Journal of Intelligent Information Systems (2021) 57:25–50
https://doi.org/10.1007/s10844-020-00627-4

A survey on data fusion: what for? in what form? what is next?

Gabrielle Karine Canalle[1] · Ana Carolina Salgado[1] · Bernadette Farias Loscio[1]

Received: 20 November 2019 / Revised: 15 October 2020 / Accepted: 15 October 2020 /
Published online: 2 November 2020

Abstract
Data fusion is the process of merging records from multiple sources which represent the same real-world object into a single representation. This review of the literature concerns Data Fusion in the context of data integration, i.e., the integration of structured and semi-structured data from the same domain, and provides an overview of this field of research. We present why data fusion is becoming increasingly necessary, what it is used for (What for?), what methods and solutions for data fusion have been proposed in the literature (In what form?), what research challenges are still open in the data fusion area and what future research directions could usefully take (What is next?)

Keywords Data integration · Data fusion · Truth discovery

1 Introduction

The Big Data era has produced petabytes of data together with several challenges, including those of attempting to identify and fuse data which represent the same real world object. In general, gathering a very significant amount of data leads to having a substantial volume of contradictory and redundant data. In this scenario, data that describe the same object can come from multiple sources and may contain conflicting information. For example, a Google search on *"What is the population of the city of Recife – Brazil?"*, will obtain different results, viz. *"1,625,583 population"*, *"1,633,697 population"* and *"1,537,704 population"*.

Due to incomplete, erroneous, and out-of-date data, data from different sources of the same domain may conflict with each other (i.e., different values of the same attribute of

✉ Gabrielle Karine Canalle
gkc@cin.ufpe.br

Ana Carolina Salgado
acs@cin.ufpe.br

Bernadette Farias Loscio
bfl@cin.ufpe.br

[1] Federal University of Pernambuco, Recife, Brazil

an entity). The main reasons for this are an increase in the volume of conflicting data that are published on the Web as well as the fact that people are using the Web to spread false information. Currently, the concepts of data quality and trustworthiness have become more important than ever. Thus, Data Fusion, the focus of this survey, has become an important topic of research that aims to detect and solve data conflicts from multiple sources. Nowadays, most Data Fusion approaches aim to resolve conflicts based on the trustworthiness of the sources that provide the data. In these approaches, the notion that guides them is that the more reliable data sources are, the more accurate the data they provide will be.

Data fusion is also applied in different fields, always with the same main purpose: to provide a unified view of data, thereby resolving conflicts and finding truth values. Examples of such applications include sensor data fusion (Sethi and Sarangi 2017), linked data fusion (Michelfeit et al. 2014; Liu et al. 2017b), knowledge fusion (Preece et al. 2001), and information retrieval (Wu 2012a).

In the literature, the fusion of unstructured data is generally referred to as Information Fusion. Typical Information Fusion problems involve the integration of multi-source information for signal and image processing, knowledge representation, and inference. These areas have been the objective of considerable research over recent years (Xu and Yu 2017). In this context, the feasibility and advantages of applying granular computing have been investigated. Granular computing is an umbrella term that covers any theories, methodologies, techniques, and tools that make use of information granules in problem solving. It is the processing of complex information entities (information granules), which gives rise to the process of data abstraction and derivations of knowledge from information (Wang 2010).

In the context of Data Integration i.e., the integration of structured and semi-structured data from multiple data sources of the same domain, the trend of using Data Fusion is increasing due to the growth in the need to integrate data. Recent decades have seen extensive research in this field. Nowadays, far from Data Fusion having been solved, researches with new approaches and methods for Data Fusion are continuously being developed (e.g., Bleiholder (2010), Dong et al. (2014), Yin and Tan (2011), Hara et al. (2013), Xie et al. (2017), Wang et al. (2017), Zhang et al. (2016), Fang et al. (2017b), Pasternack and Roth (2013), Pochampally et al. (2014), Li et al. (2014a), and Wang et al. (2015)). Early approaches were based on rules and applied functions (i.e., the average value, the most recent value) in order to resolve conflicts. Recently, approaches have used several characteristics, such as source quality (Broelemann et al. 2017; Li et al. 2016; Xiao et al. 2016; Zhao et al. 2012), copying between data sources (Fang et al. 2017b; Dong et al. 2009a; 2009b), object difficulty (i.e., recognizing how difficult it is to infer the real true value of the object and to tackle this issue) (Wang et al. 2017; Galland et al. 2010), object relationship (Nakhaei and Ahmadi 2017; Pasternack and Roth 2010), and object popularity (Fang 2017). Furthermore, there are approaches for distributed data (Wang et al. 2017) and ensemble approaches (i.e., combining various competing models) (Fang et al. 2016; Berti-Équille 2015).

In general, the most recent studies combine iteratively estimating the quality of the source with truth discovery. The basic principle is that sources that provide true information are more reliable, and that it is more likely that the information provided by reliable sources is true.

The constant research in the Data Fusion field reflects its importance. Thus, in this paper, we examine the field, by conducting a review of the literature. Other studies have presented the state-of-the-art in Data Fusion, e.g., Bleiholder and Naumann (2008) Li et al. 2012, 2015b. However, the methods for Data Fusion are in constant evolution, and a large number of studies are constantly being added to the literature. Many recently relevant

published papers are not covered in these earlier reviews. New papers address the problem of data fusion using new concepts such as object popularity and object difficulty. In addition, machine learning techniques, which were not previously used, have been used in data fusion such as Bootstrapping or Restricted Boltzmann Machines (Xiao et al. 2016; Broelemann and Kasneci 2018).

This survey sets out to review the current literature including relevant new papers on Data Fusion of structured and semi-structured data from multiple sources of the same domain as part of data integration processes. The papers selected are classified according to the methods used to undertake Data Fusion. The comparison of Data Fusion methods is based on the following criteria: supported data types, the heterogeneity of data, and the quality of the source. By using comparative analysis, we seek to provide an understanding of the methods adopted by each proposal and how data characteristics and data sources are exploited to assist the Data Fusion process. In addition, we point out some possible limitations of each method that has been proposed. Another contribution will be to identify both the existing research challenges and open questions in the Data Fusion area with a view to indicating directions for future research.

It is hoped that this review of the literature will be helpful both to researchers looking for an overview of the Data Fusion area, as well as to those wishing to know what the current research subjects are and what lines of future research will be useful to explore.

The remainder of this paper is organized as follows. Section 2 presents an overview of Data Fusion in the context of Data Integration. Section 3 describes several Data Fusion applications. In Section 4, different classifications of Data Fusion methods are presented. In Section 5, these Data Fusion methods are compared. In Section 6, research challenges are discussed and future directions are suggested. Finally, in Section 7, some conclusions are drawn.

2 Data fusion - an overview

The purpose of Data Integration is to provide unified access to data residing in multiple, autonomous, and heterogeneous data sources. Integrating these data sources can be quite tricky because the semantics of the data in each source needs to be fully understood in order to resolve ambiguities. In the context of this work, we are mainly interested in the integration of structured and semi-structured data. The integration of unstructured data (e.g., sensor data and streaming data) is not within the scope of this survey.

Data Integration is challenging for many reasons. For example, there are various schemas and models, distinct representations for the same object, as well as an enormous redundancy of data. To adequately address these challenges, considerable efforts have been made over the years with respect to data integration researchers. Given these challenges, according to Dong and Srivastava (2015a), Data Integration traditionally has three main steps:

- **Schema alignment**: In this step, the heterogeneity at the schema level is resolved by specifying the semantic relationships between the attributes of entities from different data sources
- **Record linkage**: Instance-level heterogeneity is solved by detecting records that refer to the same real-world entity
- **Data fusion**: Data Fusion aims to resolve data conflicts from heterogeneous sources that conflict with each other, and to find the truth that reflects the real world values of

an entity. Thus, Data Fusion creates a final representation for each distinct real-world entity.

The general purpose of the above three steps is to resolve the different types of conflict that arise between distinct data sources which represent the same real-world object. The conflicts that can arise are as follows: *schematic conflicts*, which involve different structures to represent the same object; *identity conflicts*, where the same object is identified in different ways; and *data conflicts*, where semantically equivalent attributes of an object in different data sources have different values. Data conflicts can be distinguished into two kinds: i) *uncertainty* is a conflict between a not null value and one or more null values related to the same attribute of a real-world entity, and ii) *contradiction* is a conflict between two or more different not null values associated with the same attribute of an entity (Dong and Naumann 2009).

According to Dong et al. (2009a, 2015), the Data Fusion problem can be defined as follows: considering a set of data sources D, and a set of objects O, an object represents a particular aspect of a real-world entity, such as the affiliation of a researcher (in the relational database, an object corresponds to a cell in a table). For each object $o \in O$, a source $s \in S$ can (but not necessarily) provide a value. Among different values provided for an object, one of them correctly describes the real world and is *true*, and the others are *false*. Thus, the purpose of Data Fusion is to decide which value among the conflicting values is the true value for each object $o \in O$ given the values provided by the data sources $s \in S$ for this object.

How to best resolve data conflicts has been a widely studied problem in the traditional Data Integration scenario especially in the Data Fusion process (Bilke et al. 2005; Fuxman et al. 2005; Motro and Anokhin 2006; Bleiholder and Naumann 2008). Early research proposed strategies for conflict resolution which were based on different ways of dealing with conflict. These strategies are classified into three main classes based on how they handle (or do not handle) conflicting data: ignorance, avoidance, and resolution (Fuxman et al. 2005; Bleiholder and Naumann 2008). Strategies to avoid and ignore conflicts became insufficient due to some limitations (e.g., these strategies did not resolve existing data conflicts). Therefore, strategies for resolving conflicts have emerged, and new solutions are continually being proposed. Further details about these strategies are presented in Section 4.

Thus, the first methods of Data Fusion for conflict resolution were generally rule-based. Therefore, the methods of Data Fusion used conflict handling functions such as average, maximum, minimum, or voting. Among these functions, voting was highlighted, and its use in different scenarios became widespread. However, simple voting does not consider the quality of data sources. Given the increasing amount of data available on the Web, in the age of Big Data, data quality has become more critical than ever. Many false data are published, thereby introducing a veracity problem which makes simple voting flawed. Veracity directly refers to inconsistency and data quality problems (Saha and Srivastava 2014). In this scenario, conventional Data Fusion techniques are not sufficient to deal with these new challenges.

Thus, there was a need to discuss and solve the problem of data veracity in the Data Fusion process. Yin et al. (2008) were the first to formally introduce the truth discovery problem, namely, the task of finding true values for conflicting data using the reliability of the data source providers. This task has received immense research support from both the Artificial Intelligence and Database communities under various names: fact-finding (Pasternack and Roth 2011), information corroboration (Galland et al. 2010), truth-finding (Li et al. 2012), and truth discovery Liu et al. (2017b, 2019).

In our view, Data Fusion may be defined as the complete process of identifying and resolving conflicting data, thereby creating the final representation for each distinct real-world object. The truth discovery task applied in this context is seen as the main component in the general process of Data Fusion. Thus, Truth Discovery and Data Fusion have practically the same main purpose, and both are cited in many studies as synonyms. Therefore, the studies considered in this review propose both Data Fusion and Truth Discovery methods. Nowadays, methods for Data Fusion are needed in several applications. So, in what follows, applications of Data Fusion will be discussed.

3 Data fusion applications (what for?)

To illustrate that Data Fusion is a broad concept and that Data Fusion methods have been used in several domains, in this section some domain applications are briefly shown. It is important to emphasize that the scope of this survey is only Data Fusion focused on structured and semi-structured data integration of data sources from the same domain.

3.1 Sensor data fusion

Sensor Data Fusion is defined by Hall and Llinas (1997) as a method of combining sensor data from multiple sensors to produce more accurate, complete, and dependable information which it would not be possible to achieve using a single sensor.

In the IoT (Internet of Things) context, Data Fusion helps building knowledge about specific events and environments which is not possible when only individual sensors are used. In this domain, Data Fusion also enabled a context-aware model, i.e., a model that considers the situational context of an object. The situational context helps to achieve one of the ultimate objectives of sensor data fusion: understanding the environment and acting accordingly. Figure 1 illustrates the process of Data Fusion in a Smart City environment.

IoT produces a substantial amount of data that is less useful unless knowledge can be derived from them. Therefore, applying sensor data fusion techniques at the different levels of the IoT application chain is essential (Wang et al. 2016).

There are studies in the IoT domain which propose different solutions. OpenIoT (Soldatos et al. 2014) is a framework for collecting and processing data from different

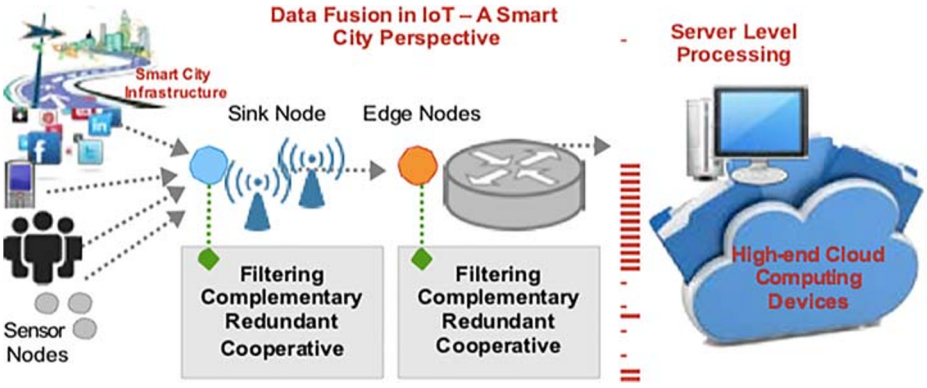

Fig. 1 Data Fusion in Smart City (Wang et al. 2016)

sources in IoT. OpenIoT includes a sensor middleware and a sensor data fusion capability in the cloud. Lau et al. (2019) e Ding et al. (2019) present a review on the state-of-the-art of data fusion in main IoT application domains.

Due to the relevance of multi-sensor data in many fields such as medical and military imaging applications and remote sensing, image fusion has become an important research area, and several lines of research are emerging. In remote sensing, the field of geographical imaging is concerned with overlapping satellite images from the same region and is interested in determining, for example, a land cover (mountain, water or woods). Image fusion in remote sensing plays a role in many essential applications, which range from detecting climate changes to managing natural disasters and preserving the environment. A general architecture of the process of integrating information from sensors is presented in (Torra and Narukawa 2007). Moreover, the authors discuss the topic and highlight methods for aggregating data. In sensor data fusion, data aggregation is considered as part of the data fusion process, whose purpose is to summarize data in order to eliminate or reduce redundancy (Chhabra and Singh 2015; Akkaya et al. 2008). In Fonseca et al. (2011), a review of fusion techniques and methods is presented, as well as a discussion of using techniques in remote sensing applications.

In Medicine, image fusion is increasingly being used for medical diagnostics, analysis, and treating patients. Moreover, the use of multi-sensor and image fusion methods offers information processing that is more robust and can reveal information that is otherwise invisible to the human eye. In order to extract more information, medical image fusion combines features of different types of images (e.g., MRI-T1 gives greater detail of anatomical structures, whereas MRI-T2 provides a higher contrast between normal and abnormal tissues) into one fused image. Medical image fusion not only helps in diagnosing diseases, but it also reduces the storage cost by reducing storage to a single fused image instead of multiple-source images (Wang and Ma 2008). A survey of medical image fusion can be found in James and Dasarathy (2014).

3.2 Linked data fusion

In the Linked Data scenario, Data Integration is also an open problem which has the same purpose, i.e., to provide a unified view of data and to simplify the creation of applications that consume Linked Data (Michelfeit et al. 2014). This process reveals different Uniform Resource Identifiers (URIs) that represent the same real-world entities while conflicting values appear due to missing data, errors, or outdated values. The Data Fusion task aims to resolve these conflicts.

In the context of Linked Data, which are represented as a Resource Description Framework (RDF), real-world objects are represented as resources. A set of RDF triples describes a resource corresponding to a "record". Conflicts are resolved, and low-quality values are purged, so as to obtain a clean representation of a resource (Michelfeit and Mynarz 2014).

According to Michelfeit et al. (2014), the issue of Linked Data Integration has its specific features and challenges, which include: i) different URIs are used to represent the same real-world entity; ii) data conflicts emerge when RDF triples which share the same subject and predicate have inconsistent values in place of the object; iii) different schemas can be used to describe data.

Several tools that implement Linked Data Fusion have come on to the market, e.g., OBResolution (Liu et al. 2017a), TruthDiscover (Liu et al. 2017b, 2019), ODCS-FusionTool (Michelfeit et al. 2014). Liu et al. (2017a) propose OBResolution to identify a true object from multiple conflicting objects using a Markov Random Field (Koller and Friedman

2009) to model all pieces of evidence under a unified framework. From the same authors, TruthDiscover (Liu et al. 2017b, 2019) is a novel system proposed to identify the truth in Linked Data with the scale-free property. Both systems leverage the topological features of the graph of the belief of a source to estimate the prior beliefs of sources, which are used to tease out the trustworthiness of sources. The Hidden Markov Random Field is used to model interdependencies among objects in order to make accurate estimates of the trustworthy values of objects. Figure 2 presents an overview of the TruthDiscover method.

In the ODCS-FusionTool (Michelfeit et al. 2014), Data Fusion is applied at query time. The system has a modular architecture in which quality computation and fusion functions are independent. It also leverages OWL to resolve schemas and identity conflicts based on mappings. The architecture of the ODCS-FusionTool is flexible, and both new resolution functions and quality assessment methods can be plugged in.

3.3 Knowledge fusion

When building a knowledge base, multiple knowledge extractors are used to extract values from various data sources. These values can be conflicting and, therefore, the degree of correctness of the extracted knowledge must be determined. This problem is called Knowledge Fusion.

The process of Knowledge Fusion consists of a succession of steps that locate and extract knowledge from multiple, heterogeneous online sources, and transform them so that a common representation can be applied as the basis for solving the problem.

Knowledge Fusion considers an additional dimension of errors (i.e. the errors made by knowledge extractors). Moreover, knowledge fusion was proposed for use on the Web, a scenario rich in resources (both memory and processing) and which has higher semantic levels of data (such as decisions) (De Oliveira Costa et al. 2018; Dong and Srivastava 2015b).

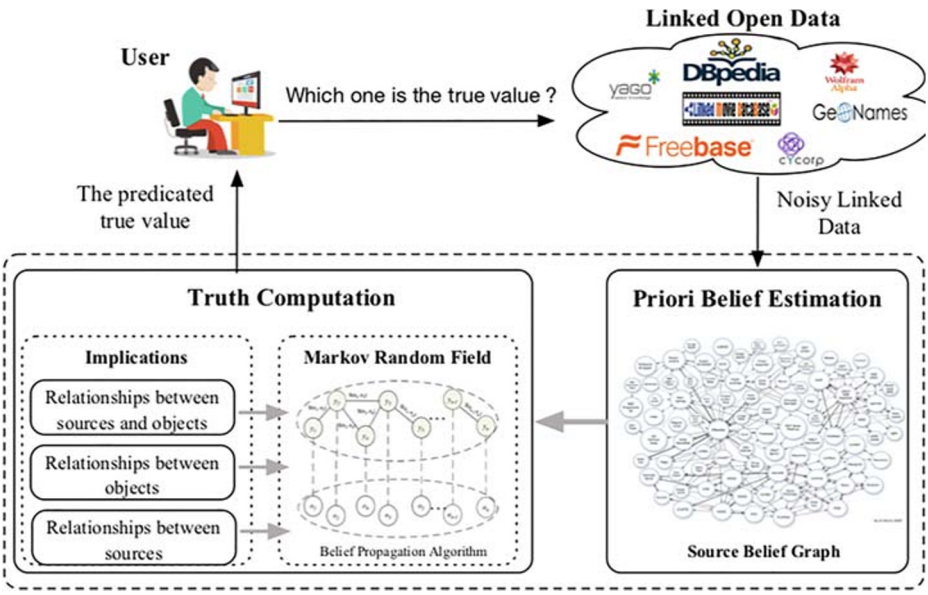

Fig. 2 Overview of the Truthdiscover method (Liu et al. 2019)

Dong et al. (2014) defines the knowledge fusion problem and presents how existing data fusion techniques can be adapted to solve it. Three existing Data Fusion methods were adapted and implemented based on MapReduce. Figure 3 shows the architecture of the system.

In De Oliveira Costa et al. (2018), the authors proposed a knowledge fusion algorithm named Athena which should support knowledge extraction. This is needed to adapt knowledge graphs for the IoT scenario. Athena applies Data Fusion techniques combined with Bayesian Decision Theory and Reinforcement Learning to enable Knowledge Fusion.

Finally, in Preece et al. (2001), the KRAFT project (Knowledge Reuse And Fusion/Transformation) aims to define a generic architecture for knowledge fusion to ease the development of knowledge fusion systems. In Dong and Srivastava (2015b), a survey on knowledge fusion is presented.

3.4 Data fusion in information retrieval

The critical point in Information Retrieval is how to rank all the documents that are retrieved. This is the task of the ranking algorithm. There may be several differences between information retrieval systems. These can cover a broad range of issues, such as different models (Boolean, Vector Space, and Probabilistic models), different treatments on many other aspects (parsing rules, phrase processing, relevance feedback techniques), documents and queries represented in distinct ways, and so on. Thus, Data Fusion is performed in order to combine the results from multiple information retrieval systems, and to obtain more effective results (Wu 2012b).

According to Bleiholder and Naumann (2008), Information Retrieval Data Fusion aims to combine the search results of different search engines into one single ranking, which is therefore also called rank merging. Data Fusion has been used successfully in many retrieval scenarios, such as meta-search and expert finding. Data Fusion models in Information Retrieval are of two types: score-based and ranked-based. Many methods have been proposed and many experiments on how to apply them in different applications scenarios have been conducted in order to evaluate them Wu et al. (2015).

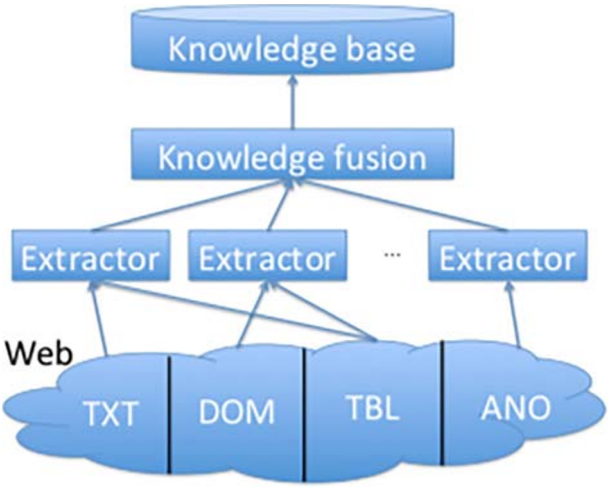

Fig. 3 Architecture of Knowledge extraction and fusion (Dong et al. 2014)

In Lillis et al. (2006), ProbFuse, a probabilistic approach to Data Fusion, is proposed. ProbFuse ranks documents based on the probability of their relevance to the given query. The probability is calculated during a training phase. The input to the fusion process are results that are produced by different information retrieval systems. For an overview of several Data Fusion methods in information retrieval see Wu (2012b).

4 Classification of data fusion methods

Data Fusion, as a part of structured and semi-structured data integration process, has been extensively studied over the years. Thus, several works have classified Data Fusion methods differently. Table 1 shows the different classifications proposed in each of them.

Initially, the Data Fusion problem has directly addressed the issue of how to integrate data from relational databases. For this reason, in Bleiholder and Naumann (2008), existing

Table 1 Classification of Data Fusion methods proposed in the literature

Published Study	Classification Proposed
Bleiholder and Naumann (2008)	– Based on conflict resolution – Based on conflict avoidance – Based on conflict ignorance
Li et al. (2012)	– Voting – Web link-based – IR-based –Bayesian-based
Dong et al. (2014)	– Voting – Relation-based – Quality-based – Web linked-based – IR-based – Bayesian – Graphical-model
Berti-Équille and Borge-Holthoefer (2015)	– Agreement-based – Maximum A Posteriori (MAP) – Analytical – Bayesian inference-based
Li et al. (2015b)	– Iterative – Optimization-based – Probabilistic graphical model-based
Classification proposed in this paper	– **Rule-based** – **Probability-based** – **Optimization-based** – **Machine Learning**

Bold entries highlight the classification of data fusion methods proposed in this paper and used in Section 5

Data Fusion systems were classified into three groups based on conflict-handling strategies. The systems initially proposed, based on conflict-handling strategies, used only conflicting data in conflict resolution, i.e. they did not consider other characteristics. Over the years, new Data Fusion methods were proposed by addressing new features and applying new techniques to solve the problem. The Data Fusion methods were classified in accordance with the characteristic exploited or the technique used. Therefore, new classifications were presented and may result in further research being conducted.

Based on existing classifications, and on our evaluation of more recent studies, we have created a slightly different rating. We divide Data Fusion methods into four categories: *Rule-based*, *Machine Learning*, *Optimization-based*, and *Probability-based*. The main difference

Table 2 Definitions of the classification categories of the Data Fusion methods

Categories	Description
Conflict resolution	Uses a variety of strategies that have different forms of implementations to resolve conflicts
Conflict avoidance	Manipulate data to avoid conflicts occurring
Conflict ignorance	Manipulate data but ignores existing conflicts
Voting/Baseline	When values from different sources conflict, one vote is given to each data source. The value which received the highest number of votes is regarded as the correct one. This is the standard strategy
Web linked	Is driven by measurement web page-based links (e.g., it uses PageRank[a])
Information Retrieval (IR)	Measures the reliability of the source and the similarity between the given values and actual values. Similarity measures such as Cosine similarity, are used and widely accepted in the area of IR
Bayesian/Bayesian inference	Uses Bayesian analysis. They rely on Bayesian modeling to calculate the accuracy of the source and the confidence of the value
Graphical-model/Probabilistic graphical model	Uses probabilistic graphical models to jointly obtain the reliability of the source and the value of correctness
Copying-affected	Discounts the votes for values copied from other sources from the total of votes initially calculated
Agreement-based	Counts the number of sources that agree/disagree with each data item
Maximum a posteriori (MAP) estimate	Calculates optimal latent variables (i.e. truth and source reliability) by means of Expectation maximization or Gibbs sampling based on available observations
Analytical	Uses matrix diagonalization in the truth discovery which is reformulated as an optimization problem
Relation-based	Also considers the relationships between sources
Iterative	Uses iteration during the steps for calculating the truth and estimating the reliability of the source until the convergence
Optimization	Considers truth discovery is an optimization problem to infer the reliability of the source and reliable information and to update truths and reliability weights from sources, iteratively to convergence. Thus methods based on optimization are similar to those iterative-based methods.
Machine learning	Uses a Machine Learning technique
Probability-based	Uses a probabilistic technique

[a]An algorithm that evaluates relevance which is used by Google to position websites in search results. For further details, see PageRank (Brin and Page 2001)

from prior classifications is that the categories of Rule-based and Machine learning have been included. In the Rule-based category, we include the methods first proposed for Data Fusion, as well as those in which the base strategy is voting. What prompted our inclusion of a Machine Learning category was the sharp increase in the number of data fusion methods proposed that use Machine Learning techniques. Another difference is the probabilistic-based category which is more general because it covers studies that use probability to perform Data Fusion, regardless of the technique used (e.g., Markov, Bayesian).

Each category that appears in the classifications presented in Table 1 is described in Table 2. The categories that appear in more than one classification (Table 1) with different names are all listed together in Table 2.

5 Data fusion methods (in what form?)

In the previous section, we have summarized several classifications of Data Fusion methods that appear in the literature. In this section, we will briefly describe several Data Fusion methods which will be classified according to the categorization proposed in this paper.

Conflict resolution was first mentioned in the literature in the relational database integration area. It was about the problem of an attribute of the same entity that has different values in different data sources. However, this was not given much importance at the time, since most of the proposed techniques used strategies to avoid conflicts or merely ignored them Bleiholder and Naumann (2008).

More recently, due to the ease of publishing and sharing data, many data sources provide incorrect information and do not have the desired reliability. Therefore, the topic of truth discovery has gained increasing attention. The first study to formally introduce the truth discovery problem was (Yin et al. 2008).

Nowadays, many data fusion methods have been proposed for resolving conflicts by discovering the truth (e.g., Zhang et al. (2018), Broelemann et al. (2017), Broelemann et al. (2018), Li et al. (2017), Xie et al. (2017), Wang et al. (2017), Zhang et al. (2016), Fang et al. (2017a), and Nakhaei and Ahmadi (2017)). These methods, while using different approaches to solve the task of Data Fusion, apply the same principle: the higher the quality of the source is, the more likely it is that it provides truth and, the more truth it gives, the higher the quality of the source is.

In this paper, the intention is to discuss the most recent methods. However, we will briefly cite some popular methods commonly used by researchers. The main studies on Data Fusion, proposed over the years, are listed in Fig. 4 (represented by squares). In this figure, the surveys on Data Fusion are represented by ellipses. Although each paper is classified by the primary strategy used, in only one category, some papers can be related to more than one category. The methods included in this paper were classified and are also listed in Fig. 4. In this paper, methods that have been proposed more recently (from the year 2016 onwards) will be discussed, none of which were included in the study in Li et al. (2015b) (the most recent survey on Data Fusion).

We now describe and compare each data fusion method briefly in accordance with the classification proposed in this paper. For further information about these studies, please consult the papers we have cited.

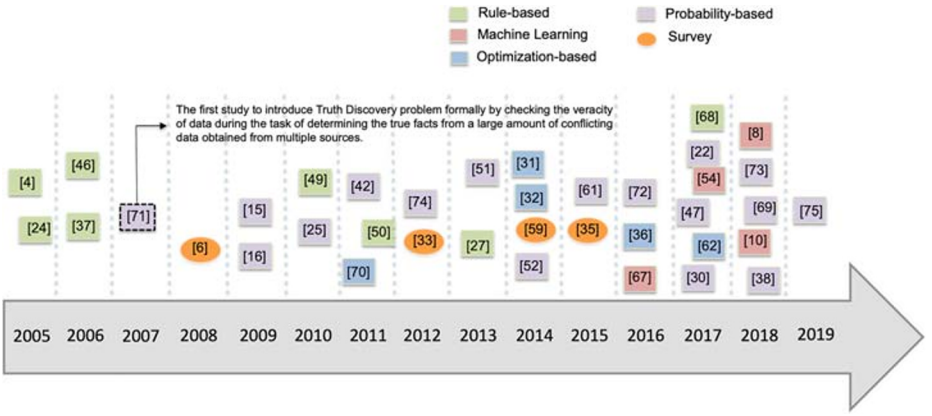

Fig. 4 The most important Data Fusion studies published since 2005

5.1 Rule-based methods

The first methods to be proposed in the literature to resolve data conflicts were rule-based methods. Generally, these methods use the mean or median (for numerical data) or a majority vote (for categorical data) to predict true values. The advantage of these approaches is that the result is generally easier to debug and to understand. Approaches have greatly evolved over time, and few recent papers are based on rules for discovering true values. **DTQ** (Xie et al. 2017), the rule-based method proposed to find true values for multi-valued attributes, is discussed below.

DTQ (Xie et al. 2017) This method is proposed with a goal to finding multiple true values for multi-valued attributes. The authors introduce the concept of quality predicates in order to differentiate true values from false ones. Quality predicates can be of three types, namely: priority predicates, status predicates, and interaction predicates. These quality predicates are in a simple form and can be found automatically by existing methods. An algorithm is proposed to infer quality vectors for each tuple, based on its quality predicates. A quality vector is an n-ary vector representing the quality of a tuple, where the real number in each dimension represents the quality of the corresponding attribute value in the tuple. True values are found based on the voting approach. For attributes labeled as multi-valued, all attributes with non-negative values in quality vectors are returned, and for attributes labeled as time-sensitive, the attribute with the highest value in quality vectors is replaced.

5.2 Probability-based methods

Probability-based approaches use probabilistic models to jointly calculate the reliability of the source and the correctness of the values. Several papers have been published that use this approach, the most prominent of which are briefly presented below.

IATD (Zhang et al. 2016) Influence-Aware Truth Discovery (IATD) is an unsupervised probabilistic method based on the Bayesian model. The authors believe that claims made by one source may be influenced by others. To model influences between sources, IATD introduces the concept of "claim trustworthiness", which fuses the trustworthiness of the

source which provides the claim, and the trustworthiness of its influencers by a parameter ensemble. By taking the source correlations as prior knowledge for deriving influence, the trustworthiness of a source can be estimated more accurately. The IATD model is divided into two stages: in the first stage, the generation of the individual trustworthiness of each source is specified as is the fusion of the influence-aware trustworthiness, given the set for each claim. Then, the second stage seeks to generate heterogeneous claims, given the "claim trustworthiness" of each claim. The model can handle both numerical and categorical types of data that are modeled in a unified manner.

FTS (Zhang et al. 2018) This method proposes the use of Silent Rate, True Rate, and False Rate to measure the quality of the source. Compared with state-of-the-art which does not consider how the source quality is affected when the source provides null, this model makes full use of all claims and null to improve the accuracy of truth discovery. However, for the truth discovery task, the authors make use of the Hub Authority method (Yu et al. 2014; Galland et al. 2010) and redesign metrics for source quality, measured by three indexes: true rate, false rate, and silent rate. The silent rate is the differential because it uses the null data provided by the source to measure source quality.

SmartMTD (Fang 2017; Fang et al. 2017b) The graph-based approach of multi-valued Truth Discovery uses Markov chain models with Bayesian inference. This approach incorporates four implications, namely: source relations, object popularity, loose mutual exclusion and the long tail phenomenon on source coverage. An overview of the framework is shown in Fig. 5. For source relations, modeling two-sided relations between sources is proposed. Supportive agreement graphs are used to capture source authority features and two-sided source precision (i.e., positive precision and negative precision), while malicious agreement graphs are used to quantify the copying relation among sources.

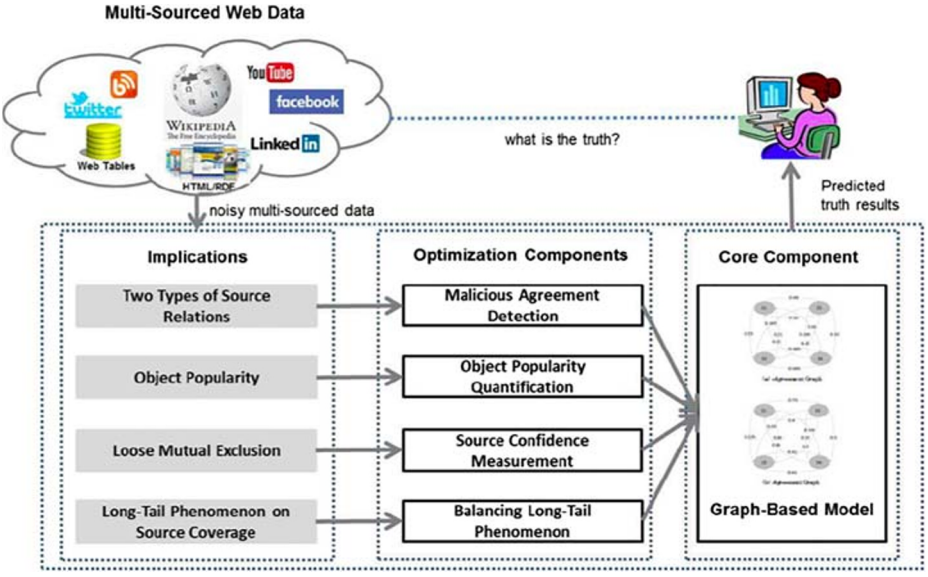

Fig. 5 The framework of SmartMTD (Fang 2017)

Object popularity is quantified by the frequency of its occurrence in the claims of sources, i.e., objects covered by more sources are usually more popular than those covered by fewer sources. With regards to loose mutual exclusion, for a multi-valued object, the mutual exclusion among values is not as strict as that of the single-valued object, for many reasons. The sources may provide partial true values, omit the values they are not sure about, or audaciously provide all potential values, even if the veracity of the values claimed is uncertain. SmartMTD applies source confidence scores to differentiate the extent to which a source believes its positive claims and negative claims. The balancing long-tail phenomenon component calculates the compensation of long-tail phenomenon on source coverage for each link in the ±supportive agreement graphs to avoid the reliability of small sources from being over- or under- estimated.

HLCR (Nakhaei and Ahmadi 2017) High-Level Conflict Resolution is based on graphical model, which performs high-level data fusion and uses relationships between objects for inferring the truth value. In short, the approach has two main goals: to find the relationship between objects and to estimate true values by using the relationship between objects. For the former, the authors introduce the concepts of attribute identifier – elements that describe attributes much more fully and are extracted from the attribute itself or others (i.e., keywords that can be extracted from the book title attribute). Two objects that have more than the specific threshold common identifiers are potentially related. For the latter goal, an undirected Conflict Resolution Graph (CRG) was introduced, and each node is an attribute object pair, and attribute identifiers estimate relationship between objects.

DART (Lin and Chen 2018) The authors propose an integrated Bayesian approach to incorporate the domain expertise of data sources and confidence scores of value sets, with the aim of finding multiple possible truths without any supervision. They believe that source reliability usually varies among different domains, and it is better to consider domains separately in the truth-finding model. Thus, they propose an integrated Bayesian approach which comprehensively incorporates the domain expertise of the data source and the confidence score of the value, to infer multiple possible truths of a data item. They also investigate the mutual influence between domains, which will affect the inference of domain expertise.

PTDCorr (Yang et al. 2018) What this paper proposes is a chain graph-based framework, Probabilistic Truth Discovery with Object Correlations (PTDCorr), in which source reliabilities, the claims of sources, and object truths are modeled as random variables. Based on PTDCorr, an incremental iPTDCorr algorithm was developed. The algorithm works efficiently in a dynamic environment. iPTDCorr is able to incorporate time-invariant correlations between different objects as well as temporal correlations for the same object and therefore can effectively infer object truths. Thus, a temporal correlation is considered for truth inference. The method is probability-based and takes into account source reliability and object correlations to find the truths.

GTFC (Zheng et al. 2019) The GTFC (Gaussian Truth Finder with Source Correlations) method is proposed in this paper. The source quality is modeled by considering accuracy and recall. In addition to the quality of sources, the correlations between sources can also affect the precision of the truth discovery algorithm. To measure the similarity between sources based on the similarity of their claims, the algorithm uses the existing DHNE (Deep Hyper-Network Embedding) model to learn the vector representation of each source, and

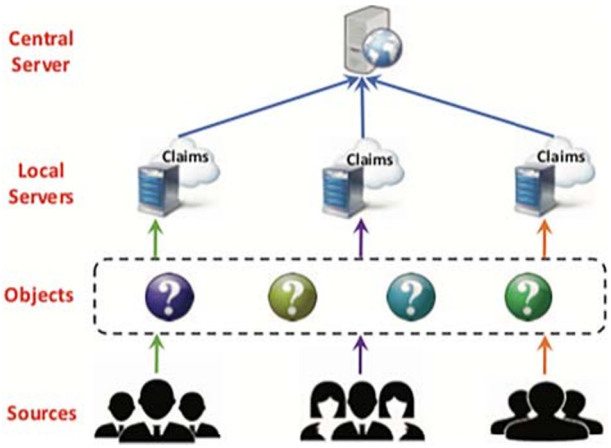

Fig. 6 The scenario of DTD (Wang et al. 2017)

then it calculates the similarity between sources according to the embedding vector. So, the duplicate data sources are removed from the process.

5.3 Optimization-based methods

Optimization-based approaches model the truth discovery as an optimization problem. Source reliability and truth discovery are modeled iteratively until convergence. Some recent papers that use the optimization-based approach are briefly discussed below.

DTD (Wang et al. 2017) A Distributed Truth Discovery (DTD) Framework is proposed for a new scenario of truth discovery, namely distributed data. In short, the information about the observed objects provided by different sources is usually distributed across a group of local servers, as shown in Fig. 6. DTD entails estimating the uncertainty inferred from the claims of each object, and consists of two components: Local truth computation which estimates the local truths and variances of objects in each local server, and Central truth estimation which is used to infer the final truths in the central server from the outputs from all the local servers. An approach called UbTD (Uncertainty-based Batch Truth Discovery) is proposed in the local truth computation step, in order to model the differences among objects as the uncertainty values that are used for estimating the truths in local servers. The central truth estimation step aims to infer the final truths of the objects by considering the quality of local servers and to infer the estimated truths as well as object variances uploaded by the local servers.

CRH (Li et al. 2016; Li et al. 2014b) The Conflict Resolution on Heterogeneous Data (CRH) framework aims to infer the truths from multiple conflicting sources, each of which involves a variety of data types. The truth discovery problem is modeled as an optimization problem that can handle heterogeneous data. The proposed algorithm solves the optimization problem by iteratively updating truths and source weights. Another version of the CRH framework is proposed for working incrementally in a streaming data scenario.

5.4 Machine learning methods

Machine Learning-based approaches use machine learning techniques in the truth-finding process. The number of studies using machine learning is growing, and some of which are discussed below.

LTD-RBM (Broelemann et al. 2017; 2018) The proposed Latent Truth Discovery (LTD) approach based on Restricted Boltzmann Machines (RBM) provides a practical inference procedure based on Contrastive Divergence and Gibbs sampling. The algorithm uses RBM for inferring the true facts and source reliabilities regarding true positive rate and false positive rate. The reason for using RBMs for latent truth discovery is their ability to learn hidden factors. Given that the sources must provide correct claims, the authors believe that the main hidden factor behind the claims of all sources is the unknown truth, which they try to discover.

GRBM (Broelemann and Kasneci 2018) This study proposes an extension of (Broelemann et al. 2018; 2017), and, unlike most methods, it incorporates arbitrary features to solve the latent truth discovery problem. An extension of the LTD-RBM is proposed that makes use of the differentiable reliability function, such as those represented by feed-forward neural networks. The reliability function enables the reliability of similar sources to be determined in a combined way and thus addresses the long-tail problem. This approach uses unsupervised training of feed-forward networks, with the contrastive divergence of the RBM on top of back-propagation in feed-forward networks, but the pre-training is conducted in a supervised way.

ETCIBoot (Xiao et al. 2016) The Estimating Truth and Confidence Interval via the Bootstrapping method is proposed to automatically construct confidence interval estimates as well as to identify the truth of objects. Most truth discovery methods focus on providing a point estimator for the truth of each object, but in many real-world applications, estimating the confidence interval of truth is more desirable. For example, two objects A and B receive the same truth estimate, e.g., 25, but the confidence in this estimate could differ significantly – A may receive 1000 claims around 25 while B only receives one claim of 25 and the confidence in the truth estimate for A is much higher. An estimated confidence interval of truth can benefit any truth discovery scenario by providing additional information in the output, with greater advantage in long-tail scenarios. ETCIBoot consists of the following three steps: Weight Update, Truth Estimation, and Confidence Interval Construction. In the first, given initialization of truths, source weights are updated. In the Truth Estimation step, for each object, the truth estimators are obtained. In the final step, for all objects, the estimation of confidence intervals for their truths are obtained.

SLiMFast (Rekatsinas et al. 2017) This paper proposes a framework that expresses Data Fusion as a statistical learning problem over discriminative probabilistic models to perform Data Fusion. An overview of the SLIMFast is shown in Fig. 7. Its main components are Compilation, Optimizer, and Data Fusion. There are two main tasks in a Data Fusion module: performing statistical learning to compute the parameters of the graphical model which are used to estimate the accuracy of data sources, and performing probabilistic inference to predict the true values of objects. SLiMFast is the first Data Fusion approach to combine

cross-source conflicts with domain-specific features as an additional signal to estimate the accuracy of sources.

5.5 Comparison of data fusion methods

We compared the methods under the various features described below. Figure 8 summarizes the studies according to the characteristics that are compared.

- **Data types -** There are several different data types. The most basic definition of data type is whether the data is continuous (quantitative) or categorical. This is used by most Data Fusion methods. Exceptions are Yang et al. (2018) and Zheng et al. (2019) which deal only with continuous data, while (Fang 2017; Zhang et al. 2018; Fang et al. 2017b; Nakhaei and Ahmadi 2017; Lin and Chen 2018) deal only with categorical data, but can be extended to encompass continuous data.
- **Heterogeneity -** Heterogeneous data types can be used to describe a real-world object. It is essential that when a data fusion method is being applied, a joint Data Fusion on all types of data is conducted simultaneously. Some approaches deal with data heterogeneity e.g., Zhang et al. (2016), Xie et al. (2017), Li et al. (2016), Xiao et al. (2016), Wang et al. (2017), and Broelemann and Kasneci (2018) all deal with more than one data type at a time.
- **Multi-truths -** Most Data Fusion approaches assume that each attribute of an object has only one truth. This assumption is formally defined as a "single-truth". However, in the real world, many attributes may have multiple true values (i.e., multiple-truths). For example, a person can have more than one phone number, and all of them can be true. In Fang et al. (2017b), Fang (2017), Xie et al. (2017), Zheng et al. (2019), and Lin and Chen (2018) the multi-truth discovery problem is discussed, and approaches to solve the problem are proposed.
- **Source quality -** In truth discovery methods, estimating the quality of sources is added as a step. This step aims to evaluate the trustworthiness of each data source according to the correctness of the values it provides. Truth-discovery methods measure the quality of a source by using metrics such as accuracy, exactness, freshness, reputation, precision, and recall. All methods discussed in this paper use source quality in Data Fusion.
- **Copying between sources -** Many truth discovery methods assume that data sources are independent. However, in the real world, copying exists between data sources. There are two types of data source: independent sources, that provide all values independently, and copier sources, which copy part (or all) of data from other sources. Detecting copying between data sources can help in the process of discovering the truth in such a way that a discounted vote count can be assigned to a copied value in voting. In Fang et al.

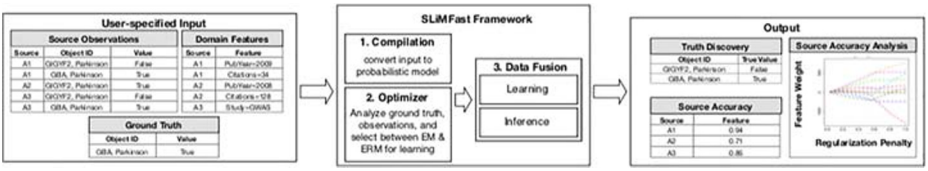

Fig. 7 An overview of SLiMFast (Rekatsinas et al. 2017)

(2017b), Fang (2017), and Zheng et al. (2019), the copy relation among sources, and the common false values shared by two sources are defined by a malicious agreement and used in the truth discovery task.

– **Object relationship -** Studies usually assume that objects are independent of each other. However, objects can have relationships. For example, the "department" in which a person studies has a strong relation with the "university" of this person, or with his/her "date of birth" and "age". In Nakhaei and Ahmadi (2017), the authors claim that two objects are related if they have some identifiers in common. Two objects that have more than the specific threshold common identifiers are potentially related. So, when recognizing such relations, it is possible to select a consonant value for attributes of correlated entities.
– **Object popularity -** Popular objects tend to be covered by more sources, as sources tend to publish popular information to attract more audiences. Therefore, objects covered by a large number of sources are usually more popular than those covered by fewer sources (Fang et al. 2017b). For example, the phone number of a restaurant is more popular and has a more significant impact than the year when it was opened because customers need to contact the restaurant. Since the size of the potential audience is generally more significant for popular objects, more people will be misled if a popular object has a false value than when a less popular object has a false value. The concept of object popularity appears in Fang et al. (2017b) and Fang (2017). These studies propose using it to determine the extent of source reliability by differentiating the popularity of objects, and therefore to minimize the number of people potentially misled by false values. Thus, sources that provide false values for popular objects can be penalized more heavily and this can be done by assigning a higher negative weight to their measurement of quality.
– **Object difficulty -** Commonly, objects are treated equally, and the characteristics of each object are not considered. However, objects have different levels of difficulty. In other words, for some objects, it is easier to find the truth than it is for others. In Wang et al. (2017), the difficulty of objects is modeled as uncertainty and considers two aspects: the inner factor and the outer factor. First, if the object is extremely hard (i.e., if it is challenging to infer the real true information of the object), then the object

Method	Year	Data type	C1	C2	C3	C4	C5	C6	C7
IATD [72]	2016	Categorical/Continuous	X		X				
CRH [32, 36]	2016	Categorical/Continuous	X		X				
ETCIBoot [67]	2016	Categorical/Continuous	X		X				
SmartMTD [19, 22]	2017	Categorical		X	X	X		X	
DTQ [68]	2017	Categorical/Continuous	X	X	X				
DTD [62]	2017	Categorical/Continuous	X		X				X
HLCR [47]	2017	Categorical			X		X		
LTD-RBM [9, 10]	2017	Categorical/Continuous			X				
SLIMFast [54]	2017	Categorical/Continuous			X				
FTS [73]	2018	Categorical			X				
GRBM [8]	2018	Categorical/Continuous	X		X				
PTDCorr [69]	2018	Continuous			X		X		
DART [38]	2018	Categorical		X	X				
GTFC [75]	2019	Continuous		X	X	X			

C1 – Heterogeneity
C2 – Multi-truths
C3 – Source
C4 – Copying between sources
C5 – Object relation
C6 – Object popularity
C7 – Object difficulty

Fig. 8 Comparison of Data Fusion methods

has a high uncertainty value. Second, if few sources provide claims on the object, then, since there is insufficient data, estimating the truth becomes more difficult, and so the object also has a high uncertainty value.

5.6 Some other studies

In this section, we discuss some studies found in the current literature that were not included in our comparative study. These papers cannot be compared to each other or to other studies due to one of the following reasons: i) the paper does not cite a specific Data Fusion method; e.g. papers that propose solutions for Data Integration, where Data Fusion is part of the process; and papers that suggest the combination of several existing Data Fusion methods that have been published in the literature, but the method has not been classified as a new original Data Fusion method, i.e., ensemble approaches; ii) the paper does not provide sufficient information about the proposed method. Even though none of them contains detailed information on the features of Data Fusion they use, these papers are included in this survey as they introduce new and exciting solutions which address Data Fusion challenges.

In Li et al. (2017), the authors propose a model, namely HYBRID, which works for multi-truth applications. HYBRID makes two decisions: it states how many truths there are for the attribute of an entity, and what they are. On the condition that there is a sequence of true values that have been selected previously, HYBRID computes the probability of a value being the next truth and the probability that there is no more truth values, based on a Bayesian model. HYBRID also works for entities with a single-truth values because it can automatically decide on the number of truths.

An ensemble approach for Truth Discovery is proposed in Fang et al. (2016), who analyzes the feasibility of the ensemble truth discovery approach, and formally defines the ensemble truth discovery problem. The authors propose to fully leverage the advantages of existing methods by extracting truth from the prediction results of these existing truth discovery methods. To implement the approach, two models are proposed: a serial and a parallel model. The parallel model unifies the format of truth discovery methods and combines their outputs into existing ensemble methods. In the serial model, the output of a method is used as if it were the input of another method for initializing prior methods. The input of the ensemble truth discovery problem is a three-dimensional data matrix, where the third dimension represents different truth discovery methods.

In Ahmed and Sadri (2018), incorporating the correctness (or confidence) measure of facts besides the accuracy of sources is proposed. The authors argue that there are several advantages to incorporating these measures, especially that of greater accuracy in the results of Data Fusion. Another contribution of this study is that it is an approach to determine the correctness threshold based on users' assessment of the correct facts. In Data Fusion, the correctness threshold is a variable that contains a probability value, since facts with probabilities equal to or higher than this value are considered true.

A novel graph-based Data Integration Framework based on the Unified Concept Model is proposed in Ma et al. (2017). The framework has three main components: UCM generation, automatic instance graph transformation, and graph analysis and visualization. UCM generation consists of extracting the schemas of entities and relationships and translates them into the Unified Concept Model (UCM), which describes a global combined schema. In UCM, concepts are represented as nodes and relationships as edges. The automatic instance graph transformation step aims to transform original data stored in heterogeneous sources that can be automatically transformed into graph instance data. In this step, data cleaning,

record linkage, and Data Fusion are performed automatically. For the graph database system, the Neo4j[1] was chosen, so that users can visually navigate through the graphs and also query them.

6 Research challenges and future directions (what is next?)

Several studies have been conducted on Data Fusion and truth discovery. Nevertheless, there are still many open issues to explore. In this section, we discuss some research challenges and future directions for Data Fusion which were identified in the analysis of the papers published and included in our survey.

6.1 Scalability

The characteristics of Big Data, such as volume, and velocity, make the Data Fusion problem even more complex. The need to access and analyze large-scale data sets efficiently is a challenge that traditional methods of Data Fusion do not consider. Most of the solutions do not scale up, because they usually demand human assistance, which is difficult to provide in an efficient manner, since the problems are large scale. Thus, distributed and scalable versions of existing algorithms are required.

In Waguih and Berti-Équille (2014), several truth-discovery algorithms were experimentally evaluated by using real-world and synthetic datasets with different configurations. The authors concluded that most algorithms have efficiency problems and are computationally expensive to apply to large scale problems.

Examining each component of Data Fusion sequentially for billions of data items and data sources can be prohibitively expensive. A natural thought is to parallelize the computation in a MapReduce-based framework. The complexity of truth discovery and estimating the reliability of sources are linear to the number of data items and data sources. Therefore, a structure based on MapReduce is useful for scaling them up Dong and Srivastava (2015a).

Some studies examine the Data Fusion scalability problem. In Dong et al. (2014), a framework for offline data fusion based on MapReduce is proposed. It covers the discovery of truth and the estimation of the reliability of the sources. However, the detection of copies between data sources has a quadratic complexity since this is done for each pair of sources, which the framework is not designed to do. Li et al. (2015a) studied the problem of how to improve the scalability of copy detection and proposed various methods for improving the efficiency and scalability of doing so by using structured data.

In Liu et al. (2011), the authors propose to fuse data from different sources at query answering time, which makes the process more efficient, since this can deal with large volumes of data, large amounts of sources, and high refresh frequency scenarios. They assumed that checking the accuracy of the sources and copy detection were done offline, while truth discovery is conducted at query answering time.

Despite these initial efforts, much research still needs to be done to solve the problem. Making the current Fusion algorithms scalable is not an easy task because of the complex parameter settings and the assumptions that must be taken into consideration. The issue of scalability in Data Fusion methods is an essential and extremely promising research topic.

[1]Neo4j - https://neo4j.com

6.2 Initializing parameters

Most truth discovery methods need to initialize input parameters. According to Waguih and Berti-Équille (2014), parameter settings can dramatically impact the quality of the truth discovery algorithms. However, many parameters, such as reliability of the source, are not known *a priori*.

Currently, most proposed methods start from the default reliability for each source and then iteratively refine this reliability while also leading to truth discovery. Therefore, the trustworthiness computed may not be precise, and it appears that knowing precise trustworthiness from the start can fix nearly half of the errors in the best fusion results (Li et al. 2012). So, can we start with some seed trustworthiness that is better than the default values currently used to improve fusion results?

From the example of the reliability parameter of the sources, note that this parameter initialization must be performed efficiently, with a view to obtaining quality results at the end of the process. How to automate parameter initialization is still an open problem.

6.3 How to combine different methods and results

Many methods are proposed to solve the data fusion problem. However, no single method outperforms the others in all scenarios, i.e., no generic solution can be applied to all scenarios. Many assumptions are made by the methods proposed, which greatly complicates generalization.

Methods assume, for example, that each object attribute has only one truth value, that the data is categorical, that the sources are independent of each other, and that the data is sufficient to evaluate the reliability of the sources. In general, each method is applied only to a few specific scenarios.

Can we combine several methods, where each method is chosen according to input data parameters? Put in another way, can we combine the results of several methods? Would the result be better? Going farther than this, could a general method be proposed that can be configured according to different scenarios?

6.4 Object relationships

Most Data Fusion methods depend on the reliability of the data sources, which, in turn, depends directly on the coverage (i.e., the greater the coverage, the greater the reliability). In scenarios where most sources are unreliable (i.e., pessimistic scenarios), or in scenarios where most sources have low coverage (i.e., long-tail phenomenon), it is ineffective to use only source reliability. Mainly for these scenarios, exploring relationships between objects can be useful.

Few papers explore the relationships among objects, and most assume that each object is independent from all others. However, in real scenarios, objects can be related, such as someone's age and date of birth. These relationships need to be further explored in order to aid the Data Fusion process.

In Pasternack and Roth (2010), a framework was proposed for incorporating prior knowledge into any truth discovery algorithm, which expresses both general "common sense" reasoning and specific facts already known to the user as first-order logic, and translates this into a linear program. Thus, prior knowledge can be added as relationships between objects or attributes.

In Nakhaei and Ahmadi (2017), a Data Fusion method at a higher level of abstraction (i.e., high-level fusion) is proposed, in which the relationships between objects are assessed. The basic idea behind this paper is that the information existing in relationships between objects can help to resolve conflicts and lead to truth discovery.

6.5 Variety

The problem of the variety of data in the Big Data era goes far beyond dealing with the heterogeneity of continuous and categorical data. Some papers propose methods of Data Fusion that can deal with this heterogeneity of data. However, with regards to the amount of data available on the web, much of it is not structured, and the main and growing need is to extract value from this data, mainly in the Big Data era. To this purpose, Data Fusion is crucial.

Most of the proposed Data Fusion methods deal with structured data, but not with semi-structured and unstructured data. It is extremely important to develop methods that are able to deal with semi-structured and unstructured data. This is a big open problem that needs much research and is far from being solved.

7 Conclusions

Data Fusion is not a new field, as we have emphasized in this paper. Over the years, researchers have proposed solutions to the problem, mainly by addressing the fusion of structured data. However, the vast amount of data that is being made available on the Web has led to Data Fusion facing new challenges. Among the main challenges, the variety of data (e.g., structured, semi-structured and unstructured), the large amount of false data and the enormous variation in the quality of data sources can be mentioned. Several types of research have been conducted to address these new scenarios.

In this paper, we have reviewed research studies on the state of the art of Data Fusion, including those related to Truth Discovery. We have also discussed the application areas of Data Fusion and have summarized objectives and research in the fields. We have presented classifications for Data Fusion methods found in the literature and have made a detailed comparison of the methods utilized, according to their characteristics. Finally, we have briefly summarized each method and its main features.

The process of automatic Data Fusion in heterogeneous and large-scale data is still new. Therefore, there is still no single and complete solution for all scenarios. Many challenges thrown down by the era of large volumes of data have been added to the traditional problem of Data Fusion. Some open issues have been mentioned, such as the need to evaluate the feasibility of scaling existing methods or to propose new scalable ways to handle large-scale data efficiently.

It is hoped that this review of the literature will prove to be both a useful guide to researchers unfamiliar with this area and as a reference for the more experienced to remind themselves of how the various fields have developed and to update themselves on recent contributions.

Acknowledgements This research was partially funded by INES 2.0, FACEPE grants APQ-0399-1.03/17 and APQ-0399-1.03/17, CAPES grant 88887.136410/2017-00, and CNPq grant 465614/2014-0.

Compliance with Ethical Standards

Conflict of interests The authors declare that they have no conflict of interest.

References

Ahmed, A.H., & Sadri, F. (2018). Datafusion: taking source confidences into account. In *ICIST, ACM, New York, NY, USA* (pp. 9:1–9:6), https://doi.org/10.1145/3200842.3200854.

Akkaya, K., Demirbas, M., Aygün, R.S. (2008). The impact of data aggregation on the performance of wireless sensor networks. *Wireless Communications and Mobile Computing*, *8*(2), 171–193.

Berti-Équille, L. (2015). Data veracity estimation with ensembling truth discovery methods. In *BigData, IEEE* (pp. 2628–2636).

Berti-Équille, L., & Borge-Holthoefer, J. (2015). *Veracity of data: from truth discovery computation algorithms to models of misinformation dynamics. synthesis lectures on data management*. New York: Morgan & Claypool Publishers.

Bilke, A., Bleiholder, J., Böhm, C., Draba, K., Naumann, F., Weis, M. (2005). Automatic data fusion with HumMer. In *VLDB, demo abstract band.* http://www.informatik.hu-berlin.de/mac/publications/VLDB2005.pdf.

Bleiholder, J. (2010). Data fusion and conflict resolution in integrated information systems. PhD thesis, Uni Potsdam.

Bleiholder, J., & Naumann, F. (2008). Data fusion. *ACM Computational Surveys*, *41*(1), 1–41. https://doi.org/10.1145/1456650.1456651.

Brin, S., & Page, L. (2001). The anatomy of a Large-Scale hypertextual web search engine. In *Proceedings of the seventh international world-wide web conference*.

Broelemann, K., & Kasneci, G. (2018). Combining restricted boltzmann machines with neural networks for latent truth discovery. arXiv:1807-10680.

Broelemann, K., Gottron, T., Kasneci, G. (2017). Ltd-rbm: Robust and fast latent truth discovery using restricted boltzmann machines. In *ICDE, IEEE computer society* (pp. 143–146).

Broelemann, K., Gottron, T., Kasneci, G. (2018). Restricted boltzmann machines for robust and fast latent truth discovery. arXiv:1801.00283.

Chhabra, S., & Singh, D. (2015). Article: data fusion and data aggregation/summarization techniques in wsns: a review. *International Journal of Computer Applications*, *121*(19), 21–30. full text available.

De Oliveira Costa, G.M., de Farias, C.M., Pirmez, L. (2018). Athena: a knowledge fusion algorithm for the internet of things. In *Q2SWinet, ACM* (pp. 92–99). http://dblp.uni-trier.de/db/conf/mswim/q2swinet2018.html#MartinsFP18.

Ding, W., Jing, X., Yan, Z., Yang, L.T. (2019). A survey on data fusion in internet of things: towards secure and privacy-preserving fusion. *Information Fusion*, *51*, 129–144.

Dong, X.L., & Naumann, F. (2009). Data fusion - resolving data conflicts for integration. *PVLDB*, *2*(2), 1654–1655. https://dblp.uni-trier.de/db/journals/pvldb/pvldb2.html.

Dong, X.L., & Srivastava, D. (2015a). *Big data integration. synthesis lectures on data management*. New York: Morgan & Claypool Publishers.

Dong, X.L., & Srivastava, D. (2015b). Knowledge curation and knowledge fusion: challenges, models and applications. In *SIGMOD conference, ACM* (pp. 2063–2066). http://dblp.uni-trier.de/db/conf/sigmod/sigmod2015.html#DongS15.

Dong, X.L., Berti-Équille, L., Srivastava, D. (2009a). Integrating conflicting data: The role of source dependence. *PVLDB*, *2*(1), 550–561. http://dblp.uni-trier.de/db/journals/pvldb/pvldb2.html#DongBS09.

Dong, X.L., Berti-Équille, L., Srivastava, D. (2009b). Truth discovery and copying detection in a dynamic world. *PVLDB*, *2*(1), 562–573. http://dblp.uni-trier.de/db/journals/pvldb/pvldb2.html#DongBS09a.

Dong, X.L., Gabrilovich, E., Heitz, G., Horn, W., Murphy, K., Sun, S., Zhang, W. (2014). From data fusion to knowledge fusion. *PVLDB*, *7*(10), 881–892.

Dong, X.L., Berti-Équille, L., Srivastava, D. (2015). Data fusion: resolving conflicts from multiple sources. arXiv:1503.00310.

Fang, X.S. (2017). Truth discovery from conflicting multi-valued objects. In *WWW (Companion Volume), ACM* (pp. 711–715). http://dblp.uni-trier.de/db/conf/www/www2017c.html#Fang17.

Fang, X.S., Sheng, Q.Z., Wang, X. (2016). An ensemble approach for better truth discovery. In *ADMA, lecture notes in computer science*, (Vol. 10086 pp. 298–311). http://dblp.uni-trier.de/db/conf/adma/adma2016.html#FangSW16.

Fang, X.S., Sheng, Q.Z., Wang, X., Barhamgi, M., Yao, L., Ngu, A.H.H. (2017a). Sourcevote: fusing multi-valued data via inter-source agreements. In *ER, Springer, lecture notes in computer science*, (Vol. 10650 pp. 164–172). http://dblp.uni-trier.de/db/conf/er/er2017.html#FangSWBYN17.

Fang, X.S., Sheng, Q.Z., Wang, X., Ngu, A.H.H. (2017b). Smartmtd: a graph-based approach for effective multi-truth discovery. arXiv:1708.02018.

Fonseca, L., Namikawa, L., Castejon, E., Carvalho, L., Pinho, C., Pagamisse, A. (2011). Image fusion for remote sensing applications. In *Image fusion and its applications, IntechOpen, Rijeka, chap 9 https://doi.org/10.5772/22899*.

Fuxman, A., Fazli, E., Miller, R.J. (2005). Conquer: efficient management of inconsistent databases. In *ACM SIGMOD international conference on management of data, ACM, New York, NY, USA* (pp. 155–166). https://doi.org/10.1145/1066157.1066176. http://www.cs.toronto.edu/afuxman/publications/sigmod05.pdf.

Galland, A., Abiteboul, S., Marian, A., Senellart, P. (2010). Corroborating information from disagreeing views. In *WSDM, ACM* (pp. 131–140). http://dblp.uni-trier.de/db/conf/wsdm/wsdm2010.html#GallandAMS10.

Hall, D., & Llinas, J. (1997). An introduction to multisensor data fusion. *Proceedings of the IEEE*, *85*(1), 6–23.

Hara, C.S., de Aguiar Ciferri, C.D., Ciferri, R.R. (2013). Incremental data fusion based on provenance information. In *In Search of elegance in the theory and practice of computation, Springer, Lecture Notes in Computer Science*, (Vol. 8000 pp. 339–365). http://dblp.uni-trier.de/db/conf/birthday/buneman2013.html#HaraCC13.

James, A.P., & Dasarathy, B.V. (2014). Medical image fusion: s survey of the state of the art. In *Information Fusion*, (Vol. 19 pp. 4–19). http://dblp.uni-trier.de/db/journals/inffus/inffus19.html#JamesD14.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. Cambridge: MIT Press.

Lau, B.P.L., Hasala, M.S., Zhou, Y., Hassan, N.U., Yuen, C., Zhang, M., Tan, U.X. (2019). A survey of data fusion in smart city applications. *Information Fusion*, *52*, 357–374.

Li, F., Dong, X.L., Langen, A., Li, Y. (2017). Discovering multiple truths with a hybrid model. arXiv:1705.04915.

Li, Q., Li, Y., Gao, J., Su, L., Zhao, B., Demirbas, M., Fan, W., Han, J. (2014a). A confidence-aware approach for truth discovery on long-tail data. *PVLDB*, *8*(4), 425–436. http://dblp.uni-trier.de/db/journals/pvldb/pvldb8.html#LiLGSZDFH14.

Li, Q., Li, Y., Gao, J., Zhao, B., Fan, W., Han, J. (2014b). Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *SIGMOD conference, ACM* (pp. 1187–1198). http://dblp.uni-trier.de/db/conf/sigmod/sigmod2014.html#LiLGZFH14.

Li, X., Dong, X.L., Lyons, K., Meng, W., Srivastava, D. (2012). Truth finding on the deep web: Is the problem solved? arXiv:1503.00303.

Li, X., Dong, X.L., Lyons, K.B., Meng, W., Srivastava, D. (2015a). Scaling up copy detection. In *ICDE, IEEE computer society* (pp. 89–100). http://dblp.uni-trier.de/db/conf/icde/icde2015.html#LiDLMS15.

Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., Fan, W., Han, J. (2015b). A survey on truth discovery. *SIGKDD Explorations*, *17*(2), 1–16. http://dblp.uni-trier.de/db/journals/sigkdd/sigkdd17.html#LiGMLSZFH15.

Li, Y., Li, Q., Gao, J., Su, L., Zhao, B., Fan, W., Han, J. (2016). Conflicts to harmony: a framework for resolving conflicts in heterogeneous data by truth discovery. *IEEE TKDE*, *28*(8), 1986–1999. http://dblp.uni-trier.de/db/journals/tkde/tkde28.html#LiLGSZFH16.

Lillis, D., Toolan, F., Collier, R.W., Dunnion, J. (2006). Probfuse: a probabilistic approach to data fusion. In *SIGIR, ACM* (pp. 139–146). http://dblp.uni-trier.de/db/conf/sigir/sigir2006.html#LillisTCD06.

Lin, X., & Chen, L. (2018). Domain-aware multi-truth discovery from conflicting sources. *PVLDB*, *11*(5), 635–647. http://dblp.uni-trier.de/db/journals/pvldb/pvldb11.html#LinC18.

Liu, W., Liu, J., Duan, H., Hu, W., Wei, B. (2017a). Exploiting source-object networks to resolve object conflicts in linked data. In *ESWC (1), lecture notes in computer science*, (Vol. 10249 pp. 53–67). http://dblp.uni-trier.de/db/conf/esws/eswc2017-1.html#LiuLDHW17.

Liu, W., Liu, J., Duan, H., Zhang, J., Hu, W., Wei, B. (2017b). Truthdiscover: resolving object conflicts on massive linked data. In *WWW (Companion Volume), ACM*, Vol. 243–246. http://dblp.uni-trier.de/db/conf/www/www2017c.html#LiuLDZHW17.

Liu, W., Liu, J., Wei, B., Duan, H., Hu, W. (2019). A new truth discovery method for resolving object conflicts over linked data with scale-free property. *Knowledge and Information Systems*, *59*(2), 465–495. http://dblp.uni-trier.de/db/journals/kais/kais59.html#LiuLWDH19.

Liu, X., Dong, X.L., Ooi, B.C., Srivastava, D. (2011). Online data fusion. *PVLDB*, *4*(11), 932–943. http://dblp.uni-trier.de/db/journals/pvldb/pvldb4.html#LiuDOS11.

Ma, B., Jiang, T., Zhou, X., Zhao, F., Yang, Y. (2017). A novel data integration framework based on unified concept model. *IEEE Access*, *5*, 5713–5722. http://dblp.uni-trier.de/db/journals/access/access5.html#MaJZZY17.

Michelfeit, J., & Mynarz, J. (2014). New directions in linked data fusion. In *ISWC (Posters & Demos), CEUR workshop proceedings*, (Vol. 1272 pp. 397–400). http://dblp.uni-trier.de/db/conf/semweb/iswc2014p.html#MichelfeitM14.

Michelfeit, J., Knap, T., Necaský, M. (2014). Linked data integration with conflicts. arXiv:1410.7990.

Motro, A., & Anokhin, P. (2006). Fusionplex: resolution of data inconsistencies in the integration of heterogeneous information sources. *Information Fusion*, *7*(2), 176–196. http://dblp.uni-trier.de/db/journals/inffus/inffus7.html#MotroA06.

Nakhaei, Z., & Ahmadi, A. (2017). Toward high level data fusion for conflict resolution. In *ICMLC, IEEE* (pp. 91–97). http://dblp.uni-trier.de/db/conf/icmlc/icmlc2017.html#NakhaeiA17.

Pasternack, J., & Roth, D. (2010). Knowing what to believe (when you already know something). In *COLING* (pp. 877–885). Tsinghua: Tsinghua University Press. http://dblp.uni-trier.de/db/conf/coling/coling2010.html#PasternackR10.

Pasternack, J., & Roth, D. (2011). Making better informed trust decisions with generalized fact-finding. In *Proceedings of the twenty-second international joint conference on artificial intelligence - Volume Three, AAAI Press, IJCAI'11* (pp. 2324–2329). https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-387.

Pasternack, J., & Roth, D. (2013). Latent credibility analysis. In *Proceedings of the 22nd international conference on World Wide Web, international world wide web conferences steering committee* (pp. 1009–1020). http://www2013.org/proceedings/p1009.pdf.

Pochampally, R., Sarma, A.D., Dong, X.L., Meliou, A., Srivastava, D. (2014). Fusing data with correlations. In *SIGMOD conference, ACM* (pp. 433–444). http://dblp.uni-trier.de/db/conf/sigmod/sigmod2014.html#PochampallySDMS14.

Preece, A.D., Hui, K.Y., Gray, W.A., Marti, P., Bench-Capon, T.J.M., Cui, Z., Jones, D.M. (2001). Kraft: an agent architecture for knowledge fusion. *International Journal of Cooperative Information Systems*, *10*(1-2), 171–195. http://dblp.uni-trier.de/db/journals/ijcis/ijcis10.html#PreeceHGMBCJ01.

Rekatsinas, T., Joglekar, M., Garcia-Molina, H., Parameswaran, A.G., Ré, C. (2017). Slimfast: guaranteed results for data fusion and source reliability. In *SIGMOD conference, ACM* (pp. 1399–1414). http://dblp.uni-trier.de/db/conf/sigmod/sigmod2017.html#RekatsinasJGPR17.

Saha, B., & Srivastava, D. (2014). Data quality: the other face of big data. In *ICDE, IEEE computer society* (pp. 1294–1297). http://dblp.uni-trier.de/db/conf/icde/icde2014.html#SahaS14.

Sethi, P., & Sarangi, S.R. (2017). Internet of things: architectures, protocols, and applications. *J Electrical and Computer Engineering*, *2017*, 9324035:1–9324035:25.

Soldatos, J., Kefalakis, N., Hauswirth, M., Serrano, M., Calbimonte, J.P., Riahi, M., Aberer, K., Jayaraman, P.P., Zaslavsky, A.B., Zarko, I.P., Skorin-Kapov, L., Herzog, R. (2014). Openiot: open source internet-of-things in the cloud. In *OpenIoT@SoftCOM, Springer, lecture notes in computer science*, (Vol. 9001 pp. 13–25). http://dblp.uni-trier.de/db/conf/softcom/openiot2014.html#SoldatosKHSCRAJ14.

Torra, V., & Narukawa, Y. (2007). *Modeling decisions - information fusion and aggregation operators*. New York: Springer.

Waguih, D.A., & Berti-Équille, L. (2014). Truth discovery algorithms: an experimental evaluation. arXiv:1409.6428.

Wang, C. (2010). Data analysis in incomplete information systems based on granular computing. In *2010 International conference on system science, engineering design and manufacturing informatization*, (Vol. 2 pp. 153–155).

Wang, M., Perera, C., Jayaraman, P.P., Zhang, M., Strazdins, P., Shyamsundar, R.K., Ranjan, R. (2016). City data fusion: Sensor data fusion in the internet of things. *IJDST*, *7*(1), 15–36. http://dblp.uni-trier.de/db/journals/ijdst/ijdst7.html#WangPJZSSR16.

Wang, X., Sheng, Q.Z., Fang, X.S., Yao, L., Xu, X., Li, X. (2015). An integrated bayesian approach for effective multi-truth discovery. In Bailey, J., Moffat, A., Aggarwal, C.C., de Rijke, M., Kumar, R., Murdock, V., Sellis, T.K., Yu, J.X. (Eds.) *CIKM, ACM* (pp. 493–502). http://dblp.uni-trier.de/db/conf/cikm/cikm2015.html#WangSFYXL15.

Wang, Y., Ma, F., Su, L., Gao, J. (2017). Discovering truths from distributed data. In *ICDM, IEEE computer society* (pp. 505–514). http://dblp.uni-trier.de/db/conf/icdm/icdm2017.html#WangMSG17.

Wang, Z., & Ma, Y. (2008). Medical image fusion using m-pcnn. *Information Fusion*, *9*(2), 176–185. http://dblp.uni-trier.de/db/journals/inffus/inffus9.html#WangM08.

Wu, H., Pei, Y., Li, B., Kang, Z., Liu, X., Li, H. (2015). Item recommendation in collaborative tagging systems via heuristic data fusion. *Knowledge-Based Systems*, *75*, 124–140. http://dblp.uni-trier.de/db/journals/kbs/kbs75.html#WuPLKLL15.

Wu, S. (2012a). *Data fusion in information retrieval., adaptation, learning, and optimization* Vol. 13. New York: Springer.

Wu, S. (2012b). *Data fusion in information retrieval., adaptation, learning, and optimization* Vol. 13. New York: Springer.

Xiao, H., Gao, J., Li, Q., Ma, F., Su, L., Feng, Y., Zhang, A. (2016). Towards confidence in the truth: a bootstrapping based truth discovery approach. In *KDD, ACM* (pp. 1935–1944). http://dblp.uni-trier.de/db/conf/kdd/kdd2016.html#XiaoGLMSFZ16.

Xie, Z., Liu, Q., Bao, Z. (2017). Sifting truths from multiple low-quality data sources. In *APWeb/WAIM (1), Springer, lecture notes in computer science*, (Vol. 10366 pp. 74–81). http://dblp.uni-trier.de/db/conf/apweb/apweb2017-1.html#XieLB17.

Xu, W., & Yu, J. (2017). A novel approach to information fusion in multi-source datasets: a granular computing viewpoint. *Information Sciences*, *378*, 410–423.

Yang, Y., Bai, Q., Liu, Q. (2018). A probabilistic model for truth discovery with object correlations. *Knowledge-Based Systems*, *165*, 360–373. http://dblp.uni-trier.de/db/journals/kbs/kbs165.html#YangBL19.

Yin, X., & Tan, W. (2011). Semi-supervised truth discovery. In *WWW, ACM* (pp. 217–226). http://dblp.uni-trier.de/db/conf/www/www2011.html#YinT11.

Yin, X., Han, J., Yu, P.S. (2008). Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, *20*(6), 796–808. http://dblp.uni-trier.de/db/journals/tkde/tkde20.html#YinHY08.

Yu, D., Huang, H., Cassidy, T., Ji, H., Wang, C., Zhi, S., Han, J., Voss, C.R., Magdon-Ismail, M. (2014). The wisdom of minority: unsupervised slot filling validation based on multi-dimensional truth-finding. In Hajic J, & Tsujii, J. (Eds.) *COLING, ACL* (pp. 1567–1578).

Zhang, H., Li, Q., Ma, F., Xiao, H., Li, Y., Gao, J., Su, L. (2016). Influence-aware truth discovery. In *CIKM, ACM* (pp. 851–860). http://dblp.uni-trier.de/db/conf/cikm/cikm2016.html#ZhangLMXLGS16.

Zhang, J., Wang, S., Wu, G., Zhang, L. (2018). A effective truth discovery algorithm with multi-source sparse data. In *ICCS (3), Springer, lecture notes in computer science*, (Vol. 10862 pp. 434–442). http://dblp.uni-trier.de/db/conf/iccS/iccS2018-3.html#ZhangWWZ18.

Zhao, B., Rubinstein, B.I.P., Gemmell, J., Han, J. (2012). A bayesian approach to discovering truth from conflicting sources for data integration. arXiv:1203.0058.

Zheng, Y., Yin, M., Luo, J., He, G. (2019). Truth discovery on multi-dimensional properties of data sources. In *ACM TUR-C, ACM* (pp. 164:1–164:8). http://dblp.uni-trier.de/db/conf/acmturc/acmturc2019.html#ZhengYLH19.