-known and carefully created dataset of Linked Data. Two other domains, including *films* and *music*, are constructed through SPARQL queries over DBpedia to draw more robust conclusions. The detailed construction process and statistics of these datasets are described in Sect. 5.2. Through a detailed quantitative analysis of these datasets, we first answer the following questions. Are object conflicts a common problem for the Linked Data community? What are the causes of object conflicts in Linked Data?

The answers obtained by observing are quite surprising. Approximately 45% of predicates have multiple inconsistent objects, and the average number of objects is 11 for a certain predicate (described in Sect. 5.2). To understand the degree of inconsistency of Linked Data, normalized entropy [32] is selected to examine the inconsistency of different objects. Generally, the higher normalized entropy is, the higher the degree of inconsistency is. Let $O = \{o_i\}_m$ denote a set of different objects for a certain predicate of a real-world entity, and $P(o_i)$ represent percentage of occurrences of o_i. The corresponding normalized entropy can be defined as follows:

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

[2] https://www.freebase.com/m/072p8.

[3] http://dbpedia.org/resource/Beijing.

[4] https://drive.google.com/open?id=1bOYI7LXTkQqTpPUb7N7RNiUjiyQn0iKC.

$$E = -\sum_{i=1}^{m} P(o_i) \log P(o_i) / \log m. \quad (1)$$

Our observations of the six datasets indicate that the average normalized entropy is 0.75. Approximately 80% of predicates have normalized entropy of more than 0.8. This result indicates that the object conflicts are a common issue in the community of Linked Data. In addition, object conflicts in Linked Data are caused by four distinct reasons. The first reason for the inconsistency is *multi-values* (32%), in which the predicate inherently has multiple objects (e.g., the predicate http://dbpedia.org/ontology/location). The second reason is attributed to being *out-of-date* (13%); the corresponding object tends to change over time because the predicate is time sensitive (e.g., http://dbpedia.org/ontology/populationTotal). The third reason for the inconsistency is *variety* (43%), which refers to different objects that may be presented in different ways or different data precision. The fourth reason for the inconsistency is *pure errors* (12%). In this study, we focus on resolving three reasons (68%) including *out-of-date*, *variety* and *pure errors*, which only have one truth for a certain predicate of a real-world entity.

An effective method that can automatically distinguish between true and false object is extremely beneficial for the Linked Data community.

1.2 Limitations of existing object conflicts resolving methods

Authoritative-source-based methods One straightforward approach for this task is regarding the object from well-known authoritative sources as the true value. However, objects from different well-known sources for the same predicate are not always consistent. For example, *Freebase* and *DBpedia* provide different objects for the *dbp:populationTotal* of *Beijing*. Therefore, selecting one of these well-known sources as a trustworthy source is difficult when confronted with the problem of object conflicts.

Majority-voting methods Majority voting is another simple method to resolve object conflicts, wherein the object with the maximum number of occurrences is regarded as the truth [18]. However, we find that these methods achieve relatively low precision (ranging from 0.3 to 0.45) on the six datasets. There are two reasons why majority voting perform poorly in Linked Data. Firstly, approximately 50% of predicates do not have a dominant object. In this case, majority voting can only randomly select one object in order to break the tie. We also examine the correlation between the dominance factor and precision to reveal the inherent reasons of the poor performance of majority voting in Linked Data, as shown in Fig. 1. The dominance factor DF of a certain predicate is defined as:

$$DF = \frac{\max_{o_i \in O} oc(o_i)}{\sum_{o_i \in O} oc(o_i)}, \quad (2)$$

where O represents a set of different objects for a certain predicate of a real-world entity, and $oc(o_i)$ is the number of occurrences o_i.

Majority voting can only achieve satisfactory precision when the dominance factor is of more than 0.7. However, this requirement is extremely stringent to achieve in Linked Data.

Secondly, majority voting assumes that all sources are equally reliable and indistinguishable. A recent research [38] has indicated that different Linked Data sources have different qualities. Therefore, majority-voting methods are not applicable in Linked Data.

Truth discovery methods Numerous truth discovery methods [7,17–19,21,33,39,40], which found the truth by simultaneously estimating source reliability and trust values of objects, have been proposed to address the limitation of majority voting. In these methods,

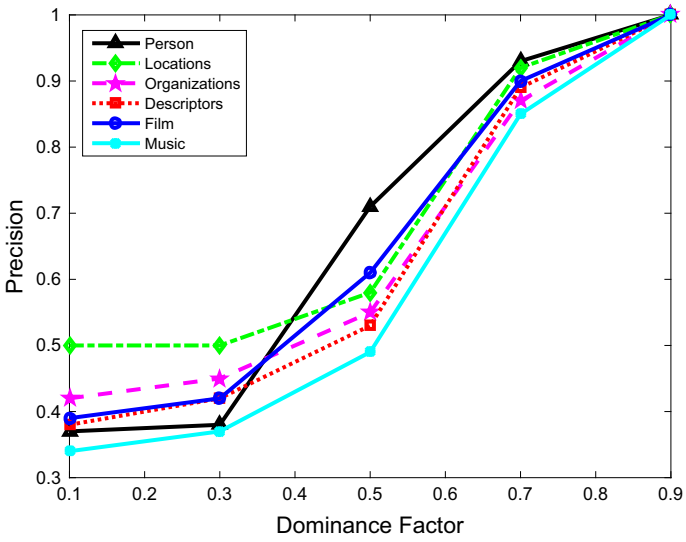


Fig. 1 Correlation between dominance factor and precision

the truth for an entity refers to the object that is assigned a maximum trust value among all different objects. A source that regularly provides trustworthy objects more often is more reliable, and an object from a reliable source is more trustworthy in truth discovery methods. The trustworthiness of each source can be simulated as the percentage of true objects provided by this source. Consequently, the more true objects a source provides, the more likely that the trustworthiness of the source is closer to their real degree. However, evaluating the reliability degrees of a few "small" sources that provide significantly few objects is difficult. Considering an extreme case when a source only provides one object, its trustworthiness is one if the object is correct, and the source is regarded as highly reliable. Otherwise, the source is considered as highly unreliable. Inaccurate estimation of source reliability inevitably has negative effects on identifying trustworthy objects. Therefore, the effectiveness of numerous truth discovery methods is significantly affected by the number of objects provided by each source.

The number of objects provided by each source in this study typically follows the approximate power law in Linked Data. This finding indicates that Linked Data has a scale-free property. This property is characterized by $p(k)$, which is the fraction of the sources possessing k objects, following the power law $p(k) \backsim k^{-\gamma}$, where γ is the exponent of the power law and ranges from 2.12 to 3.1 on the six datasets as shown in Fig. 2. In the six plots, the X- and Y-coordinates represent the number of objects provided by a source and the complementary cumulative distribution function $Pr(k) = \sum_{x=k}^{+\infty} p(x)$, respectively. Figure 2 shows that the number of objects ranges from 1 to 10 for most of the sources, and only a few sources have numerous objects. In the preceding discussion, numerous truth discovery methods are sensitive to the number of objects provided by each source. Therefore, these methods cannot be trivially extended to resolve conflicts in Linked Data with a scale-free property.

1.3 Overview of our approach

A simple method to solve the issues attributed to the scale-free property is by removing "small" sources [17]. However, the removal of "small" sources results in limited coverage

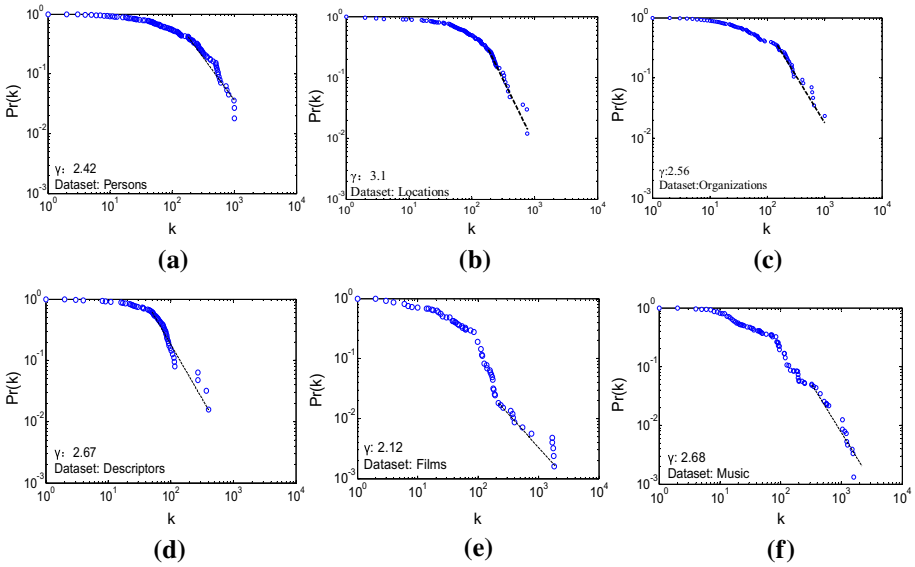


Fig. 2 Cumulative distributions of objects on the six datasets

and sparse data because most Linked Data sources are "small." A novel approach named TruthDiscover is proposed in this study to resolve object conflicts in Linked Data with a scale-free property. TruthDiscover involves the following steps.

(i) Estimating priori beliefs The non-uniform priori beliefs of all sources are computed by leveraging the topological properties of the Source Belief Graph which represents the relationship between sources.

(ii) Truth computation This component consists of two parts: (1) **Computing the Trustworthiness of Sources**. The trustworthiness of each source is automatically computed based on the trust scores of objects. Thereafter, the priori beliefs of sources are added to smooth the trustworthiness of sources using the averaging strategy. (2) **Computing the Trust Values of Objects**. The trust values of objects are computed based on pairwise Markov Random Field (pMRF) model. If the changes in all objects after each iteration are less than the threshold, then the object with maximum trust score is regarded as truth; otherwise, return to part (1).

1.4 Contributions and organization

We focus on resolving three reasons (68%) in this study including *out-of-date*, *variety* and *pure errors*. These reasons only have one truth for a certain predicate of a real-world entity. We defer handling multi-values to future work.

The main contributions of this study are as follows:

(i) The number of objects provided by multiple Linked Data sources typically follows the approximate power law. This finding indicates that only a few sources have numerous objects, whereas most of the sources provide few objects. We identify the challenges brought by the scale-free property on the task of truth discovery.

(ii) A truth discovery approach called TruthDiscover is proposed to identify the truth in Linked Data with a scale-free property. Two strategies are adopted in TruthDiscover to address the challenges attributed to the scale-free property. First, this approach leverages

the topological properties of the Source Belief Graph to estimate the priori beliefs of sources for smoothing the trustworthiness of sources. Second, we firstly formalize the object conflict-resolution problem as computing the joint distribution of variables in a heterogeneous information network called the *SourcespsObjectNetwork*, which successfully captures three correlations from objects and Linked Data sources. Then a method based on pMRF is proposed to model the interdependencies from objects and sources, and a message propagation-based method is utilized that exploits the Source-Object Network structure to infer the trust values of all objects.

(iii) We conducted extensive experiments on six real-world Linked Data datasets from the *persons*, *locations*, *organizations*, *descriptors*, *films* and *music* domains, to validate the effectiveness of our approach. Our experimental results showed that our method achieves higher precision than all baselines.

This work is the extension and continuation of our previous work [23]. The approach we update in this paper differs from our previous work in four aspects. First, we are the first to quantify and study inconsistency of Linked Data in this paper. Four reasons for inconsistency are concluded through our observations from six datasets. Second, we also find that the number of objects provided by multiple Linked Data sources typically follows the approximate power law and identify the challenges in Linked Data with power law phenomenon for the task of truth discovery. Third, a novel approach which leverages the topological properties of the Source Belief Graph is proposed to estimate the priori beliefs of sources for smoothing the trustworthiness of sources. Finally, three new experiments have been added to draw more robust conclusions including effectiveness evaluation with regard to the three reasons for inconsistency, convergence analysis and time efficiency evaluation.

The remainder of this paper is organized as follows: Related work is discussed in Sect. 2. Section 3 presents the formulation of problem, and the detail of our method is provided in Sect. 4. Section 5 shows comparative evaluation between the proposed and other approaches. Finally, Sect. 6 concludes the paper.

2 Related work

Resolving conflict from multiple sources has long been investigated [1]. Existing methods can be grouped into two categories, namely relational databases and Linked Data, depending on different data models.

2.1 Conflict in relational databases

Relational databases have the formal structure of data models. Resolving conflicts in relational databases refer to determining contradictory attribute values from different sources when integrating data [1]. This problem was first mentioned by Dayal et al. [3] in 1983. However, the problem did not receive much attention at that time because numerous of the applications adopted conflict-avoiding or conflict-ignoring strategies [1]. Subsequently, numerous methods inspired by measuring web page authority, such as Authority-Hub analysis [15], were proposed. However, authority does not indicate high precision [37]. Recently, the methods based on truth discovery have gained increasing attention because of its ability to estimate degrees of source reliability and infer trust values of objects simultaneously. These methods can be divided into three groups [20], namely iterative methods, optimization-based methods, and probabilistic-graphical-model-based methods.

Iterative methods These methods usually employ the interdependencies between the trust value of objects and the trustworthiness of sources to find true objects. The research of Yin [37] played an important role in this subfield. Yin's method utilized Bayesian analysis and the relationship between the trustworthiness of sources and the probability of each claim being true to identify truth. Since then, several methods have proposed henceforth specific scenarios based on the seminal work of Yin. For example, Dong et al. [6] proposed an iterative method by analyzing the dependency between source reliability and trust values of objects, which is different from the work of Yin, because it considers dependence between data sources.

Optimization-based methods These methods find the truth by minimizing the gap between the information provided by sources and the identified truth. For example, Li et al. [18] proposed an optimization framework among multiple sources of heterogeneous data types, wherein the trust value of objects and the trustworthiness of sources are defined as two sets of unknown variables. The truth was discovered by minimizing the optimization function.

Probabilistic-graphical-model-based methods These methods can automatically infer truth and the degree of source reliability by a probabilistic graphical model. For instance, Zhao et al. [39] developed a probabilistic graphical model to address the truth finding problem through modeling the two aspects of source reliability, namely sensitivity and specificity. This model is also the first to address the problem of multi-valued attribute types.

2.2 Conflict in Linked Data

Conflict in Linked Data can be classified into three categories, namely *identity*, *schema*, and *object conflicts* [27]. Accordingly, existing methods to resolve conflicts in Linked Data can be grouped into three groups.

Identity conflicts Identity conflicts is when different subjects from various sources denote the same real-world entity (e.g., *dbpedia:Statue of Liberty* and *freebase:m.072p8*; we use prefixes in this paper, instead of full URIs, to save space). Resolving identity conflicts is also known as entity resolution or object co-reference resolution. Two types of methods are generally adopted to resolve identity conflicts. The first methods are based on Web Ontology Language (OWL) semantics inference. For instance, Glaser et al. [9] implemented a co-reference resolution service based on *owl:sameAs*. The other methods are based on the assumption that two subjects denote the same real-world entity if they share several common property-value pairs. For instance, Wang et al. [35] proposed a concept mapping method based on the similarities between concept instances.

Schema conflicts Schema conflicts indicate that different schemata are utilized to describe the same predicate (e.g., *rdfs:label* and *skos:prefLabel*). Numerous methods have been introduced to solve schema conflicts through schema mapping. These methods are further divided into two categories, namely linguistic matching-based and structural matching-based. Linguistic matching-based methods usually employ string similarity computation based on names, labels, and several other descriptions. For instance, Qu et al. [30] presented a method to resolve schema conflicts by computing the similarity between documents of a domain entity (e.g., a class or a property). Structural matching-based methods usually employ graphs to represent different schemata and determine the structural similarity between graphs. For example, Hu et al. [13] proposed a method based on RDF bipartite graphs to resolve schema conflicts by computing structural similarities between domain entities and statements using a propagation procedure over the bipartite graphs.

Object conflicts Object conflicts occur when multiple inconsistent objects exist for a certain predicate of the same real-world entity. Resolving object conflicts is a key step for

Linked Data integration and consumption. However, to the best of our knowledge, research on resolving object conflicts has not elicited enough attention in the Linked Data community. According to our survey, existing methods to resolve object conflicts in Linked Data can be grouped into three major categories: conflict ignoring, conflict avoidance and conflict resolution.

The conflict-ignoring methods ignore the object conflicts and defer conflict resolution to users. For instance, Wang et al. [34] presented an effective framework to fuse knowledge cards from various search engines. In this framework, the fusion task involves card disambiguation and property alignment. For the value conflicts, this framework only adopts deduplication of the values and groups these values into clusters.

The conflict-avoidance methods acknowledge the existence of object conflicts, but does not resolve these conflicts. Alternatively, they apply a unique decision to all data, such as manual rules. For instance, Mendes et al. [26] presented a Linked Data quality assessment framework called Sieve. In this framework, the strategy "Trust Your Friends," which prefers the data from specific sources, was adopted to avoid conflicts.

The conflict-resolution methods focus on how to solve a conflict regarding the characteristics of all data and metadata. A straightforward method is to resolve object conflicts by conducting the majority voting as shown in [28]. The drawback of majority voting is that it assumes that all Linked Data sources are equally reliable as discussed in Sect. 1.2. In order to consider the quality of data source, some methods based on truth discovery techniques have been proposed. For example, Michelfeit et al. [27] presented an assessment model that leverages the quality of the source, data conflicts, and confirmation of values for determining which value should be the true value. Liu et al. [23] proposed a truth discovery approach, ObResolution, which utilizes the Source-Object network to infer the true object. This network successfully captures three correlations from objects and Linked Data sources. One shortcoming of these approaches is that the effectiveness of these methods is significantly affected by the number of objects provided by each source as discussed in Sect. 1.2.

Previous work enlightens us on resolving object conflicts. Our approach is different from the approaches mentioned above in two aspects. First, we find that the number of objects provided by multiple Linked Data sources typically follows the approximate power law and identifies the challenges in Linked Data with power law phenomenon for the task of truth discovery. A novel approach which leverages the topological properties of the Source Belief Graph is proposed to estimate the priori beliefs of sources. Second, we formalize the problem of object conflict resolution through a heterogeneous information network, which successfully captures all the correlations from objects and Linked Data sources.

3 Preliminaries

3.1 Basic definitions

Several important notations utilized in this study are introduced in this subsection. Thereafter, the problem is formally defined.

Definition 1 (*RDF Triple*) [24] We let I denote the set of IRIs (Internationalized Resource Identifier), B are the set of blank nodes, and L are the set of literals (denoted by quoted strings, e.g., "*Beijing City*"). An RDF triple can be represented by $\langle s, p, o \rangle$, where $s \in I \cup B$ is a *subject*, $p \in I$ is a *predicate*, and $o \in I \cup B \cup L$ is an *object*.

Definition 2 (*SameAs Triple*) A SameAs triple can be represented by $\langle s, \text{owl:sameAs}, o\rangle$, which connects two RDF resources through the *owl:sameAs* predicate.

Definition 3 (*SameAs Graph*) Given a set of sameAs triples T, a SameAs Graph SG can be represented by (V, E), where $V = \{s|\langle s, \text{owl:sameAs}, o\rangle \in T\} \cup \{o|\langle s, \text{owl:sameAs}, o\rangle \in T\}$ is a set of vertices (i.e., subjects and objects), $E \subseteq V \times V$ is a set of directed edges with each edge corresponding to a sameAs triple in T.

Definition 4 (*Source Belief Graph*) Given a SameAs Graph SG, the Source Belief Graph can be denoted by $SBG = (\mathcal{W}, R)$, where $\mathcal{W}$ is a set of vertices with each vertex corresponding to the source name of the vertex in SameAs Graph SG; R is a multiset of $\mathcal{W} \times \mathcal{W}$ formed by pairs of vertices (μ, ν), $\mu, \nu \in \mathcal{W}$ and each pair (μ, ν) corresponds to an edge in SameAs Graph SG.

Definition 5 (*Trustworthiness of sources*) [37] The trustworthiness of a source ω_j is the confidence of the objects provided by ω_j, which is denoted by $t(\omega_j)$.

Definition 6 (*Trust values of objects*) [37] The trust value of an object o_i is the probability of being correct, which is denoted by $\tau(o_i)$.

Therefore, the process of resolving object conflicts in Linked Data is formally defined as follows: given a set of different objects O, TruthDiscover will produce one truth for a certain predicate of a real-world entity. The truth is represented by $o^* = \arg\max_{o_i \in O} \tau(o_i)$.

3.2 Problem analysis

Through the observation and analysis of the object conflicts in our sample Linked Data, we found three helpful correlations from Linked Data sources and objects to effectively distinguish between true and false objects.

Correlations among Linked Data sources and objects If an object comes from a reliable source, it will be assigned a high trust value. Thus, a source that provides trustworthy objects often has big chance to be selected as a reliable source. For example, the object provided by DBpedia is more reliable than objects supported by many small sources because DBpedia is created from Wikipedia. This condition also serves as a basic principle for many truth discovery methods [17–19,21,33,39].

Correlations among objects If two objects are similar, they should have similar trust values, which indicates that similar objects appear to have mutually support. For example, we assume that one source claims that the *dbp:height* of *Statue of Liberty* is "46.0248" and another says that it is "46.2". If one of these sources has a high trust value, the other should have a high trust value as well. Meanwhile, if two objects are mutually excluded, they cannot be both true. If one of them has a high trust value, the other should have a low trust value. For instance, if two different sources claim that the *dbp:height* of *Statue of Liberty* are "80" and "46," respectively. If the true object is "46", then "80" should be a wrong object. Two similarity functions are adopted to determine the similarity $S(o_i, o_k)$ between objects o_i and o_k to validate the second assumption in this study.

The most commonly used similarity function for numerical data is defined as:

$$S(o_i, o_k) = \frac{1}{1 + d(o_i, o_k)}, \tag{3}$$

$$d(o_i, o_k) = \begin{cases} 1 & \text{if } o_i = o_k = 0, \\ \frac{|o_i - o_k|}{max(|o_i|, |o_k|)} & \text{others.} \end{cases} \tag{4}$$

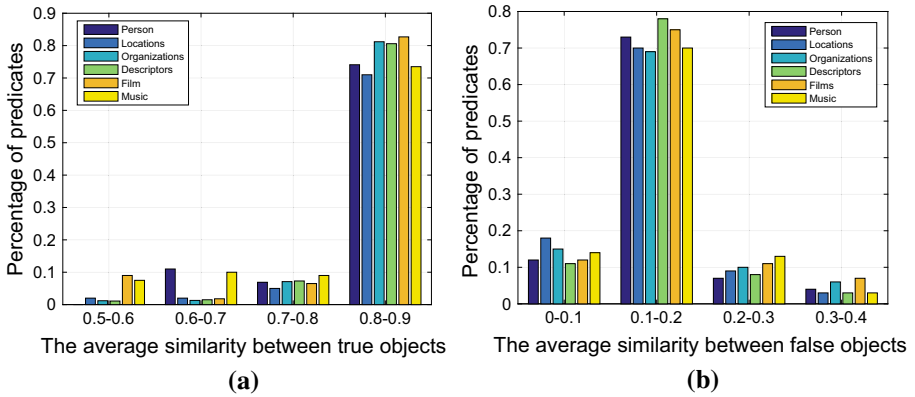


Fig. 3 Distributions of average similarities between objects on the six datasets

The Levenshtein distance is adopted to describe the similarities of objects for string data. The similarity function is defined as follows:

$$S(o_i, o_k) = 1 - \frac{ld(o_i, o_k)}{max(len(o_i), len(o_k))}, \tag{5}$$

where $ld(o_i, o_k)$ denotes the Levenshtein distance between objects o_i and o_k; $len(o_i)$ and $len(o_k)$ are the length of o_i and o_k, respectively.

The distribution of average similarities between true objects on the six datasets is shown in Fig. 3a. Approximately 90% of predicates have average similarities of more than 0.8. Figure 3b shows the average similarities between false objects on the six datasets. The average similarities range from 0 to 0.4, and approximately 80% of predicates whose average similarities are less than 0.2. This finding indicates that the truths provided by different sources appear to be similar, and false objects are generally less consistent than true objects.

Correlations among Linked Data sources In many truth discovery methods, the trustworthiness of a source is formulated as the probability of the objects provided by this source being the truth. Therefore, the more same objects two different sources provide, the more similar is the trustworthiness of the two sources. Consider an extreme case when two sources provide the same objects for each predicate, and the trustworthiness of these two sources is the same.

As discussed, these three principles can be used to infer the trust values of objects. A key problem for object conflicts resolution is how to model these principles under a unified framework.

4 TruthDiscover method

Based on these analyses, we propose a method called TruthDiscover to resolve object conflicts in Linked Data with a scale-free property. Given a set of different objects $O = \{o_i\}_m$, Fig. 4 illustrates the framework of generating truth o^* by TruthDiscover, which mainly includes the following three modules.

(1) Module I. Priori belief estimation (described in Sect. 4.1). This module produces priori belief for each source by leveraging the topological properties of the Source Belief Graph.

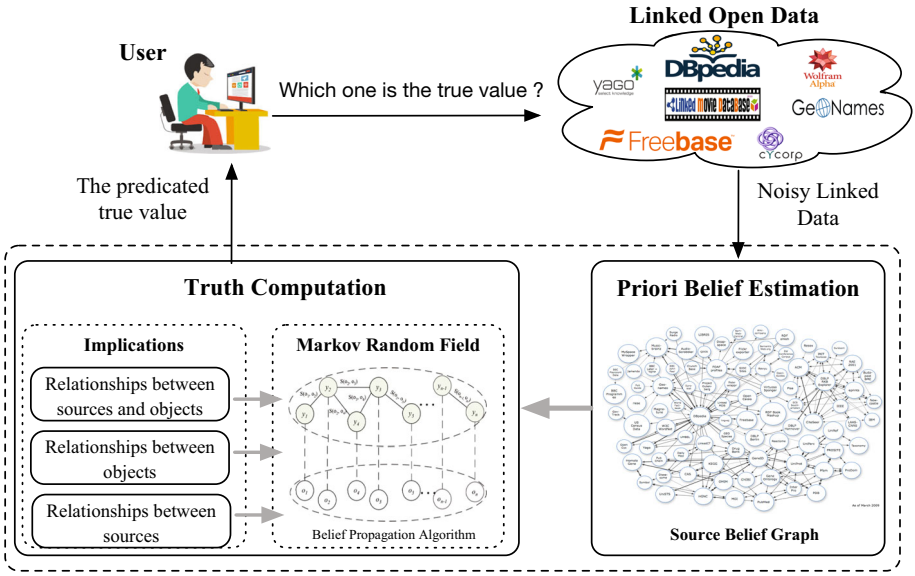


Fig. 4 Visualization of the whole method

(2) Module II. Truth Computation (presented in Sect. 4.2). First, this module computes the trustworthiness of each source based on the trust scores of objects. Thereafter, the priori beliefs of sources are added to smooth the trustworthiness of sources by using an averaging strategy. Finally, the MRF is adopted to model the relationships between objects for accurately computing trust values of objects. In this study, the Loopy Belief Propagation (LBP) is applied to estimate the marginal probabilities of each unobserved variable in MRF. The object with the maximum trust score is regarded as the truth when the changes in all objects after each iteration are less than the preset threshold.

The pseudo-code of this method is shown in Algorithm 1.

4.1 Priori belief estimation

This section describes a method called BeliefRank to estimate the priori beliefs of all sources by leveraging the topological properties of the Source Belief Graph.

The *owl:sameAs* property in OWL [25] indicates that two subjects actually refer to the same thing. Utilizing this property further enriches the space of Linked Data by declaratively interconnecting "equivalent" objects across distributed Linked Data sources [4]. The *owl:sameAs* property have been extensively utilized in many Linked Data sources recently, such as DBpedia, Freebase, YAGO, and GeoNames.[5] Figure 5 shows a fragment of *owl:sameAs* triples in *dbpedia:Beijing*. [6] Numerous *owl:sameAs* triples are obtained together and form a directed graph called SameAs Graph [5], as defined in Definition 3. Many researchers investigated *owl:sameAs* triples and sameAs graph [5,10] because of the importance of *owl:sameAs* in the integration of Linked Data. However, estimating the reliability degree of Linked Data sources through SameAs Graph analysis has not been attempted to the best of our knowledge.

[5] www.geonames.org/.

[6] http://dbpedia.org/data/Beijing.

Algorithm 1 TruthDiscover

Input: a set of conflicting objects $O = \{o_1, \ldots, o_m\}$, a set of Linked Data sources $\Omega = \{\omega_1, \ldots, \omega_n\}$ and the mapping relations between O and Ω
Output: trust value $\tau(o_i)$, $o_i \in O$; trustworthiness of source $t(\omega_j)$, $\omega_j \in \Omega$
// The purpose of 1∼ 2 is to generate the priori beliefs of sources
1: Priori belief estimation:
$\forall \omega_j \in \Omega$, compute $BR(\omega_j)$ through BeliefRank (described in Sect. 4.1) ;
2: Initialize the trustworthiness of sources by the normalized priori beliefs:
$\forall \omega_j \in \Omega, t(\omega_j) = NBR(\omega_j)$(described in Sect. 4.2.1);
3: **for** $o_i \in O$ **do**
4: Compute trust value $\tau(o_i)$ with Eq. 10;
5: **end for**
6: $\forall o_i, o_k \in O$: Calculating their similarity $S(o_i, o_k)$ (Described in Sect. 3.2);
//Initialize the message
7: $\forall o_i, o_j \in O$: $m_{i\to j}(y_i) = 1$;
//Message propagation
8: **repeat**
9: **for** $j \leftarrow 1$ to $m + n$ **do**
10: **for** $i \leftarrow 1$ to $m + n$ **do**
11: $m_{i\to j}(y_j) = \sum_{y_i \in \{0,1\}} U(y_i, y_j)\psi_i(y_i) \prod_{y_k \in N(y_i)\cap Y \setminus \{y_j\}} m_{k\to i}(y_i)$.
12: **end for**
13: **end for**
14: **until** the convergence criterion is satisfied;
//Belief assignment
15: **for** $i \leftarrow 1$ to $m + n$ **do**
16: $P(y_i) = \psi_i(y_i) \prod_{y_j \in N(y_i)\cap Y} m_{j\to i}(y_i)$.
17: **end for**
18: **return** $\tau(o_i), \forall o_i \in O; t(\omega_j), \omega_j \in \Omega$.

dbpedia:Beijing
owl:sameAs: geo:1816670;
owl:sameAs: opencys:Mx4rvVjVPZwpEbGdrcN5Y29ycA;
owl:sameAs: freebase:m.01914;
owl:sameAs: geovocab:3_20168;
owl:sameAs: linkedgeodata:node25248662;
owl:sameAs: yago:Beijing.

Fig. 5 Fragment of *owl:sameAs* triples in *dbpedia:Beijing*

Data publishers usually add new *owl:sameAs* triples that points to the external equivalent subject when they publish their data as Linked Data on the web. Data publishers logically select a subject provided by the source they trust. That is, the *owl:sameAs* property indicates that the data publishers place their trust to the subject provided by a source they trust. Typically, the data publisher of a subject can be represented by the name of the source [5]. For example, "DBpedia" is an abstract representation of the data publisher for *dbpedia:Beijing*. That is, the SameAs Graph can be converted to a directed multigraph called the Source Belief Graph, which represents the relationship between sources. Formally, the Source Belief Graph SBG is a pair of sets $(\mathcal{W}, R)$, where $\mathcal{W}$ is a set of vertices with each vertex corresponding to the source name of the vertex in SameAs Graph SG; R is a multiset of $\mathcal{W} \times \mathcal{W}$ formed by pairs of vertices (μ, ν), $\mu, \nu \in \mathcal{W}$ and each pair (μ, ν) corresponds to an edge in SameAs Graph SG. Figure 6 shows a fragment of a SameAs Graph and the corresponding Source Belief Graph.

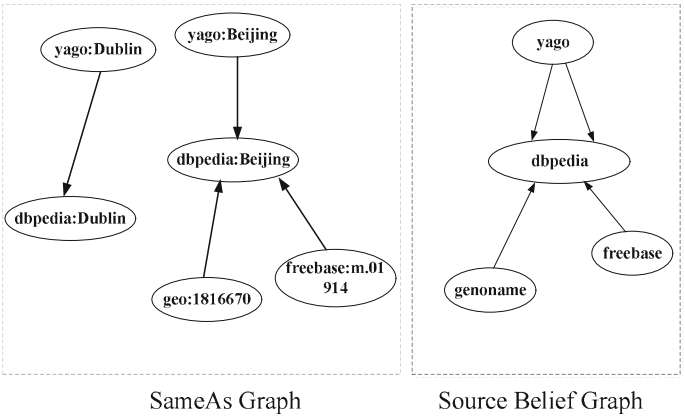


Fig. 6 Example of SameAs Graph and its corresponding Source Belief Graph

The Source Belief Graph indicates that the trustworthiness of different sources can be propagated through the edges. The edge structure of the Source Belief Graph is utilized to produce a global reliability ranking of each source. We let B_{ω_j} denote the set of sources that point to ω_j, $C(\omega_j)$ denote the number of edges coming out of source ω_j and $L(\omega_j, \omega_l)$ present the number of edges that ω_j point to ω_l. The priori belief $BR(\omega_j)$ of source ω_j can be defined as follows:

$$BR(\omega_j) = (1-d) + d \times \sum_{\omega_l \in B_{\omega_j}} \frac{BR(\omega_l)L(\omega_l, \omega_j)}{C(\omega_l)}, \tag{6}$$

where parameter d is a damping factor.

Recent research [10] has indicated that the *owl:sameAs* property does not always mean that the two subjects refer to the same thing in practice. Four incorrect usages of *owl:sameAs* have been identified in Linked Data, including *Same Thing As But Different Context*, *Same Thing As But Referentially Opaque*, *Represents and Very Similar To*. Intuitively, the damping factor d in BeliefRank can be considered that the probability that the usage of owl:sameAs is correct. The experimental results of [10] show that approximately 51% of the usages of owl:sameAs are correct. Therefore, the damping factor is set to 0.51 in this study.

The effectiveness of BeliefRank is significantly affected by the total number of sameAs triples. We extracted 18 millions of sameAs triples from BTC2012 [11], which covers a significant portion of Linked Data, to produce a global reliability of source. BeliefRank reaches a stable stage after 14 iterations when the threshold is set to 0.001. The priori beliefs of 1,402 sources[7] are obtained in this study by using BeliefRank. The top 3 are dbpedia.org with 14.1648, freebase with 11.15 and FOAF with 3.58.

4.2 Truth computation

This section shows the accurate inference of the trustworthiness of the source and the trust value of an object in Linked Data with a scale-free property.

[7] http://1drv.ms/1M2PHoG.

4.2.1 Computing the trustworthiness of sources

A simple method to compute the precision of a source is that regarding the trustworthiness of a source as the percentage of true objects provided by this source. However, we do know for sure which objects are truth. Therefore, we instead compute the trustworthiness $t(\omega_j)$ as the average probability of the object provided by ω_j being true as defined below:

$$t(\omega_j) = \frac{\sum_{o_i \in F(\omega_j)} \tau(o_i)}{\left|F(\omega_j)\right|}, \tag{7}$$

where $F(\omega_j)$ is the set of objects provided by source ω_j.

Accurately estimating the real reliability degree of source ω_j when $\left|F(\omega_j)\right|$ is "small" is difficult for Eq. 7 considering the scale-free property of Linked Data, as discussed in Sect. 1.2. In this study, the trustworthiness $t(\omega_j)$ of source ω_j is smoothed by priori belief $BR(\omega_j)$ based on the averaging strategy as defined as follows:

$$t'(\omega_j) = \frac{NBR(\omega_j) + t(\omega_j)}{2}, \tag{8}$$

$$NBR(\omega_j) = \frac{BR(\omega_j) - min}{max - min}, \tag{9}$$

where $NBR(\omega_j)$ represents the normalized priori belief of source ω_j; max and min indicate the maximum and minimum values of all priori beliefs, respectively.

4.2.2 Computing the trust values of objects

This subsection describes the computation of trust values of objects. First, we analyze a simple case wherein all sources are independent. The trust value $\tau(o_i)$ of object o_i can be defined as follows:

$$\tau(o_i) = \frac{\sum_{\omega_j \in \Omega(o_i)} t'(\omega_j)}{|\Omega(o_i)|}, \tag{10}$$

where $\Omega(o_i)$ represents the set of sources that providing object o_i.

However, Sect. 3.2 shows three helpful correlations from Linked Data sources and objects to effectively distinguish between true and false objects. In this study, we first formulate the object conflict-resolution problem as the Source-Object network analysis problem, which successfully captures all the correlations from objects and Linked Data sources. Subsequently, a message propagation-based method that exploits the Source-Object network structure is introduced to solve this problem. Finally, several important issues that make this method practical are discussed.

In general, the input to our problem includes three parts: (i) objects, which are the values of a certain predicate for the same real-world entity, (ii) Linked Data sources, which provide these objects, e.g., Freebase; and (iii) mappings between objects and Linked Data sources, e.g., which Linked Data sources provide which objects for certain predicate of the same real-world entity. Thus, a set of objects and sources can be structured into a bipartite network. In this bipartite network, source nodes are connected to the object nodes, in which links represent the "provider" relationships. For ease of illustration, we present example network of six sources and four conflicting objects as shown in Fig. 7a. According to the first principle, an object from a reliable source is more trustworthy and thus a source that providing trustworthy objects than other sources. The "provide" relationship between a source and an object also indicates

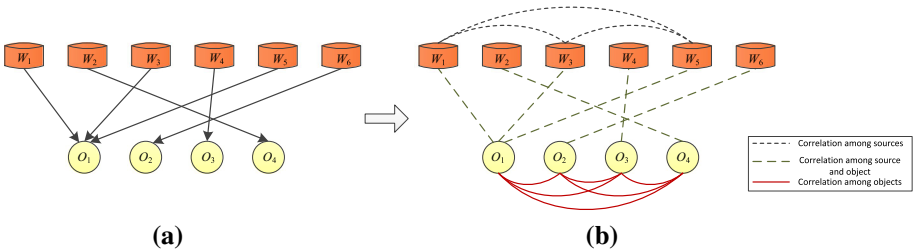


Fig. 7 Illustration of an example source-Object Network. **a** The bipartite network of input data. **b** The Source-Object Network

the interdependent relationship between the trust value of the object and the trustworthiness of the source. Besides the "provider" relationship between the source and object, among objects and among Linked Data sources also have correlations. For instance, because sources $\omega_1, \omega_3, \omega_5$ provide the same object o_1 in Fig. 7a, they have a correlation for any two of these three sources. Therefore, the bipartite network in Fig. 7a can be converted to a heterogeneous information network called the Source-Object Network as shown in Fig. 7b.

The Source-Object Network $G = (V, E)$ contains n Linked Data source nodes $\Omega = \{\omega_1, \ldots, \omega_n\}$ and m conflicting object nodes $O = \{o_1, \ldots, o_m\}$, $V = \Omega \cup O$, connected with edge set E. Owing to three types of correlations of objects and Linked Data sources, the Source-Object Network G has three types of edges $E = E_\Omega \cup E_O \cup E_{\Omega \to O}$, where $E_\Omega \subseteq \Omega \times \Omega$ represents the correlations between sources, $E_O \subseteq O \times O$ indicates the correlations among objects and $E_{\Omega \to O}$ represents the "provided" relationships between sources and objects.

Given a Source-Object Network, which successfully captures three correlations from objects and Linked Data sources, the task is to estimate the reliability of sources and the trust values of all conflicting objects. Each node in G is a random variable that can represent the trust values of objects and trustworthiness of sources. However, we find that the trust values of objects and trustworthiness of sources are assumed to be dependent on their neighbors and independent of all the other nodes in this network. This condition motivates us to select a method based on pMRF, which is a powerful formalism used to model real-world events based on the Markov chain and knowledge of soft constraints. Therefore, the Source-Object Network is represented by pMRF in this study. In fact, pMRF is mainly composed of three components: an unobserved field of random variables, an observable set of random variables, and the neighborhoods between each pair of variables. We let all the nodes $V = \Omega \cup O$ in G be observation variables. Thus, the unobserved variables $Y = Y_\Omega \cup Y_O$ have two types of labels.

1. The unobserved variable y_i is the label of an object node. It indicates whether the corresponding object is the truth, which follows the Bernoulli distribution defined as follows.

$$P(y_i) = \begin{cases} \tau(o_i) & \text{if } o_i \text{ is true, i.e., } y_i = 1, \\ 1 - \tau(o_i) & \text{if } o_i \text{ is false, i.e., } y_i = 0. \end{cases} \tag{11}$$

2. The unobserved variable y_i is the label of Linked Data source node which represents whether the corresponding source is a reliable source and also follows the Bernoulli distribution.

$$P(y_j) = \begin{cases} t(\omega_j) & \text{if } \omega_j \text{ is a reliable source, i.e., } y_j = 1, \\ 1 - t(\omega_j) & \text{if } \omega_j \text{ is a unreliable source, i.e., } y_j = 0. \end{cases} \tag{12}$$

The problem of inferring the trust values of conflicting objects and trustworthiness of sources can be converted to compute the joint distribution of variables in pMRF, which is factorized as follows:

$$P(y_1, \ldots, y_m, \ldots, y_{m+n}) = \frac{\prod_{c \in \mathrm{C}} \psi_c(x_c)}{\sum_{x_c \in X} \prod_{c \in \mathrm{C}} \psi_c(x_c)}, \tag{13}$$

where C denotes the set of all maximal cliques, the set of variables of a maximal clique is represented by x_c ($c \in$ C), and $\psi_c(x_c)$ is a potential function in pMRF.

The belief propagation algorithm [29] is proven to be an exact solution for estimating the marginal probabilities of an unobserved variable when the graph does not have loops. LBP is an approximate algorithm for a loopy graph. LBP (Loopy Belief Propagation) process is designed in this study to estimate the marginal probabilities of the unobserved variable y_i considering the loops. Estimating the marginal probabilities of the unobserved variable is a process of minimizing the graph energy in belief propagation. The key steps of the propagation process are shown as follows.

- **Step I: Initialization** The trust value of object o_i and the probability distribution of $P(y_i)$ are initialized with Eqs. 10 and 11, respectively.
- **Step II: Spreading the belief message** The message from variable y_i to y_j is represented by $m_{i \to j}(y_j)$, $y_j \in \{0, 1\}$. The message $m_{i \to j}(\eta)$ is defined as follows:

$$m_{i \to j}(y_j) = \sum_{y_i \in \eta} U(y_i, y_j)\tau(o_i) \prod_{y_k \in N(o_i) \cap O \setminus \{o_j\}} m_{k \to i}(y_i), \tag{14}$$

  where $N(o_i)$ is the set of neighbors of 0_i; $U(y_i, y_j)$ is a unary energy function, which indicates that if y_i and y_j are the same, then such propagation requires low energy (easy to propagate). Otherwise, high energy, which is difficult to propagate, is required.
- **Step III: Belief assignment** The marginal probability $P(y_i)$ of unobserved variable y_i is updated based on its neighbors and is defined as follows:

$$P(y_i) = \tau(o_i) \prod_{y_j \in N(o_i) \cap O} m_{j \to i}(y_i). \tag{15}$$

The algorithm updates all messages in parallel and assigns the label until the messages stabilizes, i.e., achieve convergence. Although convergence is not theoretically guaranteed, the LBP has been shown to converge to beliefs within a small threshold fairly quickly with accurate results [31]. After they stabilize, we compute the marginal probability $P(y_i)$. Thus, we can obtain the trust values of object and the trustworthiness of source. Given only one truth for a certain predicate of a real-world entity, the true object is o_i when $\tau(o_i)$ is the maximum. To date, we have described the main steps of LBP, but one problem occurs in the algorithm, energy function. This problems are discussed as follows.

Energy function The energy function $U(y_i, y_j)$ denotes the likelihood of a node with label y_i to be connected to a node with label y_j through an edge. The following three types of energy functions exist depending on the types of edges:

- The energy function between sources and objects. A basic principle between sources and objects is that the reliable source tends to provide true objects and unreliable sources to false objects. However, a reliable sources may also provide false objects as unreliable

Table 1 Energy function from sources and objects

Source	Object		Source	
	True	False	Reliable	Unreliable
Reliable	β	$1-\beta$	ε	$1-\varepsilon$
Unreliable	$1-\delta$	δ	$1-\varepsilon$	ε

Table 2 Energy function between objects

Object	Object	
	True	False
True	$S(o_i, o_j)$	$1-S(o_i, o_j)$
False	$1-S(o_i, o_j)$	$S(o_i, o_j)$

sources to true objects. In this study, we let β denote the likelihood between reliable sources and true objects, whereas δ denotes the likelihood between unreliable sources and false objects. Therefore, the energy function between sources and objects is shown in Table 1.

– The energy function among objects. The more similar the two objects are, the greater is the probability of them having the same trust values. Therefore, a positive correlation exists between the energy function and the similarity $S(o_i, o_j)$ between object o_i and o_j, as shown in Table 2.
– The energy function among sources. We assume that the more same objects two different sources provide, the more similar the trustworthiness of the two sources are. The coefficient $\varepsilon = |F(\omega_i) \cap F(\omega_j)|/max(|F(\omega_i)|, |F(\omega_j)|)$ is used to denote the likelihood between sources ω_i and ω_j, where $F(\omega_i)$ is the set of objects provided by source ω_i as shown in Table 1.

4.3 Practical issues and time complexity

In this subsection, we discuss several important issues, including similarity functions and missing values, to ensure the practicality of our method. At the end of this section, we analyze the time complexity of the proposed method.

Similarity functions The energy function between objects depends on the similarity function. We respect the characteristic of each data type and adopt different similarity functions to describe the similarity degrees. We have discussed two similarity functions for numerical and categorical data, which are the two most common data types in Sect. 3.2.

Apart from these two functions, we also adopt a most commonly used similarity function based on depth information of concepts to measure the semantic similarity [14,36] in this study. Following the Description Logic Terminology [12] in the Semantic Web community, the nodes of RDF Graph consist of a set of concepts denoting conceptual abstractions of things, and a set of instances representing real-world entities. Concepts in RDF Graph contains terminology box (TBox) which describes constraints on the structure of the domain, similar to the conceptual schema in database setting, are used to denote concept hierarchies and usually referred as ontology classes. Assertion box (ABox) about entity instances are usually referred as ontology instances. A tiny example using the above notions is shown in Fig. 8.

Semantic proximities between concepts are defined based on how closely they are related in the hierarchies of the rdfs:subClassOf relationships. Let c_i, c_j be two concepts in a given

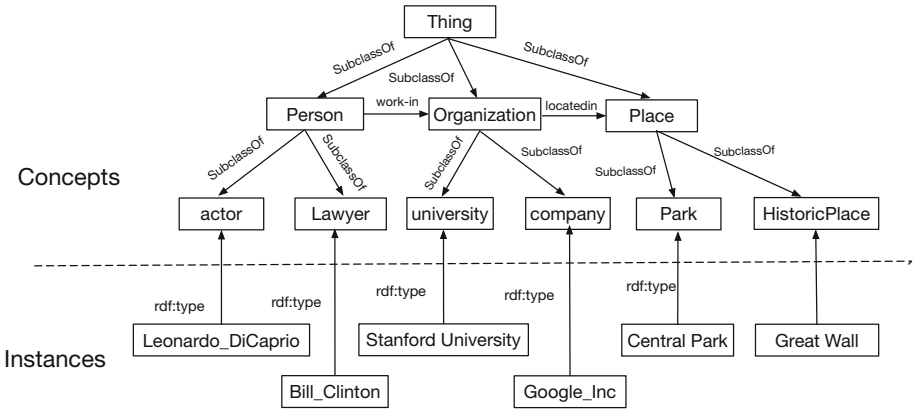


Fig. 8 A example of concepts hierarchy

linked data sources. The semantic proximities function between c_i and c_j is defined as follows:

$$sim(c_i, c_j) = \frac{2 * depth(c_{ij})}{depth(c_i) + depth(c_j)}, \tag{16}$$

where c_{ij} is the most specific concept that is a shared ancestor of the two concepts, and $depth(c_i)$ gets the depth of c_i in the original concepts hierarchy (i.e., without inferencing).

This similarity function gives two meanings: (1) the similarity between lower-level concepts should be considered more similar than those concepts between upper-level concepts. For example in Fig. 4, the concept pair lawyer and actor are more similar than the concept pair person and organization; (2) the structural proximity of two concepts strengthens as their depths increase.

Missing values Linked Data are built on the Open World Assumption, which states that what is not known to be true is simply unknown. Therefore, for the sake of simplicity in this study, we assume that all missing values are not known to be true.

Time complexity We let m denote the total number of different objects, n is the number of Linked Data sources, and r is the number of iterations of TruthDiscover. The time complexity of TruthDiscover is $O((m+n)^2 \times r)$. BeliefRank can produce a global reliability ranking of each source through an offline process. Therefore, the time complexity of TruthDiscover is $O((m+n)^2 \times r)$ and is experimentally validated in Sect. 5.4.

5 Experiments

Three experiments are conducted in six real datasets to validate the effectiveness of our approach. The experimental results show that TruthDiscover outperforms the existing approaches in resolving object conflicts when confronts with the challenge of data having a scale-free property. The experiment setup is discussed in Sect. 5.1, and the experimental results are presented in Sects. 5.3 and 5.4.

5.1 Experiment setup

5.1.1 Multi-values filtering

As discussed in Sect. 1.4 , TruthDiscover focuses on three reasons for object conflicts, whereas the fourth (multi-valued predicates) is left for the future. A method to distinguish multi-valued predicates is necessary to assess the applicability of TruthDiscover. Two effective rules in this study is that 1) if the type of predicates is "owl:FunctionalProperty", this predicate only have one unique object according to the OWL Web Ontology Language, 2) if a source provides more than one objects for a predicate of a real-world entity, this predicate is the multi-valued predicate that is used to filter other multi-valued predicate automatically. The method based on these two rules achieves relatively high precision (ranging from 0.96 to 0.98) on the six datasets. Therefore, this method meets the desired objectives compared with manual annotation method.

5.1.2 Performance measures

In the experiments, we have two types of data in our datasets: numerical data and string data. For these two types of data, only one truth is selected from multiple different objects. The precision is computed as the percentage of the output objects that are consistent with a gold standard. Meanwhile, the recall is computed as the percentage of the values in the gold standard being output as correct. However, the recall is equivalent to the precision when all sources have been fused [19]. Therefore, the precision as a unified measure is adopted in the experiments for the two types of data.

5.1.3 Baseline methods

We select six well-known truth discovery methods as baselines. These methods are evaluated using the same datasets in the experiments.

Voting Voting regards the object with the maximum number of occurrences as truth. This method is a straightforward method.

Sums (Hubs and Authorities) [16] This method regards the object which supported by the maximum number of reliable sources as true. In this study, a source is recognized as a reliable source if its trustworthiness score exceeds 0.5.

TruthFinder [37] This method is a seminal work that is used to resolve conflicts based on the estimation of source reliability. TruthFinder adopts the Bayesian analysis to infer the trustworthiness of sources and the probabilities of a value being true.

ACCUCOPY [6] This method is a popular truth discovery algorithm, which obtains the highest precision among all methods in [19]. ACCUCOPY considers the copying relationships between the sources, the accuracy of data sources, and the similarity between values.

F-quality assessment [27] This method is a popular algorithm used to resolve conflicts in Linked Data. Three factors, namely the source quality, data conflicts, and confirmation of values from multiple sources, are leveraged to decide which value should be the true value.

ObResolution [23] This method is a latest algorithm used to resolve object conflicts in Linked Data. This method models all the clues from sources and objects by a heterogeneous information network called the Source-Object Network in a unified framework.

The parameters of the baseline methods are set according to the suggestions of the author. The experiments are performed on a desktop computer with Intel Core i5-3470 CPU 3.2

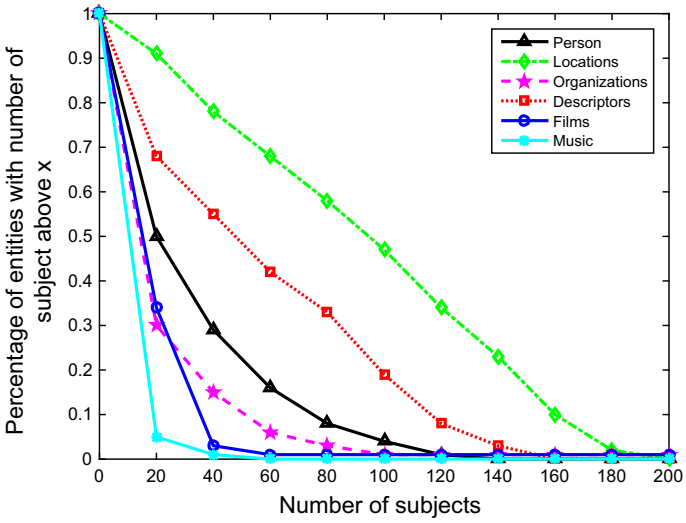


Fig. 9 Distribution of subjects

GHz with 4 GB main memory, and Microsoft Windows 7 professional operating system. All baseline methods are executed in the Eclipse (Java) platform[8] by a single thread.

5.2 Real-world dataset collection

The experimental results for the six datasets show that TruthDiscover outperforms the baseline methods in determining the truth from multiple different objects in Linked Data with a scale-free property.

Data collection Three experiments are conducted on the six datasets including *persons*, *locations*, *organizations*, *descriptors*, *films* and *music*. The first four datasets are constructed based on the OAEI2011 New York Times dataset, which is a well-known and carefully created dataset of Linked Data. Two other domains, including *films* and *music*, are constructed through SPARQL queries over DBpedia to draw more robust conclusions. The construction process of datasets mainly involves the following steps:

(i) **Identity subjects** For each entity of the six domains, we perform entity co-reference resolution through the API of a well-known tool sameas.org,[9] to identify subjects for the same real-world entities. The cumulative distribution function distribution of the number of subjects on the six datasets is shown in Fig. 9, wherein the x-axis shows the number of subjects per entity (for instance a location or film) in our datasets and the y-axis shows the percentage of the entities in our datasets. We observe from Fig. 9 that the number of subjects for most entities in our datasets ranges from 0 to 100 and the entity whose number of subjects is more than 140 is a small proportion, for example more than 170 different subjects can be found in LOD for the entity *Beijing*.

(ii) **Schema mapping** Schema matching aims to combine a few predicates for different sources with the same meaning together (e.g., *rdf:labels* and *foaf:names*), which is an important step for constructing our datasets. We adopted a method combining automatic matching

[8] https://www.eclipse.org/.

[9] http://sameas.org/.

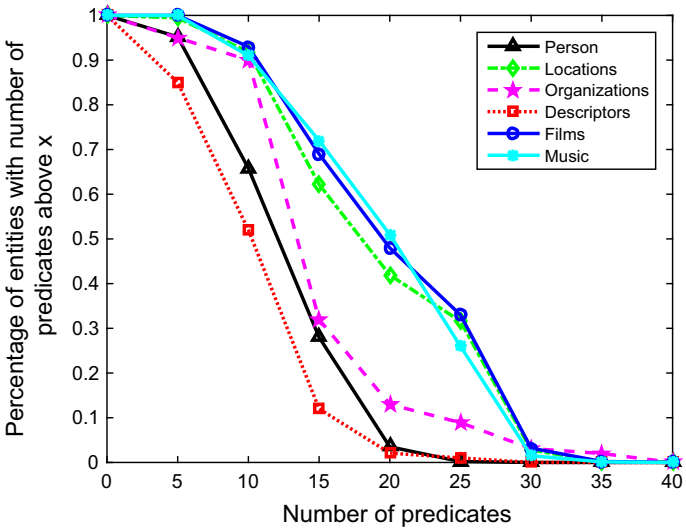


Fig. 10 Distribution of global predicates

Table 3 Top 8 global predicates on the six datasets

ID	Global predicates	Numbers
1	http://www.w3.org/2000/01/rdf-schema#label	125,515
2	http://dbpedia.org/ontology/producer	59,346
3	http://dbpedia.org/ontology/releaseDate	46,389
4	http://dbpedia.org/ontology/runtime	35,876
5	http://dbpedia.org/ontology/birthDate	32,934
6	http://dbpedia.org/ontology/weight	23,123
7	http://dbpedia.org/ontology/height	19,233
8	http://dbpedia.org/ontology/populationTotal	19,234

and manual annotation to produce more accurate schema mapping results. Firstly, features (*Property Similarity*, *Value Overlap Ratio*, *Value Match Ratio* and *Value Similarity Variance*) are selected based on the description provided by [34]. These selected features can achieve good performance in Linked Data. Then after, we leverage Random Forest and SVM model which achieve the best F1-Measure in [34] as classifier for schema matching. Manual annotation is used to break the tie when an agreement is unreachable on a predicate between these two classifiers. This method achieves relatively high precision (ranging from 0.92 to 0.97) on the six datasets.

A strict manual annotation process is established to ensure the quality of the annotation in this paper. We have nine undergraduate students in their junior or senior year from the computer science department. They were asked to combine the predicates with the same meaning together by using their background knowledge. This process mainly involved the following steps:

(a) We provided the annotators annotated examples and annotation guidelines.
(b) Every two annotators are asked to label the same dataset predicates.

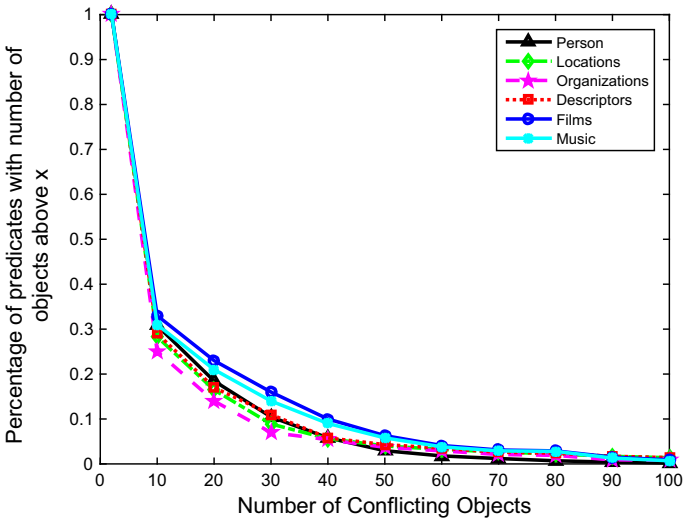


Fig. 11 Distribution of objects

Table 4 Statistics of the six datasets

Datasets	#Entities	#Sources	#Global predicates	#Triples
Person	4978	222	51,054	637,184
Locations	1910	184	36,099	498,510
Organizations	3044	194	53,270	608,800
Descriptors	498	170	3250	39,840
Films	7542	138	150,840	2,257,200
Music	7131	114	164,013	2,281,920

(c) Occasionally, two annotators have different opinions on a predicate. A third annotator is asked to break this tie. The annotation results from the two annotators are measured by Cohens kappa coefficient [2]. The agreement coefficient of the six datasets is set to be at least 0.75.

Predicates that are combined with the same/similar meaning are called *global predicate* in this study. After manually matching the predicates, the average number of global predicates for *people*, *locations*, *organizations*, *descriptors*, *films* and *music* is 10, 19, 17, 9, 20 and 23, respectively. The distribution of number of global predicates in our datasets is shown in Fig. 10. We find that the number of global predicates for most entities (approximate 90%) in *Person*, *Organizations* and *Descriptors* ranges from 0 to 15 and ranges from 0 to 35 for datasets Location, *Films* and *Music*. Only less than 1% entity has more than 30 global predicates, such as the entity *Cook_Islands* has 35 global predicates. We also provide a list about the top 8 global predicates according to the number of triples as shown in Table 3. Figure 11 shows the cumulative distribution of the number of objects for every global predicate in our dataset. We find that the number of objects for most global predicates (approximate 90%) ranges from 2 to 30, and only less than 1% global predicate have more than 60 different objects.

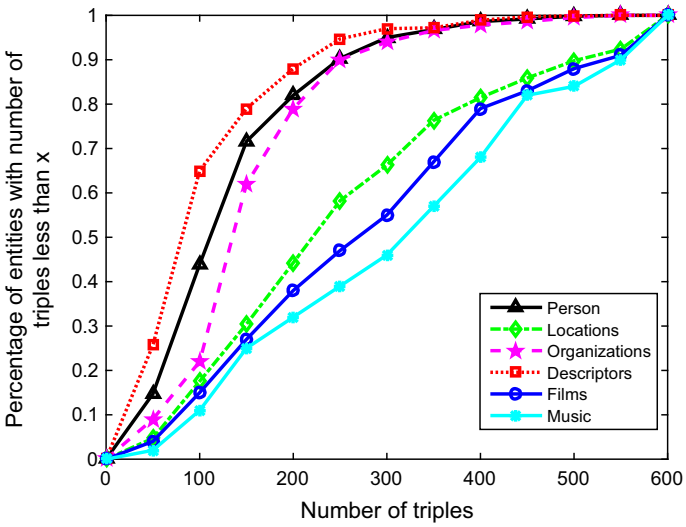


Fig. 12 Distribution of triples

The statistics of the six datasets are shown in Table 4. The column "#Entities" represents the number of entities for each of the six domains. The column "#Sources" shows the number of sources that we have crawled originatively. The column "#Global Predicates." represents the total number of global predicates that have different objects and the column "#Triples" shows the number of triples for each dataset. In addition, we provide the cumulative distribution of the number of triples for each entity on the six datasets as shown in Fig. 12 to show more details. We find that the number of triples for most entities (approximate 90%) in datasets *Person*, *Organizations* and *Descriptor*s ranges from 0 to 250 and ranges from 0 to 500 for datasets *Location*, *Films* and *Music*.

One truth is selected from multiple different objects for experimental verification. The manually labeled results are regarded as the ground truth used in the evaluation in this study. A strict manual annotation process, which is similar to the process adopted in **Schema mapping**, is established to ensure the quality of the annotation.

5.3 Effectiveness evaluation

In the first experiment, except for the six baseline methods as discussed in Sect. 5.1.3, two other baseline methods, including Baseline1 and Baseline2, are selected in order to evaluate the effectiveness of the two strategies adopted in TruthDiscover. Baseline1 removes the priori belief of all sources, and Baseline2 ignores the interdependencies between objects used in TruthDiscover. The following observations are drawn from the statistical data presented in Fig. 13.

1. TruthDiscover outperforms the first six baseline methods, including Voting, Sums, TruthFinder, ACCUCOPY, F-Quality Assessment and ObResolution with regard to precision. This finding can be attributed to the difficulty in accurately estimating the reliability degree of small sources in Linked Data. Two strategies are adopted to reduce the effect of scale-free property in TruthDiscover. One strategy involves the topological properties of the Source Belief Graph to estimate the priori beliefs of sources for smoothing the trust-

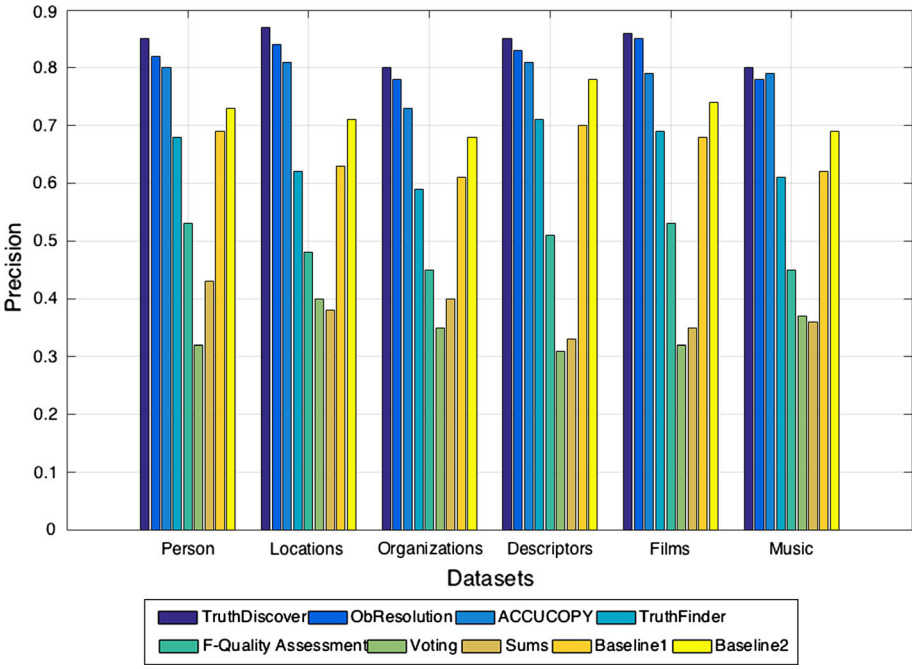


Fig. 13 Performance comparison of the six datasets

worthiness of sources. The other strategy uses MRF to infer the trust values of objects by modeling the interdependencies between the objects.

2. In addition, the precision of Baseline2 and Baseline1 is lower than TruthDiscover, which indicates that BeliefRank and Source-Object network are effective in reducing the effect of "small" sources.
3. The Baseline1 has higher precision than TruthFinder. In fact, Baseline1 adopts Bayesian analysis to infer the trustworthiness of sources as TruthFinder does. The most important difference between Baseline1 and TruthFinder is that two different methods are adopted to model the interdependencies between objects. TruthFinder uses a fixed parameter to control the influence of related objects; however, an appropriate fixed parameter for all objects is hard to determine. Therefore, TruthFinder is not effective. Baseline1 considers influence in a principled fashion and can automatically adjust the influence between objects depending on MRF model. Therefore, Baseline1 outperforms the TruthFinder in six datasets.

The second experiment is conducted to validate the effectiveness of four baseline methods including TruthFinder, ACCUCOPY, F-Quality Assessment and ObResolution which obtain top four highest precision in the first experiment, with regard to the three reasons for inconsistency. The following observations are drawn from the statistical data presented in Table 5.

1. The average precision of the five methods varies for the different reasons. These methods achieve lowest precision in reasons of *out-of-date*, which indicates these methods based only on source reliability estimation are insufficient to resolve conflicts of *out-of-date*; thus, additional information is required.

Table 5 Performance comparison with regard to three reasons for inconsistency

Reasons	Methods	Datasets					
		Persons	Locations	Organizations	Descriptors	Films	Music
Out-of-date	ObResolution	0.47	0.31	0.41	0.43	0.47	0.45
	ACCUCOPY	0.46	0.29	0.39	0.43	0.44	0.42
	TruthFinder	0.42	0.29	0.35	0.33	0.35	0.35
	F-Quality	0.31	0.32	0.33	0.24	0.18	0.36
	Our Method	0.49	0.51	0.45	0.49	0.50	0.47
Variety	ObResolution	0.90	0.75	0.81	0.85	0.85	0.83
	ACCUCOPY	0.90	0.73	0.79	0.87	0.87	0.80
	TruthFinder	0.73	0.68	0.59	0.78	0.72	0.61
	F-Quality	0.57	0.48	0.42	0.51	0.53	0.42
	Our Method	0.93	0.90	0.88	0.91	0.94	0.87
Pure Errors	ObResolution	0.93	0.76	0.88	0.87	0.90	0.87
	ACCUCOPY	0.92	0.73	0.87	0.88	0.93	0.86
	TruthFinder	0.78	0.81	0.81	0.83	0.93	0.88
	F-Quality	0.63	0.68	0.71	0.84	0.91	0.75
	Our Method	0.95	0.97	0.89	0.90	0.96	0.89

2. For the three reasons, TruthDiscover outperforms the four baseline methods with regard to precision because two effective strategies are adopted.

5.4 Efficiency evaluation

5.4.1 Convergence analysis

Two experiments are conducted to validate the convergence of TruthDiscover. The first experiment is conducted to analyze the convergence of TruthDiscover. The second experiment is performed to show the relation between precision and iteration.

We formulate the problem of resolving conflicts as an iterative computation problem because of the interdependencies between the trust value of objects and the trustworthiness of sources. Therefore, convergence significantly affects the performance of TruthDiscover. Figure 14 shows the average change in the trust value of objects after each iteration. The change rapidly decreases in the first five iterations and then reaches a stable stage until the convergence criterion is satisfied. The average number of iterations for *persons*, *locations*, *organizations*, *descriptors*, *films* and *music* is 23, 24, 25, 13, 28 and 29, respectively.

The second experiment is conducted to analyze the relationship between precision and iteration. The results are shown in Fig. 15. The precision of TruthDiscover increases with the number of iterations increases and reaches a stable stage until the convergence criterion is satisfied.

5.4.2 Sensitivity analysis

We also studied the effect of the parameter β, δ on our methods. As discussed in Sect. 4.2.2, β indicates the likelihood between reliable source and true object, whereas δ denotes the

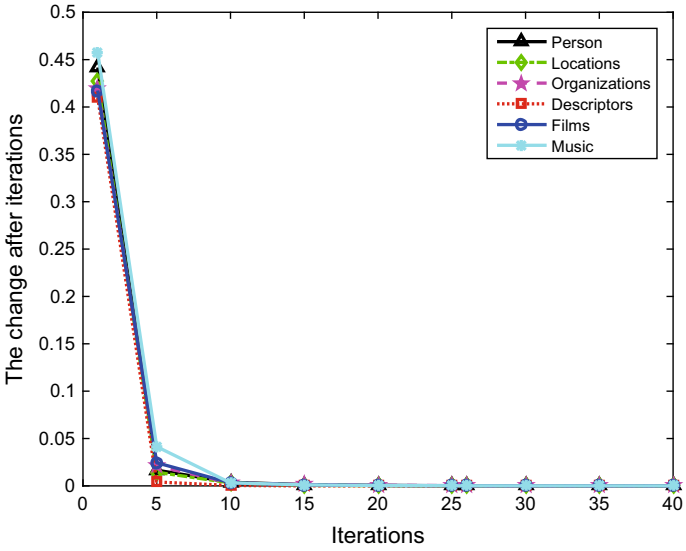


Fig. 14 Change in the trust values of objects after each iteration

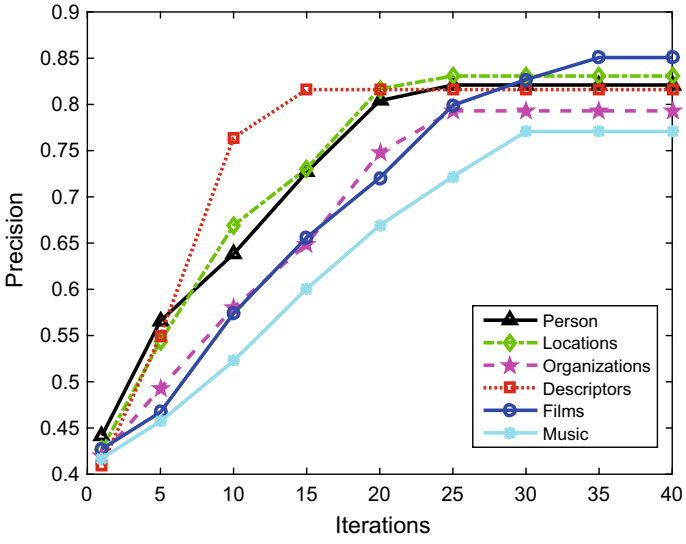


Fig. 15 Relation between precision and iteration

likelihood between unreliable source and false object. Figure 16 shows that the precision of our method varies in different values of β, δ in the same dataset, and TruthDiscover achieves best precision on six datasets with different values of β, δ ($\beta = 0.9, \delta = 0.7$ for *Persons*, for *Music* $\beta = 0.7, \delta = 0.9$). Therefore, parameters β, δ are sensitive to different datasets because different Linked Data datasets have different qualities [38]. TruthDiscover uses different β, δ for different datasets to optimize the performance of our method.

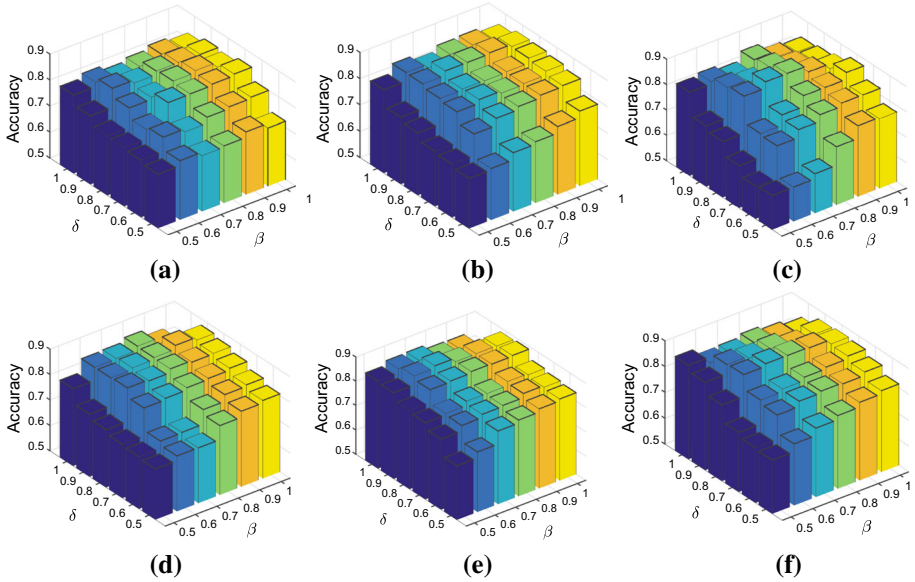


Fig. 16 Sensitive analysis in six Linked Data datasets. **a** Persons, **b** locations, **c** organizations, **d** descriptors, **e** films, **f** music

5.4.3 Time efficiency evaluation

We sample different numbers of different objects to determine the computational complexity of TruthDiscover in a single machine. Figure 17 shows the running time for different objects. The power law function is adopted to fit the relationship between running time and number of objects. We find that the relationship between running time and the number of objects typically follows the power law $y = a * x^b$, where a is 39.844 and b is 2.037, which verifies the analysis of the time complexity of TruthDiscover discussed in Sect. 4.3.

The experimental results in Sects. 5.3 and 5.4 show that two strategies are effective in reducing the effect of scale-free property. These results indicate that the performance of TruthDiscover is robust and consistent in various domains.

In addition, an easy-to-use system has been developed to visualize the Source Belief Graph and process of truth computation [22]. It allows users to search their interested subject via a Web-based interface.

6 Conclusion and future work

In this study, the observations on six datasets reveal that Linked Data has a scale-free property. This property indicates that only a few sources have numerous objects, and most of the sources provide significantly few objects. Thus, the existing work cannot be extended trivially to resolve object conflicts in Linked Data. In this study, the problem of resolving object conflicts in Linked Data is formulated as a truth discovery problem. A truth discovery approach called TruthDiscover is proposed to determine the most trustworthy object, which leverages the topological properties of the Source Belief Graph and the interdependencies between objects

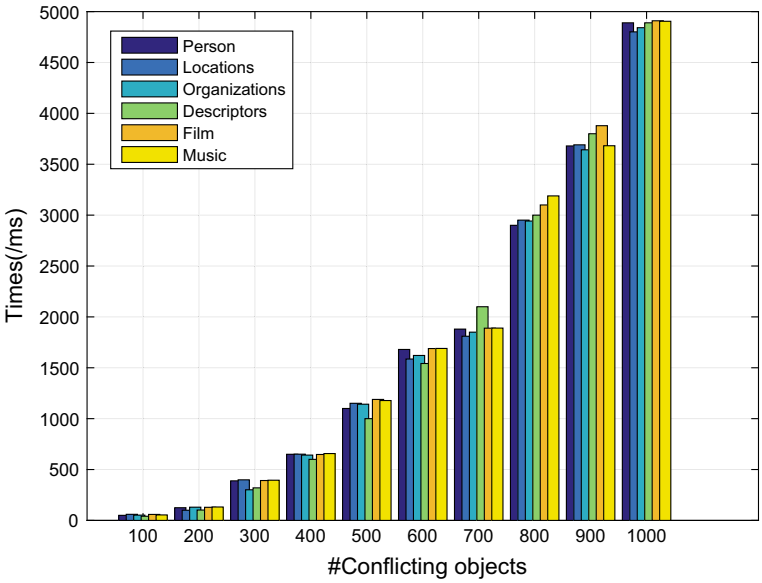


Fig. 17 Running time of different numbers of entities

to infer the trustworthiness of sources and the trust values of objects. The experimental results in six real-world datasets show that this method exhibits satisfactory precision.

A potential direction for future research is to focus on resolving *out-of-date* conflicts by leveraging truth discovery and provenance information. Another potential future direction is to identify the copying relations of different sources to improve performance.

Acknowledgements This work is sponsored by "The Fundamental Theory and Applications of Big Data with Knowledge Engineering" under the National Key Research and Development Program of China with grant number 2016YFB1000903; National Science Foundation of China under Grant Nos. 61370019, 61502377, 61672419, 61672418, 61532004, and 61532015; Project of China Knowledge Centre for Engineering Science and Technology; MOE Research Center for Online Education Funds under Grant No. 2016YB165; Innovative Research Group of the National Natural Science Foundation of China (61721002); Ministry of Education Innovation Research Team No. IRT17R86.

References

1. Bleiholder J, Naumann F (2008) Data fusion. ACM Comput Surv 41(1):137–153
2. Carletta J (1996) Assessing agreement on classification tasks: the kappa statistic. Comput Linguist 22(2):249–254
3. Dayal U, Center FC (1983) Processing queries over generalization hierarchies in a llultidatabare system. In: PVLDB, Florence, Italy
4. Ding L, Shinavier J, Finin T, McGuinness DL. (2010) owl: sameas and linked data: an empirical study
5. Ding L, Shinavier J, Shangguan Z, McGuinness DL (2010) Sameas networks and beyond: analyzing deployment status and implications of owl: sameas in linked data. In: ISWC, Shanghai, China. Springer, pp 145–160
6. Dong XL, Berti-Equille L, Srivastava D (2009) Integrating conflicting data: the role of source dependence. In: PVLDB, Lyon, France, vol 2. VLDB Endowment, pp 550–561

7. Dong XL, Gabrilovich E, Murphy K, Dang V, Horn W, Lugaresi C, Sun S, Zhang W (2015) Knowledge-based trust: estimating the trustworthiness of web sources. In: PVLDB, Hawai'i, USA, vol 8. VLDB Endowment, pp 938–949
8. Dutta A, Meilicke C, Ponzetto SP (2014) A probabilistic approach for integrating heterogeneous knowledge sources. In: ESWC, Crete, Greece. Springer, pp 286–301
9. Glaser H, Jaffri A, Millard IC (2009) Managing co-reference on the semantic web. In: WWW, Madrid, Spain. Citeseer
10. Halpin H, Hayes PJ, McCusker JP, McGuinness DL, Thompson HS (2010) When owl: sameas isn't the same: an analysis of identity in linked data. In: ISWC, Shanghai, China. Springer, pp 305–320
11. Harth A (2012) Billion triples challenge data set. http://km.aifb.kit.edu/projects/btc-2012/
12. Horrocks I (2008) Ontologies and the semantic web. Commun ACM 51(12):58–67
13. Hu W, Jian N, Qu Y, Wang Y Gmo (2005) A graph matching for ontologies. In: K-CAP, Banff, Canada, pp 41–48
14. Hu W, Qu Y, Cheng G (2008) Matching large ontologies: a divide-and-conquer approach. Data Knowl Eng 67(1):140–160
15. Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5):604–632
16. Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5):604–632
17. Li Q, Li Y, Gao J, Su L, Zhao B, Demirbas M, Fan W, Han J (2014) A confidence-aware approach for truth discovery on long-tail data. In: PVLDB, Hangzhou, China, vol 8. VLDB Endowment, pp 425–436
18. Li Q, Li Y, Gao J, Zhao B, Fan W, Han J (2014) Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In: SIGMOD, Utah, USA. ACM, pp 1187–1198
19. Li X, Dong XL, Lyons K, Meng W, Srivastava D (2012) Truth finding on the deep web: is the problem solved? In: PVLDB, Istanbul, Turkey, vol 6. VLDB Endowment, pp 97–108
20. Li Y, Gao J, Meng C, Li Q, Su L, Zhao B, Fan W, Han J (2015) A survey on truth discovery. arXiv preprint arXiv:1505.02463
21. Li Y, Li Q, Gao J, Su L, Zhao B, Fan W, Han J (2015) On the discovery of evolving truth. In: ACM SIGKDD, Sydney, Australia. ACM, pp 675–684
22. Liu W, Liu J, Duan H, Jian Z, Wei H, Wei B (2017) Truthdiscover: Resolving object conflicts on massive linked data. In: WWW[Demo], Perth, Australia
23. Liu W, Liu J, Duan H, Wei H, Wei B (2017) Exploiting source-object network to resolve object conflicts in linked data. In: ESWC, Portoroz, Slovenia. Springer
24. Manola F, Miller E, McBride B Rdf1.1 primer. http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/
25. McGuinness DL, Van Harmelen F et al (2004) Owl web ontology language overview. http://www.w3.org/TR/owl-ref/#sameAs-def
26. Mendes PN, Mühleisen H, Bizer C (2012) Sieve: linked data quality assessment and fusion. In: EDBT/ICDT Berlin, Germany. ACM, pp 116–123
27. Michelfeit J, Knap T, Nečaskỳ M (2014) Linked data integration with conflicts. arXiv preprint arXiv:1410.7990
28. Nolle A, Meilicke C, Chekol MW, Nemirovski G, Stuckenschmidt, H (2016) Schema-based debugging of federated data sources. In: ECAI, pp 381–389
29. Pearl J (1982) Reverend Bayes on inference engines: a distributed hierarchical approach. In: AAAI, Pennsylvania, USA, pp 133–136
30. Qu Y, Hu W, Cheng G (2006) Constructing virtual documents for ontology matching. In: WWW, Edinburgh Scotland, United kingdom. ACM, pp 23–31
31. Rayana S, Akoglu L (2015) Collective opinion spam detection: bridging review networks and metadata. In: SIGKDD, Melbourne, Australia. ACM, pp 985–994
32. Srivastava D, Venkatasubramanian S (2010) Information theory for data management. In: SIGMOD, Indiana, USA. ACM, pp 1255–1256
33. Vydiswaran V, Zhai C, Roth D (2011) Content-driven trust propagation framework. In: ACM SIGKDD, CA, USA. ACM, pp 974–982
34. Wang H, Fang Z, Zhang L, Pan JZ, Ruan T (2015) Effective online knowledge graph fusion. In: ISWC, Pennsylvania, USA. Springer, pp 286–302
35. Wang S, Englebienne G, Schlobach S (2008) Learning concept mappings from instance similarity. In: ISWC, Karlsruhe, Germany, vol 5318. Springer, p 339
36. Wu Z, Palmer M (1994) Verbs semantics and lexical selection. In: ACL, New Mexico, USA. Association for Computational Linguistics, pp 133–138
37. Yin X, Han J, Yu PS (2008) Truth discovery with multiple conflicting information providers on the web. IEEE Trans Knowl Data Eng 20(6):796–808

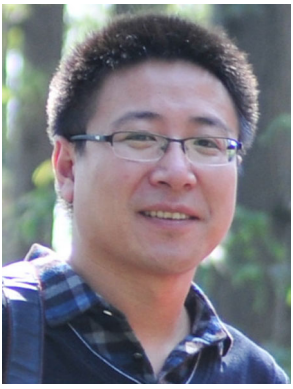
38. Zaveri A, Rula A, Maurino A, Pietrobon R, Lehmann J, Auer S (2016) Quality assessment for linked data: a survey. Semantic Web 7(1):63–93
39. Zhao B, Rubinstein BI, Gemmell J, Han J (2012) A Bayesian approach to discovering truth from conflicting sources for data integration. In: PVLDB, Istanbul, Turkey, vol 5. VLDB Endowment, pp 550–561
40. Zheng Y, Li G, Li Y, Shan C, Cheng R (2017) Truth inference in crowdsourcing: is the problem solved? In: PVLDB, Munich, Germany, vol 10, pp 541–552

Wenqiang Liu is currently working toward the PhD degree in computer science at Xian Jiaotong University. He received the B.S. degree in Computing Science from Qingdao University of Science and Technology in 2012. His research interests include Linked Data, knowledge base and data mining.

Jun Liu is currently a professor in the Department of Computer Science, Xian Jiaotong University. He has published more than 70 research papers in various journals and conference proceedings. He received the B.S., M.S., and PhD degrees from Xian Jiaotong University, China, in 1995, 1998, and 2004, respectively, all in computer science. His main research interests include text mining, data mining and e-learning.

Bifan Wei is currently an engineer in the Department of Computer Science, Xian Jiaotong University. He received the B.S. degree in Aircraft Dynamics Engineering from Beijing University of Aeronautics and Astronautics in 2000, the PhD degree in computer science in 2014 from Xian Jiaotong University, China. His research interests include web data mining, faceted search and taxonomy learning.

Haimeng Duan is currently working toward the M.S. degree in computer science at Xian Jiaotong University. She received the B.S. degree in Computing Science from Central South University in 2015. Her research interests include Linked Data, knowledge base and data mining.



Wei Hu is currently a professor in the Department of Computer Science, Nanjing University. He has published more than 100 research papers in various journals and conference proceedings. He received his PhD degree in Computer Software and Theory in 2009, and his B.S. degree in Computer Science and Technology in 2005, both from Southeast University. His main research interests include data integration and web application.