Generating Actionable Knowledge from Big Data: Knowledge Extraction and Truth Discovery

MACQUARIE University

Xiu (Susie) Fang

Department of Computing

Macquarie University

This dissertation is submitted for the degree of

Doctor of Philosophy

Supervisors: Prof. Michael Sheng, Prof. Anne H.H. Ngu,

and Prof. Jian Yang

January 2018

To my mother and father,

my husband and my little prince,

who made all of this possible,

for their endless encouragement and patience.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the Macquarie University and where applicable, any partner institution responsible for the joint-ward of this degree. I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968. I also give permission for the digital version of my thesis to be made available on the Web, via the University's digital research repository, the Library Search and also through Web search engines, unless permission has been granted by the University to restrict access for a period of time.

Xiu (Susie) Fang

January 2018

Acknowledgements

At the beginning of this dissertation, I would like to take time to thank all the people without whom this project would never have been possible. Although it is just my name on the cover, many people have contributed to the research in their own particular way and for that I want to give them special thanks.

First and foremost I would like to express my sincere gratitude to my principal supervisor, Prof. Michael Sheng. He is such a person with goodness, intelligence, integrity, generosity, and patience, it has been a great honor to be his Ph.D. student. He has created the invaluable space for me to do this research and develop myself as a researcher in the best possible way. I greatly appreciate the freedom he has given me to find my own path and the guidance and support he offered when needed. He is the person who always reminds me to keep a positive attitude towards rejecting comments of my research submissions and to never lose heart. Without him, I could not be successful at any point of time. I am also thankful for the excellent example he has provided as a successful researcher. Secondly, I would like to thank my co-supervisor, Prof. Anne H.H. Ngu. The joy and enthusiasm she has for her research was contagious and motivational for me, even during tough times in the Ph.D. pursuit. Her insightful suggestions and comments inspired my research works. Thirdly, I am very thankful to Prof. Jian Yang, who is another co-supervisor of mine, for her continuous support, advice and motivation.

The members of our research group have contributed immensely to my personal and professional time at the University of Adelaide and Macquarie University. The group has

been a source of friendships as well as good advice and collaboration. I am especially grateful for Dr. Xianzhi Wang's precious suggestions and comments on my research. I would also like to acknowledge my fellow colleagues, including Dr. Lina Yao, Dr. Yongrui Qin, Dr. Yihong Zhang, and Dr. Wei Zhang, for their creative ideas which inspired my research.

I sincerely appreciate the University of Adelaide and Macquarie University, who provided the Postgraduate Research Scholarship to financially support my Ph.D. study.

Lastly, I would like to thank my family for all their love and encouragement. For my parents who raised me with a love of science and supported me in all my pursuits by devoting all that they can offer. For my parents-in-law who helped me a lot with taking care of my little baby so that I can concentrate on my research. For my little prince whose angel-like smile is always my motivation. Most of all for my loving, supportive, encouraging, and patient husband Mengyang Sun whose faithful support during the final stages of this Ph.D. is so appreciated. Thank you.

Abstract

To revolutionize our modern society by utilizing the wisdom of Big Data, considerable knowledge bases (KBs) have been constructed to feed the massive knowledge-driven applications with Resource Description Framework (RDF) triples. The important challenges for KB construction include extracting information from large-scale, possibly conflicting and different-structured data sources (i.e., the knowledge extraction problem) and reconciling the conflicts that reside in the sources (i.e., the truth discovery problem). Tremendous research efforts have been contributed on both problems respectively. However, the existing KBs are far from being comprehensive and accurate.

In this dissertation, we first propose a system for **g**ene**r**ating **a**ctionable k**n**owledge from Big **D**ata, and use this system to construct a comprehensive KB, called GrandBase. Then we solve the raised research issues regarding GranbBase construction by developing a series of methodologies: Firstly, we study predicate extraction and implement ontology augmentation for knowledge base expansion. Secondly, we address truth discovery (on both single-valued and multi-valued objects or predicates) and performance evaluation on truth discovery methods for knowledge base purification. In particular, we first propose a framework for extracting new predicates from four types of data sources, namely Web texts, Document Object Model (DOM) trees, existing KBs, and query stream to augment the ontology of the existing KB (i.e., Freebase). We use query stream and two major KBs, DBpedia and Freebase, to seed the predicate extraction from Web texts and DOM trees. Then, to estimate value veracity for multi-valued objects, we model the endorsement relations

among sources by quantifying their two-sided inter-source agreements. Two aspects of source reliability are derived from the two graphs constructed by modeling the inter-source relations. To more precisely estimate source reliability for effective multi-valued truth discovery, our graph-based model incorporates four important implications, including *two types of source relations*, *object popularity*, *loose mutual exclusion*, and *long-tail phenomenon on source coverage*. After that, to fully leverage the advantages of the existing truth discovery methods and achieve more robust and better truth discovery, we propose to extract truth from the prediction results of those methods. Our ensemble approach distinguishes between the single-valued and multi-valued truth discovery problems. Finally, for performance evaluation of truth discovery methods, as the ground truth may be very limited or even impossible to obtain, we make the attempt towards conducting evaluation without using ground truth.

For each of the models and approaches presented in this dissertation, we have conducted extensive experiments using either real-world or synthetic datasets. Empirical studies show the effectiveness of our approaches.

Finally, we also discuss the future research directions regarding GrandBase construction and extension in this dissertation.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

According to IBM[1], 2.5 quintillion bytes of data are created every day and 90% of data in the world has been created in the past two years. Thanks to the unprecedented information explosion, the modern Web has gradually evolved into a huge data repository. To exploit the full potential and support unified representation of such data, knowledge base (KB) construction has become an important research topic to both database and knowledge management communities. Recent years have witnessed a proliferation of large-scale KBs [1], including academic KBs, such as YAGO [2], NELL [3], DBpedia [4], Elementary/DeepDive [5, 6], KnowItAll [7, 8], ImageNet [9], BabelNet [10], ConceptNet [11], Wikidata [12], WikiNet [13], and industrial KBs, such as Satori constructed by Microsoft[2] to enhance Bing's semantic searching, Google's Knowledge Graph[3], which is a replacement of Freebase [14], served as the backbone of many Google applications, and "Entity Graph" built by Facebook[4] to boost the social network searching. Moreover, there are also some commercial projects on knowledge base construction, including Google Knowledge Graph

[1] http://www-01.ibm.com/software/data/bigdata/

[2] http://blogs.bing.com/search/2013/03/21/understand-your-world-with-bing/

[3] https://twitter.com/jeffjarvis/status/783338071316135936

[4] https://www.fastcompany.com/3006389/where-are-they-now/entity-graph-facebookcalls- users-improve-its-search

and related work at Google [15–17], the EntityCube and Probase projects at Microsoft Research [18, 19], and IBM's Watson project [19], of which some projects are still ongoing.

The majority of current KBs store data in the form of {subject, predicate, object}, or *Resource Description Framework* (RDF) triples, which we call *actionable knowledge*. So far, more than 100 billion of SPO (subject-predicate-object) facts about the real world, including named entities, their semantic classes, and their mutual relationships, are collected or extracted from more than $1,000$ sources, to build those KBs[5] (see Figure 1.1). Such triples can be utilized to efficiently and effectively change human lives by enabling applications such as semantic search and question answering, natural language understanding, recommender systems, text analytics, data cleaning, disambiguation, deep reasoning, and machine reading.

Despite the large scale of the existing KBs, they are still far from complete and accurate. Take the top two largest KBs, Freebase and Knowledge Graph, as examples. The former covers 40 million entities, but only $4,000$ properties (note that in Freebase, predicates are referred to as properties). The latter covers 20 billion facts about 600 million entities. While the type *University* has only 9 properties in Freebase and the type *CollegeOrUniversity*[6] contains only 59 properties in Knowledge Graph, a person can easily spot more properties for a university in the real life. Another example is that a large amount of people in Freebase have no known place of birth or nationality, due to the conflicts reside in the multi-source data. When it comes to the rare or multi-valued predicates, the lack of values is more serious.

As KB construction involves extracting information from large-scale, possibly conflicting, and different-structured data sources and determining the data veracity by estimating the reliability of data sources given the conflicting multi-source data [20], two of the major reasons regarding the unsatisfied coverage and accuracy of the existing KBs are the unsolved *knowledge extraction* and *truth discovery* problems [21]. Specifically, knowledge extraction techniques (i.e., refiners) aim at obtaining machine-readable and interpretable knowledge

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

[6] http://schema.org/CollegeOrUniversity

from structured (e.g., relational databases), semi-structured (e.g., Extensible Markup Language (XML)) and/or unstructured sources (e.g., texts, documents, images). Truth discovery is a fundamental research topic, with the goal of estimating data veracity automatically by resolving the conflicts in multi-source data. In this dissertation, we focus on those two problems to effectively and efficiently generate actionable knowledge from big data.

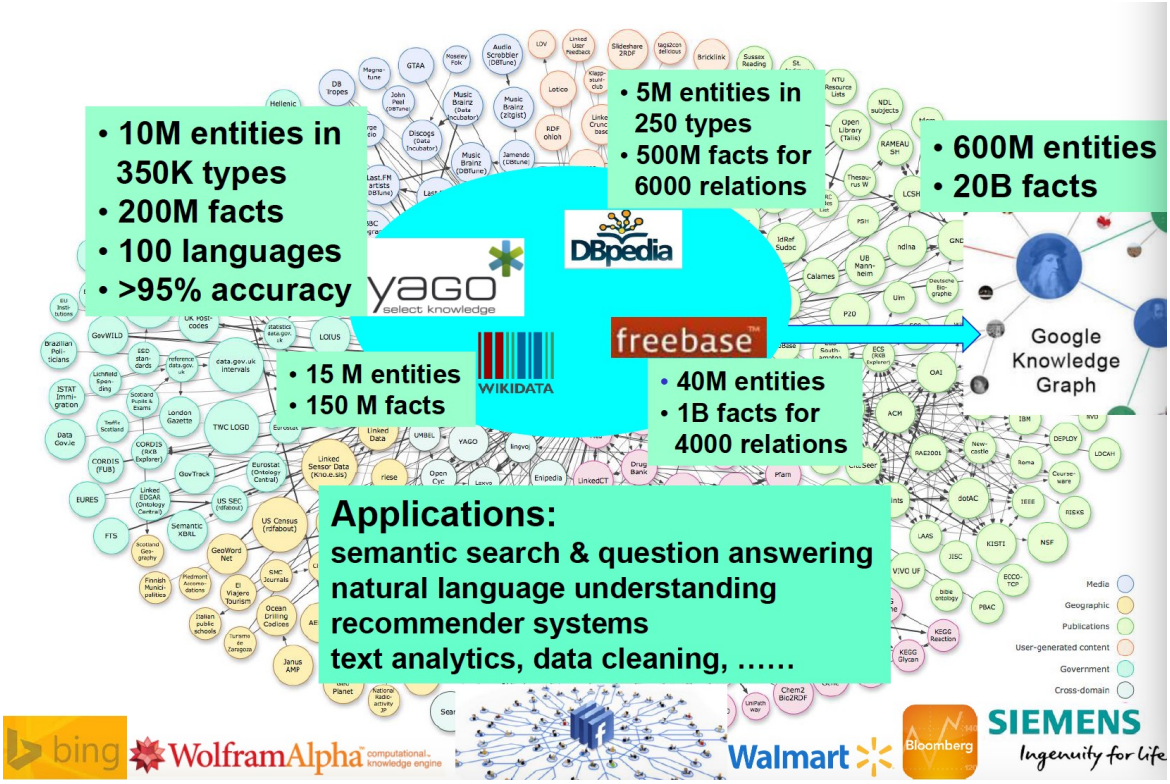


Fig. 1.1 Web of Knowlege

This chapter is organized as follows. In Section 1.1, we present the specific research issues to be addressed in this dissertation. In Section 1.2, we outline our contributions by tackling those research issues. In Section 1.3, we enumerate the pulications by the author that are related to this work. Finally, in Section 1.4, we present the stucture of this dissertation.

1.1 Research Issues in Knowledge Extraction and Truth Discovery

The works in this dissertation tackle a number of research issues in *knowledge extraction* and *truth discovery* for knowledge base construction.

Tremendous knowledge extraction techniques (i.e., refiners) have been proposed to obtain knowledge from the Open Web [22–26]. However, there are two limitations with the current approaches: i) most existing KBs, such as Freebase, DBpedia, and DeepDive, are constructed by applying refiners that focus on extracting knowledge from a single type of data sources (e.g., Web texts). In particular, these KBs simply remove tags and extract data from plain texts, and ignore the knowledge contained in the DOM tree structures formed by the tags. For this reason, these KBs cannot exploit the full knowledge contained in the data sources, leading to limited coverage and quality of the extractions. In fact, various types of data sources, such as DOM trees, HTML tables, and human annotated pages [15], can be used for more accurate and complete knowledge extraction; ii) previous research efforts commonly focus on extracting facts of entities in a *predefined ontology*, which limits the coverage of extractions. Although several approaches, such as open information extraction (Open IE) [27], manage to add new entities and relations to the extractions, they fail to distinguish synonyms, therefore introducing extra redundancy to the results. Under such circumstance, ontology augmentation by extracting knowledge from multiple types of sources becomes a fundamental research issue for KB construction.

While extracting knowledge from the Web, we can easily observe that multiple sources often provide *conflicting* descriptions on the same objects (in this dissertation, when we discuss the truth discovery problem, we all refer to predicates as objects) of interest, due to typos, out-of-date data, missing records, or erroneous entries, making it difficult to determine which data source should be trusted. For example, in online healthcare systems

that collect reviews from patients, the posted reviews for the same drug may vary due to the diverse physical conditions of patients [28]; in social networks that provide opportunities for individuals to comment on the physical world whenever and wherever they want to, the posts may have varied reliability regarding the same event (such as gas shortage after a disaster, or physical conditions after an earthquake) and result in conflicting observations [29, 30]; in crowdsourcing systems that solicit labels from worldwide workers, the labels of the same task may be diverse due to workers' varied skills, expertise, and biases [31–33]. Another example is that in systems that extract information about their topics of interest from the Web, the outputs of different refiners may differ on the same topic, due to the varied capability of the refiners and the various corpora they focus on [20, 21, 34]. Even worse, some sources may intentionally provide false data or copied data from other sources to misguide people. Being misled by those conflicting data could lead to considerable damages and financial loss in many applications such as drug recommendation in healthcare systems or price prediction in the stock markets [35]. It is thus urgent and important to discover the truth from those conflicting data.

Due to the large-scale of data, it is unrealistic to expect a human to be able to manually determine which data is true. Therefore, a fundamental research topic for KB construction named *truth discovery* (also known as *information corroboration* [36], *information credibility* [37], *conflicting data integration* [38], *fact-checking* [39], *data fusion* [40, 41], and *knowledge fusion* [20]) has emerged. Though Considerable research efforts have been conducted [36, 42–44, 40, 45, 46], by applying different formulas and models while incorporating different additional factors (such as data types, source dependency, source quality, object properties, and value implications), to solve the truth discovery problem, there are still several research issues waiting to be better sovled: i) most of the existing methods commonly assume that each object has exactly one true value (i.e., *single-valued* assumption). However, in real world, multi-valued objects—such as the children of a person, the authors

of a book—widely exist. To conduct truth discovery while taking multi-valued objects into consideration, becomes an important research topic, which is also known as *multi-valued truth discovery* (MTD); ii) Several surveys [47–49] have shown that a "one-fits-all" truth discovery method is not achievable due to the limitations of the existing methods. Therefore, combing various competing methods could be an effective alternative for conducting high-quality truth discovery; iii) For the purpose of performance evaluation of truth discovery methods, the ground truth information is always assumed to be available. Unfortunately, we cannot make this assumption in pratice, especially in the Big Data era. How to evaluate the performance of various truth discovery methods without using ground truth becomes another big challenge for the truth discovery applications.

1.2 Our Contributions

Based on the aforementioned research issues, this dissertation makes the following contributions to the domain of knowledge extraction and truth discovery for KB construction.

GrandBase

To revolutionize our modern society by utilizing the wisdom of Big Data, considerable knowledge bases (KBs) have been constructed to feed the massive knowledge-driven applications with RDF triples, such as Google Knowledge Graph [50] and the IBM Watson question answering system [51]. The important challenges for KB construction include extracting information from large-scale, possibly conflicting and different-structured data sources (i.e., the *knowledge extraction* problem) and reconciling the conflicts that reside in the sources (i.e., the *truth discovery* problem). Tremendous research efforts have been contributed on both problems respectively. However, the existing KBs are far from being comprehensive and accurate. Aiming at **G**ene**r**ating **a**ctionable k**n**owledge from Big **D**ata, we propose a

system, which consists of two phases, namely *knowledge extraction* and *truth discovery*, as an overall solution to construct a comprehensive KB, called *GrandBase*. Empirical studies demonstrate the effectiveness of our approaches and the potential of GrandBase.

Predicate Extraction for Ontology Augmentation

A comprehensive ontology can ease the discovery, maintenance and popularization of knowledge in many domains. As a means to enhance existing ontologies, predicate extraction has attracted tremendous research attention. However, most existing attribute extraction techniques focus on exploring a single type of sources, such as structured (e.g., relational databases), semi-structured (e.g., Extensible Markup Language (XML)) or unstructured sources (e.g., Web texts, images), which leads to the poor coverage of knowledge bases (KBs). Our contribution is that we propose a novel framework that extracts and merges the predicates from four types of sources, existing KBs (i.e., Freebase and DBpedia), query stream, Web texts, and DOM trees, for comprehensive ontology augmentation. In particular, we first extract predicates from existing KBs and query stream as seeds. We adopt new patterns and filtering rules for better query stream extraction. Then, we utilize those seeds to learn tag path patterns (from DOM trees), and lexical and parse patterns (from Web texts). Those patterns are in turn leveraged to extract new predicates from DOM trees and Web texts. Experiments show the capability of our approach in augmenting existing KB ontology.

Multi-Valued Truth Discovery

Most of the current truth discovery methods assume only one true value for each object, while in reality objects with multiple true values widely exist. The few existing methods that cope with multi-valued objects still lack of accuracy. To tackle this issue, we first propose a novel approach, which models the endorsement relations among sources by quantifying their two-sided inter-source agreements, for multi-valued truth discovery. Based on this approach,

we further propose a full-fledged graph-based model, to pursue more accurate and complete results. Our model incorporates four important implications, including two types of source relations, object popularity, loose mutual exclusion and long-tail phenomenon on source coverage. Empirical studies on two large real-world datasets demonstrate the effectiveness of our approach.

Combining Existing Methods for Better Truth Discovery

Surveys on truth discovery methods show that none of the existing methods is a clear winner that consistently outperforms the others due to the varied characteristics of different methods. In addition, in some cases, an improved method may not even beat its original version as a result of the bias introduced by limited ground truths or different features of the applied datasets. To realize an approach that achieves better and robust overall performance, we propose to combine the existing methods by adopting two models, namely *serial model* and *parallel model*, to extract truth from the outputs of these methods. Extensive experimental results show that our approach outperforms traditional methods on both real-world and synthetic datasets.

Performance Evaluation without Using Ground Truth

Previous comparative studies on truth discovery methods are based on real-world datasets with sparse ground truth. Such sparse ground truth is not statistically significant to be legitimately used for evaluating and comparing existing methods in a systematic way. To tackle this problem, we propose an approach for comparing truth discovery methods without using ground truth. We conduct extensive experiments on both synthetic and real-world datasets to demonstrate the effectiveness of our proposed approach. Our approach consistently achieves more accurate rankings of the twelve evaluated methods than traditional evaluation approach based on sparse ground truth.

1.3 Dissertation Publications

In this section, I would like to list the publications that are produced from this dissertation (out of the 17 publications and submissions during the author's PhD study). The list of the papers, including all accepted, revised and submitted manuscripts is as follows:

Journals

1. **Xiu Susie Fang**, Quan Z. Sheng, Xianzhi Wang, Anne H.H. Ngu, and Yihong Zhang. "GrandBase: Generating Actionable Knowledge from Big Data". International Journal on PSU Research Review, Emerald, 1(2):105–126, 2017 (Invited Paper).

2. **Xiu Susie Fang**, Quan Z. Sheng, Xianzhi Wang, and Anne H. H. Ngu. "SmartVote: A Full-Fledged Graph-Based Model for Multi-Valued Truth Discovery". Submitted to World Wide Web Journal (WWWJ). Under minor revision.

3. Xianzhi Wang, Michael Sheng, Lina Yao, **Xiu Susie Fang**, Xiaofei Xu, and Xue Li. "Generalizing Truth Discovery by Incorporating Multi-Truth Features". Submitted to TWeb.

Conferences

1. **Xiu Susie Fang**, Quan Z. Sheng, Xianzhi Wang, Wei Emma Zhang, Anne H.H. Ngu. "How to Compare Truth Discovery Methods When Ground Truth is Missing?". Submitted to SIGIR 2018.

2. **Xiu Susie Fang**, Quan Z. Sheng, Xianzhi Wang, and Anne H. H. Ngu. "SmartMTD: A Graph-Based Approach for Effective Multi-Truth Discovery". Submitted to SIGIR 2018.

3. **Xiu Susie Fang**, Quan Z. Sheng, Xianzhi Wang, Mahmoud Barhamgi, Lina Yao, and Anne H.H. Ngu. "SourceVote: Fusing Multi-Valued Data via Inter-Source Agreements". In Proceedings of the 36th International Conference on Conceptual Modeling (ER 2017). 6-9 November 2017, Valencia, Spain.

4. **Xiu Susie Fang**. "Truth Discovery from Conflicting Multi-Valued Objects". In Proceedings of the 26th International World Wide Web Conference (WWW 2017 Companion), 3 - 7 April 2017, Perth, Australia.

5. **Xiu Susie Fang**, Quan Z. Sheng, Xianzhi Wang, and Anne H.H. Ngu. "Value Veracity Estimation for Multi-Truth Objects via a Graph-Based Approach". In Proceedings of the 26th International World Wide Web Conference (WWW 2017 Companion), 3 - 7 April 2017, Perth, Australia.

6. **Xiu Susie Fang**, Quan Z. Sheng, and Xianzhi Wang. "An Ensemble Approach for Better Truth Discovery". In Proceedings of the 12th Anniversary of the International Conference on Advanced Data Mining and Applications (ADMA 2016) , 12 - 15 Dec 2016, Gold Coast, Australia.

7. Xianzhi Wang, Quan Z. Sheng, Lina Yao, Xue Li, **Xiu Susie Fang**, and Xiaofei Xu. "Truth Discovery via Exploiting Implications from Multi-Source Data". In Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM 2016). 24 - 28 October, 2016, Indianapolis, USA.

8. Xianzhi Wang, Quan Z. Sheng, Lina Yao, Xue Li, **Xiu Susie Fang**, and Xiaofei Xu. "Empowering Truth Discovery with Multi-Truth Prediction". In Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM 2016). 24 - 28 October, 2016, Indianapolis, USA.

9. Xianzhi Wang, Quan Z. Sheng, **Xiu Susie Fang**, Lina Yao, Xiaofei Xu, and Xue Li. "An Integrated Bayesian Approach for Multi-Truth Discovery". In Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM 2015). 19 - 23 October 2015, Melbourne, VIC, Australia.

10. Xianzhi Wang, Quan Z. Sheng, **Xiu Susie Fang**, Xue Li, Xiaofei Xu, and Lina Yao. "Approximate Truth Discovery via Problem Scale Reduction". In Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM 2015). 19 - 23 October 2015, Melbourne, VIC, Australia.

11. **Xiu Susie Fang**, Xianzhi Wang, and Quan Z. Sheng. "Ontology Augmentation via Attribute Extraction from Multiple Types of Sources". In Proceedings of the 26th Australasian Database Conference (ADC 2015). 4 - 6 June 2015, Melbourne, VIC, Australia.

12. **Xiu Susie Fang**. "Generating Actionable Knowledge from Big Data". In Proceedings of the 2015 SIGMOD PhD Symposium (SIGMOD 2015), May 31 - June 04 2015, Melbourne, VIC, Australia.

1.4 Dissertation Organization

The reminder of this dissertation is organized as follows:

In Chapter 2, we review the literature that are closely related to our work. Specifically, we present an overview of the big data integration and knowledge bases as well as the techniques used in knowledge extraction and truth discovery. We also present the system of GrandBase construction as an overview of our work.

In Chapter 3, we present our work on ontology augmentation via predicate extraction from multiple types of sources. We first introduce our framework. Then, the methods for merging the attribute extractions from Freebase and DBpedia, the method for query stream

extraction, and the algorithm for extracting attributes from DOM trees using the above extractions as seeds, are sequentially described.

In Chapter 4, we describe our novel approach for multi-valued truth discovery. Our graph-based model captures the endorsement relations among sources by quantifying their two-sided inter-source agreements. Two graphs are constructed based on those relations, from which we derive two aspects of source reliability. The source reliability quantification can also be utilized to initialize existing truth discovery methods.

In Chapter 5, to further improve the model introduced in Chapter 4, we propose a full-fledged graph-based model for better multi-valued truth discovery. We first discuss the observations that motivate our work, and validate the claim that the agreement among sources indicate endorsement of source trustworthiness. Then we present the framework of our model, which incorporates four implications into one graph-based core component. The methodology for each component is also introduced to facilitate the accurate multi-valued truth discovery.

In Chapter 6, we first formally define the ensemble truth discovery problem. Then, we analyze the feasibility of the ensemble approach. We present two implementation models for the approach. Our approach also distinguishes between two types of truth discovery problems, i.e., the single-valued truth discovery and multi-valued truth discovery problems.

In Chapter 7, we discuss the bias introduced by sparse ground truth in evaluating the truth discovery methods, by conducting experiments on synthetic datasets. As a key contribution, we propose a novel approach for comparing truth discovery methods without using ground truth.

Finally, in Chapter 8, we provide concluding remarks of this dissertation and discuss future work directions.

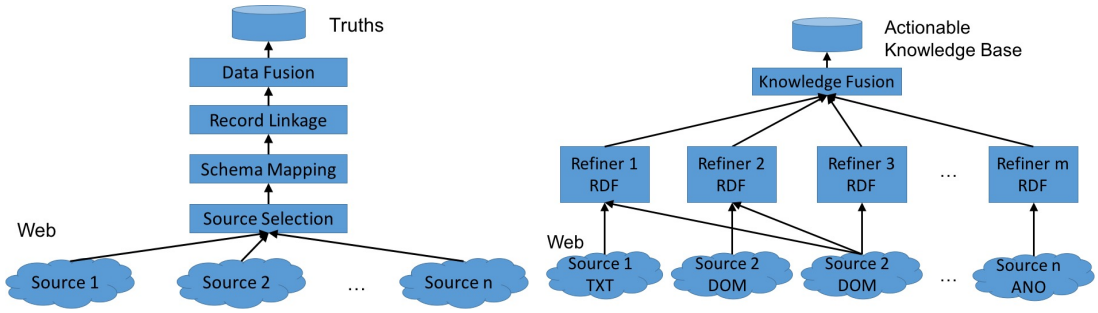
Chapter 2

Background

In this chapter, we give an introduction to the research fields related to our works, including big data integration, knowledge bases, knowledge extraction, and truth discovery. We also present the overview of our GrandBase construction, to help readers gain a better understanding of the works described in this dissertation. The chapter is organized as follows: In Section 2.1, we present an overview of knowledge bases. In Section 2.2, techniques used to extract data from the Web are introduced. Then, we overview the truth discovery methods in Section 2.3, Section 2.4, and Section 2.5. We introduce the overview of GrandBase construction in Section 2.6. In Section 2.7, we summarize this chapter.

2.1 Overview of Big Data Integration and Knowledge Bases

Nowadays, the 5V-dimension (volume, velocity, variety veracity and value) of big data becomes a hot topic of great importance, which inspires a significant number of research directions. With advanced data extraction and collection techniques, we can now easily collect data from various data sources to support all types of novel applications. For example, social analysis systems collect posts of diverse contents from social networks to predict social events [29, 52]; crowdsourcing systems get reports from worldwide workers on the



(a) Big Data Integration (b) Knowledge Base Construction

Fig. 2.1 Comparison of Big Data Integration and KB Construction

same set of tasks [31, 53, 54]; the emerging Internet of Things (IoT) systems gather signals from distributed sensors for a wide spectrum of applications, ranging from environmental monitoring [55], traffic management [56], to assisted living [57]. Modern applications are increasingly dependent on the multi-source data to gain valuable insights towards their interested domains [20]. During recent years, worldwide researchers have contributed their effort to the area of big data integration, which involves combining data residing in different sources and providing users with a unified view of these data.

Generally, big data integration faces three challenges [58, 59]: i) value heterogeneity, there may be different and conflicting values for the same objects in different data sources, which inspires the research work regarding truth discovery [60]. ii) Instance heterogeneity, each instance may be described by many different data records. Therefore, the techniques of string matching and object matching are required [61, 62]. iii) Structure heterogeneity, different sources store similar data with different schemas. Thus, we need to apply schema matching, model management and other relevant techniques [63, 64]. Specifically, in order to solve these issues, the workflow of data integration consists of four steps (see Figure 2.1a), including source selection [65, 66], schema alignment [67], record linkage [68, 69] and data fusion [38]. However, as the scale of data increases unprecedentedly, it becomes more urgent and important to extract full knowledge from the big data. As a result, knowledge base construction becomes another hot topic on big data integration.

KBs are built to provide actionable knowledge for both human use and feeding knowledge-driven applications, by fusing extractions from different types of Web sources [20] (see Figure 2.1b). Formally, a KB is a comprehensive and semantically organized machine-readable collection of universally relevant or domain-specific entities, classes, and SPO facts (attributes, relations), it also contains spatial and temporal dimensions, commonsense properties and rules, contexts of entities and facts, etc. Since 1985, early KBs have been built manually for human use, such as Wikipedia[1], and the seminal projects Cyc [70, 71] and WordNet [72], comprehensive automatic methods or algorithms for large-scale machine-readable KB construction and curation have been proposed in recent years, based on knowledge extractions from the Web sources for machine use. The research efforts on KB construction can be generally divided into four main groups. First, some researchers focus on constructing KBs based on high quality structured sources, such as Wikipedia infoboxes, including YAGO [2], YAGO2 [73], DBpedia [4], and Freebase [14]. Second, some KBs are built by using open information (schema-less) extraction techniques (Open IE), and extract data from the entire Web, including Reverb [8], OLLIE [74], and PRISMATIC [75]. These techniques can obtain lots of new facts, new entities from the Web. However, they work at the lexical level, and usually result in redundant facts which are worded differently but have the same semantic. Third, some techniques, such as NELL/ ReadTheWeb [3], PROSPERA [76], DeepDive/ Elementary [5], and Knowledge Vault [15], construct KBs by using a fixed ontology, and also extract data from the entire Web. These techniques generate smaller amount of entities from the Web than the Open IE techniques. However, the quality of data which are generated by these techniques are much higher than data generated by the Open IE techniques. Fourth, compared to general KBs with multiple types of predicates, there are also some methods, such as Probase [19], which construct taxonomies (is-a hierarchies).

Most of these KBs represent their data by using RDF triples, which provide large scale knowledge of the real world, such as named entities, their semantic classes, and their

[1]https://en.wikipedia.org/wiki/Main_Page

mutual relationships. Resource description framework, which is usually called RDF, depicts resources (particularly Web resources) in the forms of {subject, predicate, object} triples. The subject represents the resource. The predicate represents the property of the resource or the relationship between the resources. The object represents the value of the property of the certain resource or the resource has correlation with certain subject [77]. The data structure of RDF is so simple that it has been widely used to model disparate, abstract concepts, and fed to knowledge management applications. A dataset of RDF triples are essentially a large labelled, directed multi-graph which is very expressive. As such, an RDF-based data model is more naturally applicable to represent certain kinds of knowledge than the other ontological models. Moreover, many existing knowledge bases store RDF triples which can be used as priors for broader knowledge base construction [15]. Therefore, we refer to the collection of RDF triples as *actionable knowledge*. The backbone of the Web of Linked Data[2] is formed by interlinking those RDF-style KBs at the entity level [78].

In general, knowledge base construction follows these steps: knowledge extraction (discovering data sources, tapping unstructured data, connecting structured and unstructured data sources), truth discovery (making sense of heterogeneous, dirty, or uncertain data). To our knowledge, although knowledge base construction has been studied for many years, this research area is still far from mature. Both knowledge extraction and truth discovery techniques need to be further improved.

2.2 Overview of Knowledge Extraction

As a huge amount of data sources are available on the Web, which provide data in different styles (semi-structured, or unstructured), in order to make use of these data, many Web data extraction techniques have been proposed. The goal of Web data extraction systems is to extract information from the Web in an efficient manner [79, 80], and convert the Web data

[2] http://linkeddata.org/

to the user required structured formats [81, 82]. In our works, we focus on extracting data in RDF-format, that is, knowledge extraction.

2.2.1 Web Data Extraction

Existing techniques apply different methods to extract structured data from the Web, including Markov chains, graph theory, neural network approaches, association mining, statistical methods, etc. These techniques can be divided into four groups (see surveys in [80, 83, 84]):

Tree-based Techniques: These techniques are based on the Document Object Model (DOM tree). By using this model, the semi-structured or unstructured Web data can be represented in a hierarchical structure. This technique is the easiest and cheapest way to extract data from the Web [79].

Web Wrappers: These techniques implement several classes of algorithms to semi-automatically or automatically find out required data, and extract them from unstructured or semi-structured Web sources [79]. They are more costly but faster than other Web data extraction techniques, because they need to develop different programs to extract data from different Web sources, which consist of a life-cycle, including wrapper generation, wrapper execution and wrapper maintenance.

Machine Learning Approaches: These techniques are applied on semantic Web which is based on machine learning systems. They are designed for automatically extracting domain-specific information from the Web, which rely on training sessions. Statistical Machine Learning systems [85, 86] have been proposed for this group of methods.

Web Data Mining: These techniques are applications of data mining techniques. They are used for discovering hidden information and patterns from the Web pages. Web mining can

be divided into three types [87]: first, Web content mining, which extract knowledge from Web page contents which required by users. Some approaches [88] have been proposed for this purpose, such as statistical, neural network approaches, rapid miner, Web data extractor etc. Second, Web structure mining, of which the goal is to search the link structure of the Web and rank Web pages. Approaches including Page Rank [89], Weighted Page Rank [90] and HITS [91] have been proposed for these goals. Third, Web usage mining, of which the purpose is to understand user behaviour in interacting with a certain Web site.

2.2.2 Knowledge Extraction

The task of knowledge extraction is to obtain data in the machine-readable and machine-interpretable format from structured (relational databases, XML) and unstructured sources (text, documents, images). Currently, RDF is one of the most popular knowledge representation languages. Based on RDF, a large amount of work has been done in this area. For instance, the RDB2RDF W3C group has designed a standard language to extract RDF triples from relational databases. Also, there are some groups focusing on extracting RDF triples from Wikipedia. They construct knowledge bases, such as DBpedia and Freebase based on these triples. Specifically, the current research work covers transforming relational databases into RDF, identity resolution, knowledge discovery and ontology learning. Many researchers have been inspired and contribute to extract Web data into semantic Web format (RDF triples). For instance, DEiXTo[3], which is based on DOM Tree, creates extraction rules to convert Web information to any structured format, including RDF triples. Virtuoso Sponger[4] and Semantic Fire[5], which also support extract RDF triples from the Web sources. Some *refiners*(see e.g., [15, 92, 93]) also assign a confidence score to each triple to represent

[3]http://deixto.com/
[4]http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VirtSponger/
[5]https://code.google.com/p/semantic-fire/

the uncertainty about this extraction. The techniques designed for knowledge harvesting can be divided into four groups according to the types of knowledge:

Taxonomic Knowledge *Refiners*: These *refiners* search for individual entities, and organize them into semantic classes. This group of methods contains two kinds of methods: first, Wikipedia-centric methods, such as the methods proposed by [2, 94] link Wikipedia categories to WordNet, and Kylin Ontology generator [95] which learns more mappings by applying advanced ML methods (SVM's, MLN's). Second, Web-based methods, such as Watson [19] constructs a taxonomy from the Web. However, the coverage and the quality of the extractions from these *refiners* are not high.

Factual Knowledge *Refiners*: These *refiners* focus on given binary relations from the Web. There are several methods have been proposed for this purpose [15, 3, 96], including Regex-based extraction, pattern-based harvesting, consistency reasoning, probabilistic methods and Web-table methods. However, these *refiners* are not robust with both high precision and recall, not scalability enough, which need to be further improved.

Emerging Knowledge *Refiners*: Unlike the above two groups of *refiners*, which based on fixed ontology, this group of *refiners* use open information (schema-less) extraction techniques (Open IE) [97, 27], which seek for new relationships and new entities from the Web. However, they work at the lexical level, and usually result in redundant facts which are worded differently but have the same semantic. Moreover, although several methods like frequent sequence mining, Map-Reduce-parallelized on Hadoop have been designed, the scalability of this group of *refiners* still need to be developed.

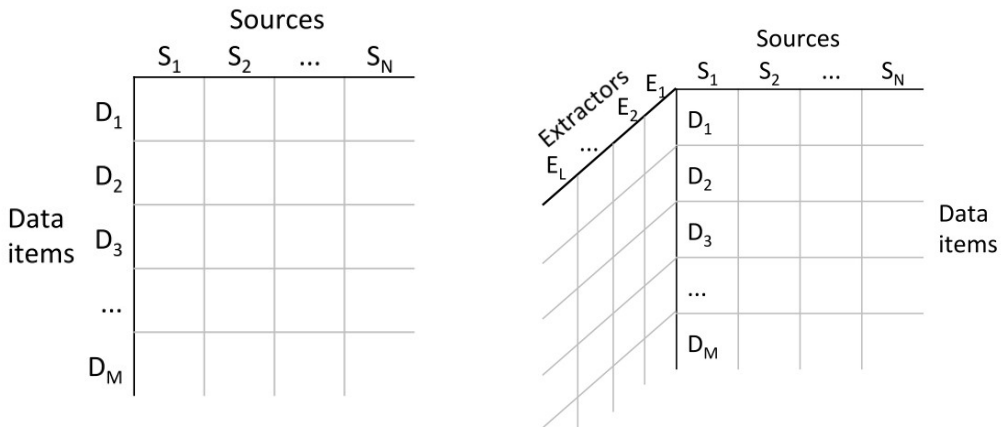
Temporal Knowledge *Refiners*: This group of *refiners* identify the facts for given relations for different time points, which observe data in the dynamic world. Several methods have been proposed for extracting temporal knowledge, such as [98–102]. As for temporal

knowledge, we need to extract the validity time of facts additionally, the solutions are much more complex. We will first focus on the static world, improving the refiners for extracting temporal knowledge will be our future work.

According to the above analysis, the knowledge extraction techniques involve three key steps: triple identification (identifying which part of the data indicate a predicate and its value); entity linkage (linking any entities that are mentioned to the corresponding entity identifier); and predicate linkage (linking any relations that are mentioned to the corresponding knowledge base schema). All three steps are error prone, and the coverage and accuracy of the existing *refiners* are not sufficient.

2.3 Overview of Truth Discovery

For knowledge extraction tasks, such as slot filling [103] and entity profiling [104], related data can be collected from various corpora and multiple refiners can be applied to extract desired information. The outputs of different extractors can be conflicting, therefore, how to fuse these data and store the true values becomes one of the most important topics in database and knowledge management communities [105–107]. In order to solve this issue, researchers setup a scenario that assumes there are M objects, each representing a certain properties of an entity (e.g., the profession of Xiu Fang), and N data sources (e.g., Web Sources). In this case, we can refer to the raw data as an $M \times N$ data matrix (see Figure 2.2a). Due to the facts that some sources may not provide certain objects and the scale of the raw data is huge, this data matrix is sparse and large. Each row represents multiple values for a certain object, and the values in the row are conflicting ones. Based on this scenario, researchers named the problem as data fusion (or truth discovery), of which the goal is to identify the true values for each row, given the noisy observations in the data matrix, while deciding the quality and correlations of the data sources (columns). Recently a new problem called knowledge fusion



(a) Data Fusion Input (b) Knowledge Fusion Input

Fig. 2.2 Input Comparison for Data Fusion and Knowledge Fusion

has been introduced, the goal of which is to calculate the probabilities of the truthfulness of the extractions extracted from multiple sources by different *refiners* [20, 41]. In this case, we can consider the scenarios as adding a third dimension to the data matrix (see Figure 2.2b). This problem is out of the scope of our work, we may study this research issue in our future work.

For truth discovery, the *primitive methods* are typically *rule-based*, such as: i) regarding the latest edited values as true; ii) conducting *majority voting* (for categorical data), i.e., predicting the values with the highest number of occurrences as the truth; iii) naively taking the *mean/minimum/maximum* as the true values (for continuous data). These methods focus on improving the efficiency in database queries [108, 109], see [110, 111] for surveys, but they show low accuracy for cases that many sources provide low quality data, due to the fact that the sources may not be equally reliable [112]. Yin et al. [44] first formulate the truth discovery problem in 2008. Since then, many advanced solutions have been proposed by applying unsupervised or semi-supervised learning techniques while additionally taking various implications of multi-source data into consideration (see [47, 37, 113, 48, 49] for

surveys). According to the models the existing truth discovery methods adopt, we roughly classify them into five categories[6], to be discussed in the following sections.

2.3.1 Web-Link Based Methods

These methods are basically inspired by *PageRank* [89], which uniformly assign weights to each outgoing links based on the number of those links of a vertex in a graph, and evaluate page worthiness based on the link structure of the Web. They conduct random walks on the bipartite graph between sources and values of objects. They measure webpage authority based on their links to the claimed values, and estimate source reliability and value correctness based on the bipartite graph. For example, *Sums* [43] employs the idea of *authority-hub analysis* proposed by Kleinberg et al. [91]. It lacks accuracy because it overestimates the sources that make larger coverage of objects. To overcome this disadvantage, *Average-Log*, *Investment*, and *PooledInvestment* [43, 114] have been proposed, each of which applies different calculation to assess source reliability. Specifically, Average-Log uses a non-linear function to assess sources. Investment conducts in a way that each source uniformly invests/distributes its reliability to the values it provides, while sources collect credits back from the confidence of their claimed values. PooledInvestment follows a similar procedure with Investment, except it uses a linear function to estimate the confidence of values instead of the non-linear function. *SSTF* (Semi-Supervised Truth Finder) [115] solves truth discovery in a semi-supervised manner based on a small set of labeled true values. It also incorporates both mutual exclusivity for categorical data and mutual support for continuous data to capture the relations among multi-source data.

[6]Note that there are overlaps among those categories. For example, Investment belongs to both Web-link based methods and iterative methods.

2.3.2 Iterative Methods

For iterative methods, value veracity and source reliability are iteratively calculated from each other until certain convergence condition is met [44, 43, 36]. This type of model is easier to understand and interpret than the other models (to be detailed in the following subsections), all of which can be conducted in an iterative manner to achieve better accuracy. For example, the coordinate descent in optimization based methods (Section 2.3.5) and the parameter inference in probabilistic graphical model based methods (Section 2.3.4) require iterative updating, while iterative methods can be reformulated as a parameter inference or an optimization task [49]. There are several representative iterative methods. For example, *TruthFinder* [44] iteratively estimates *trustworthy of source* and *confidence of fact* from each other and additionally considers the *influences between facts*. However, its special definition of *confidence of fact* for simplifying the computation leads to overestimation of this measurement. Inspired by similarity measurements in *Information Retrieval*, Galland et al. proposed a series of methods including *Cosine, 2-Estimates, 3-Estimates* [36]. They all take mutual exclusion for categorical data into consideration, while 3-Estimates additionally incorporates *hardness of fact* to improve 2-Estimates.

2.3.3 Bayesian Point Estimation Methods

This type of methods adopts *Bayesian analysis* to compute the maximum a posteriori or *MAP* value for each object. Dong et al. [38, 65] proposed a series of Bayesian methods, including *Depen, Accu, AccuPR, AccuSim, AccuFormat, AccuNonUni, PopAccu*, based on the single-valued assumption by considering the copying relations among sources. They measure the trustworthiness of a source by its accuracy, which essentially indicates the probability of each of its values being true. In particular, Depen estimates value veracity while conducting copy detection. Accu improves Depen by relaxing the equal source reliability assumption. It assumes there are N uniformly distributed false values and only one true value for each

object. AccuPR augments Accu by additionally considering the probability of a value being true. AccuSim improves Accu by tackling value similarity. AccuFormat further augments AccuSim by considering value formats. AccuNonUni relaxes the assumption made by Accu that false values are uniformly distributed. Different from their previous methods, PopAccu computes value distribution from the observed data. To solve the MTD problem, Wang et al. [116] proposed a multi-truth Bayesian model (*MBM*).

2.3.4 Probabilistic Graphical Model Based Methods

Truth discovery methods in this category apply probabilistic graphical models to jointly reason about source trustworthiness and value correctness. They make strong assumptions about prior distributions for the latent variables [117, 118, 46, 119], rendering their models inhibitive and intractable to incorporating various implications to improve their performance. Moreover, Waguih et. al conclude with extensive experiments that this type of methods is generally of poor scalability in [48]. There are several representative methods belong to this type. For example, *GTM* (Gaussian Truth Model) [118] is specially designed for continuous data. In *LCA* (Latent Credibility Analysis) [37], source reliability is modeled by a set of latent parameters. It enriches the meaning of source reliability by tackling the difference between telling the truth and knowing the truth. There are four models in their work, namely *SimpleLCA, GuessLCA, MistakeLCA, LieLCA*. In *LTM* (Latent Truth Model) [46], both false positives and false negatives are considered. They measure sources in terms of precision and recall, making LTM capable to discover multiple true values simultaneously for each object. *IATD* (Influence-Aware Truth Discovery) [120] is an unsupervised probabilistic model, which takes source correlations as prior for influence derivation. This method is applicable for both categorical data and continuous data.

2.3.5 Optimization Based Methods

Truth discovery can also be formulated as an optimization problem. Several methods are based on this formulation: *CRH* (Conflict Resolution on Heterogeneous Data) [40] is proposed for tackling heterogeneous data, in which different types of distance functions can be plugged in to capture the features of various data types, such as categorical data and continuous data. The goal of CRH is to minimize the weighted deviation of the multi-source data from the predicted true values; *MLE* (Maximum Likelihood Estimation) [121] is based on the *EM* (Expectation Maximization) algorithm [122] to quantify source reliability and value correctness. It only deals with Boolean positive claims, and ignores the negative claims; *CATD* (Confidence-Aware Truth Discovery) [123] is designed specially for continuous data with the awareness of long-tail phenomenon; a recent work [124] considers the inherent correlations among entities for truth discovery in crowd sensing scenarios. This work addresses the truth discovery problem via a weight squared minimization model, where the effect of correlations is modeled as a regularization term.

2.4 Overview of Multi-Valued Truth Discovery

Despite active research in the field, multi-valued truth discovery is rarely studied by the previous work. LTM (Latent Truth Model) [46], a probabilistic graphical model based method, is the first solution to the MTD. In this work, Zhao et al. measure two types of errors (false positive and false negative) by modeling two different aspects of source reliability (*specificity* and *sensitivity*) in a generative process. Pochampally et al. [41] study various correlations among sources by taking information extractors into consideration. To rebalance the distributions of positive claims and negative claims and to incorporate the implication of values' co-occurrence in the same claims, Wang et al. [125] propose a probabilistic model that takes multi-valued objects into consideration. Waguih et al. [48] conclude with extensive

experiments that these probabilistic graphical model-based methods cannot scale well. Zhi et al. [126] also consider the mutual exclusion between sources' positive claims and negative claims. They model the silence rate of sources to tackle the possible non-truth objects rather than multi-valued objects. To relax unnecessary assumptions, Wang et al. [116] analyze the unique features of MTD and propose an MBM (Multi-truth Bayesian Model), which incorporates source confidence and finer-grained copy detection techniques in a Bayesian framework. Recently, Wang et al. [127] design three models (i.e., the *byproduct* model, the *joint* model and the *synthesis* model) for enhancing existing truth discovery methods. Their experiments show that those models are effective in improving the accuracy of multi-valued truth discovery using existing truth discovery methods. Wan et al. [128] propose an uncertainty-aware approach for continuous data where the number of true values is unknown.

2.5 Other Work Related to Truth Discovery

Besides the basic truth discovery issue, many advanced issues have been actively studied. For example, the relation-based truth discovery methods additionally take the relationships between sources into account. There are two types of the relation-based methods: the first type only considers the copying relationships between pairs of sources [129, 130, 45, 38, 131, 132]. They assign a discounted vote count for the values provided by copiers. The second type comprehensively considers the complex relationships among a subset of sources [41, 113, 120], such as positive correlations and negative correlations. The attribute decomposition methods [133, 134] differentiate sources' quality on different (groups of) attributes of the same object. In [135], a problem scale reduction framework is proposed to improve truth discovery efficiency. An assembling method [136] that combines the results of multiple existing methods to deliver better results. The probabilistic [137], the factorization [138], and the optimization-MAP combination [139] approaches for discovering truth incrementally from data streams. A Hidden Markov Model (HMM) [131] for truth dis-

covery from sources' update history. A privacy-preserving truth discovery framework [140] for privacy-preserving in truth discovery tasks. *ETCIBoot* (Estimating Truth and Confidence Interval via Bootstrapping) [141] for the real-world where confidence interval estimation of truths is more desirable than point estimation. Despite the research efforts conducted on improving the accuracy and efficiency of truth discovery methods, nowadays, truth discovery methods have been successfully applied in many real-world applications. While the source reliabilities estimated by truth discovery can be used to access the quality of web-pages [142], truth discovery techniques can also be utilized in the following areas.

Healthcare. People post reviews about various drugs in online health communities. This user-generated information is valuable for both patients and physicians. However, the quality of such information is a big issue to address. Mukherjee et al. [28] adopt the truth discovery technique to automatically find reliable users and identify trustworthy user-generated medical statements.

Crowdsourcing aggregation. Crowdsourcing platforms such as Amazon Mechanical Turk[7] provide a cost-efficient way to collect labels from crowd workers. However, workers' capabilities are quite diverse, which leads to the important task of identifying true labels from the labeling efforts of multiple workers [143–152]. Thus crowdsourcing aggregation approaches focus on learning true labels or answers to certain questions. The main difference between crowd-sourcing aggregation and truth discovery is that the former is an active procedure (one can control what and how much data to be generated by workers) while the latter is a passive procedure (one can only choose from available data sources).

Social sensing. With the explosive growth of online social networks, users can provide observations about physical world for various social sensing tasks, such as gas shortage report after a disaster, or real-time information summarization of an evolving event. For these participatory social sensing tasks, users' information may be unreliable. Recently, a

[7] https://www.mturk.com/mturk/welcome

series of approaches [153–158, 121] have leveraged truth discovery methods to improve the aggregation quality of such noisy sensing data.

2.6 Overview of GrandBase Construction

In our system, knowledge base construction involves two main phases, namely *knowledge extraction* and *truth discovery*. Generally, the knowledge extraction phase contains three tasks [20]: *triple identification*, *entity linkage*, and *predicate linkage*. Due to the diverse reliability of different sources and the varied capacity of various extractors, it is common to observe conflicts in the extracted triples. The truth discovery phase aims at reconciling those conflicts. Fig. 2.3 shows an overview of the framework for GrandBase construction. This section is based on our research reported in [21, 159].

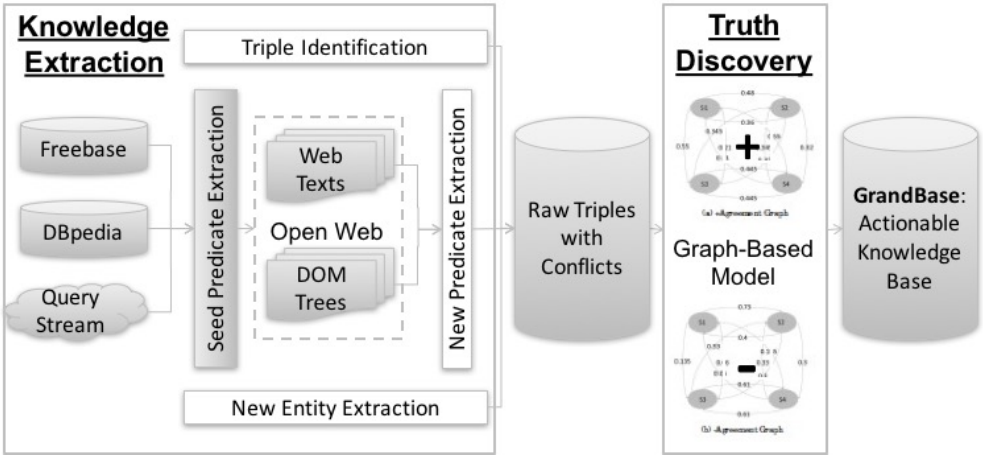


Fig. 2.3 The framework of GrandBase construction: white rectangles with underlined labels represent the two main phases of GrandBase construction, the three small white rectangles inside the knowledge extraction rectangle depict the three tasks of knowledge extraction.

Specifically, for the knowledge extraction phase, we apply the open IE approach to extract RDF triples from four types of sources including query stream, existing KBs (Freebase and DBpedia), Web texts and DOM trees. We will report this work in Chapter 3. To construct a more complete KB, we propose to augment the ontology of Freebase, since Freebase

contains the largest number of entities and *isA* pairs. In particular, we change the traditional tasks of knowledge extraction, namely *predicate linkage* and *entity linkage*, to *new predicate extraction* [16] (i.e., discovering new predicates from the Web content and attaching them to the corresponding classes to augment the ontology) and *new entity extraction* [160] (i.e., identifying new entities described in the Web content and attaching them to the corresponding classes to augment the ontology). For new predicate extraction, since the data in query stream and existing KBs would be more accurate, we first extract predicates from those sources. Then, we utilize the extractions as seed to learn extraction patterns of the open Web (Web texts and DOM trees), which are in turn used to extract more new predicates from the Web. Due to the differed features of Web texts (often presented by natural languages) and DOM trees (semi-structured data described by tags), we apply different extractors on them. In particular, as Web text extraction has been widely studied, we focus on DOM tree extraction in our work. For Web texts, we first perform standard natural language processing (NLP) technique, then apply distant supervision to induce lexical and parse patterns, which are unified syntax rules over the Web, and finally leverage these patterns to extract predicates from Web texts. For DOM tree extraction, since Web sites are different from each other in display style and format, no unified *tag path pattern* could be found that is applicable to all the Web pages. To this regard, our extractor learns tag path patterns for each Web page and then uses these patterns to extract new predicates from the Web pages. There is a work [160] related to new entity extraction in the literature, which jointly solves entity-linking and entity-discovery, our framework seeks to incorporate this technique to broaden the amount of entities accommodated in GrandBase. To further enhance the ontology, we also conduct misspelling, synonym, sub-predicate identification [16]. Finally, we propose to apply this enhanced ontology to explore more facts from the open Web[8].

At the truth discovery phase, our work relaxes the single-valued assumption commonly made by the previous work and tackles a more general problem, i.e., multi-valued truth

[8]New entity extraction and triple identification will be the focuses of our future work.

discovery (MTD). Given a multi-valued predicate, the value sets provided by sources may be the same, totally different, or overlapping. Sources may cautiously provide partial true values and omit the values they are not sure about, or audaciously provide all potential values, even if the veracity of the claimed values is uncertain. To differentiate the cautious and audacious sources and be aware of false positives and false negatives, we propose to measure source reliability by *positive precision* and *negative precision*. Given a source, the positive (resp., negative) precision represents the probability of the positive (resp., negative) claims being true (resp., false). Intuitively, if the positive (resp., negative) claims of a source are agreed by the majority of other sources, this source is likely to have high positive (resp., negative) precision. This means that the inter-source agreements indicate source reliability endorsement. This intuition motivates us to measure the two-sided source reliability by quantifying the two-sided agreements among sources regarding their positive claims and negative claims. We design graph-based approaches to fusing the conflicts in the raw triples extracted by the extractors at the knowledge extraction phase. Our graph-based approaches for MTD will be reported in Chapter 4 and Chapter 5.

2.7 Summary

In this chapter, we have introduced some background knowledge, state-of-the-art techniques, and research challenges regarding big data integration and knowledge bases. In particular, we overviewed the extraction techniques for extacting data in user required structured formats from the Web (i.e., Web data refiners), and the methods for extracting RDF data from structured and unstructured sources (i.e., knowledge refiners). Then, we discussed the research efforts on truth discovery. We classified the existing methods into five categories, and introduced state-of-the-art MTD methods that are more close to our work. For completeness, we also presented some interesting recent works on this research topic. Finally, we overviewed

the framework for GrandBase construction. From next chapter, we will introduce our works that related to the concepts demonstrated in this chapter.

Chapter 3

Attribute Extraction for Knowledge Base Expansion

In this chapter, we introduce our approach for ontology augmentation via attribute (or predicate) extraction from multiple types of sources. As a means to enhance existing ontologies, *attribute extraction* has attracted tremendous research attention. However, most existing attribute extraction techniques focus on exploring a single type of sources, such as structured (e.g., relational databases), semi-structured (e.g., Extensible Markup Language (XML)) or unstructured sources (e.g., Web texts, images), which leads to the poor coverage of knowledge bases (KBs). To this regard, we present a framework for *ontology augmentation* by extracting attributes from four types of sources, namely existing knowledge bases (KBs), query stream, Web texts, and Document Object Model (DOM) trees. In particular, we use query stream and two major KBs, DBpedia and Freebase, to seed the attribute extraction from Web texts and DOM trees. We specially focus on exploring the extraction technique from DOM trees, which is rarely studied in previous works. Algorithms and a series of filters are developed. Experiments show the capability of our approach in augmenting existing KB ontology. This chapter is based on our research reported in [34].

Table 3.1 Statistics of Representative KBs

KB	# Entities(million)	# Attributes
YAGO	10	100
DBpedia	4	6,000
Freebase	25	4,000
NELL	0.3	500

3.1 Overview

With the sheer amount of data produced and communicated over the Internet and the Web in the last few years, the Web has gradually evolved into a huge information repository with hidden knowledge. To explore this knowledge, researchers have developed various extraction techniques (i.e., *extractors*) to augment ontologies or enhance existing knowledge bases (KBs). While the existing ontologies have already included a wide range of entities, the number of attributes contained in these KBs is still small (see Table 3.1 for some statistics we have done). For example, Freebase has 25 million entities, but only 4,000 attributes. The type *University* in Freebase (note that in Freebase, classes are referred to as *types* and attributes are referred to as *properties*)[1] has only 9 properties, while in reality we can easily identify more attributes. For this reason, it becomes urgent to find more attributes of classes for ontology augmentation.

Although tremendous previous efforts have been conducted, they mostly extract attributes from a single type of Web sources, such as Web texts (e.g., [25]), Web tables (e.g., [24]) DOM trees (e.g., [22, 23]), or relational databases. These approaches often lead to a poor coverage of the results.

There are generally three challenges. First, the single pattern (e.g., "*what is the A of E*" or "*the A of (the/a/an) E*") used by previous attribute extraction systems [16, 161] no longer applies due to the poor coverage and incapability of filtering noisy inputs (such as "*what is*

[1]Hereafter, we will use the terms class and type, attribute and property interchangeably.

the plural of apple"). Second, since the same entity may have different attributes in different KBs, the attributes should be consolidated for better usage. To the best of our knowledge, there is no previous work on merging the attributes from different KBs. Third, as the DOM trees of Web pages may differ from one and another, it is tricky to develop a generic solution to exploring new attributes from such sources.

In our work, we propose more comprehensive attribute extraction from four types of sources, namely existing KBs (Freebase and DBpedia in our case), query stream, Web texts, and DOM trees, to resolve the above challenges. In a nutshell, this work makes the following contributions:

- We propose a novel framework that extracts and merges the attributes from four types of sources, KBs (Freebase and DBpedia) and query streams, Web texts, and DOM trees, for comprehensive ontology augmentation.

- We develop an improved query stream extraction technique that adopts new patterns and filtering rules to improve the coverage and quality of the extractions.

- We develop an algorithm for extracting attributes from DOM trees. We use the attributes extracted from the query stream and existing KBs as seeds to learn the tag path patterns from Web pages, and use these patterns to extract more attributes from the DOM trees.

The remainder of this chapter is organized as follows. Section 3.2 gives a brief overview of the related work. Section 3.3 demonstrates our approach. Specifically, Section 3.3.1 introduces our framework and particularly compares with a recent research work named Biperpedia [16]. The methods for merging the attribute extractions from Freebase and DBpedia are also described in this section. Section 3.3.2 presents our method for query stream extraction. Section 3.3.3 describes the algorithm for extracting attributes from DOM

trees using the above extractions as seeds. Finally, Section 3.4 reports the experimental results, and Section 3.5 provides some concluding remarks.

3.2 Related Work

In this section, we overview the related work on attribute extraction, particularly the approaches on multiple types of sources and DOM trees.

3.2.1 Extracting Attributes from Multiple Types of Sources

While attribute extraction has been widely studied in recent years, few works have been conducted to extract attributes from multiple types of data sources. Pasca et al. [162] are the first to exploit attributes from query streams. By using a head-to-head qualitative comparison, they conclude that extracting attributes from query stream achieves 45% higher accuracy than that from Web texts. Based on this insight, Pasca et al. [163] extract attributes from both query logs and query sessions, and Kopliku et al. [24] combine extractions from structured data sources including Web tables, search hit counts, Wikipedia, and DBpedia.

Comparing with above efforts, our approach is the first to extract attributes from four different types of sources, namely *query stream*, *Web texts*, *DOM trees*, and *existing KBs*. Our work is inspired by a very recent work conducted by Gupta et al. [16], which proposes a novel ontology named Biperpedia. We will discuss the differences between Biperpedia and our work in Section 3.3.1. Another related work is proposed by Lee et al. [161], which extracts attributes from query logs, Web documents, and external KBs independently to compute the typicality for a class (resp. attribute) given an attribute (resp. class). In contrast, our system extracts attributes from DOM trees and Web texts seeded by the attributes extracted from query stream and two major KBs (i.e., Freebase and DBpedia).

3.2.2 Extracting Attributes from DOM Trees

Extracting attributes from DOM trees is not completely new. Early supervised approaches [164, 165] use manually defined wrappers to extract attributes from each Website, which are time-consuming and non-scalable. Wrapper learning techniques (e.g., [86] proposed by Turmo et al.) can help reduce human intervention, but additionally requires labeled data for the training, and are inapplicable to new websites that have not been handled before. Generative models designed in [166] alleviate this problem by segmenting and labeling the training samples, but they can only extract the attributes that are predefined in the training data. Interactive learning techniques developed by Irmak et al. [82] and Kristjansson et al. [167] can also help reduce human efforts on preparing the training data, but they are still not automated.

Unsupervised methods include *template-based* methods and *pattern-based* methods. The template-based methods, represented by RoadRunner (designed by Crescenzi et al. [168]) and EXALG (developed by Arasu et al. [169]), detect website-specific templates to extract attribute values. The pattern-based methods proposed by Liu et al. [22] and Bing et al. [23] extract data records from a single list page, based on some patterns that repeatedly occur in multiple data records. Both methods however require re-implementation for new websites. Comparing with previous works, our approach enables more accurate and extensive attribute extraction from DOM trees automatically.

3.3 The Extraction Approach

3.3.1 The Framework for Ontology Augmentation

Our framework (see Figure 3.1) contains two main phases: *attribute extraction* and *ontology enhancement*. At the *attribute extraction* phase, we extract attributes from DBpedia and Freebase, which are then combined.The details will be introduced in the remainder of this section. We also extract attributes from query stream (see Section 3.3.2). The resulting

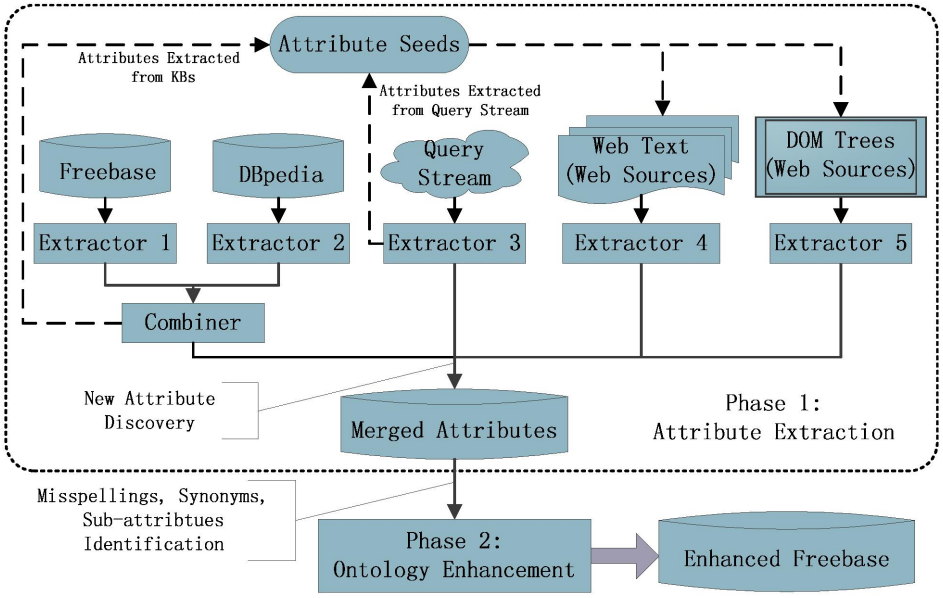


Fig. 3.1 The Framework for Ontology Augmentation

attributes will be used as seed to extract more attributes from the open Web. We adopt different methods to deal with Web texts and DOM trees. For Web texts, we first perform standard natural language processing (NLP), and then apply distant supervision to induce lexical patterns and use these patterns to extract more attributes. For DOM trees, we learn the tag path patterns and define several filters to extract attributes (see Section 3.3.3). At the *ontology enhancement* phase, we identify the misspellings, synonyms, and sub-attributes within the extracted attributes. Because our work focuses on the first phase, we simply reuse the methods in Biperpedia [16] for the ontology enhancement. Specifically, we use search engines to identify misspellings, a Support Vector Machine (SVM) to identify synonyms, and two heuristics[2] to identify sub-attributes.

Attribute extraction from Freebase and DBpedia. We use two dominant KBs, Freebase and DBpedia, for attribute extraction. Since Freebase contains the largest number of entities and `isA` pairs, we use Freebase as the basis for ontology augmentation.

[2] i) The former includes a modifier over the latter, such as *English teacher* and *teacher*. ii) The two attributes have relation like "*A1 is a A2*", e.g., *supervisor* is a *teacher*.

Table 3.2 Statistics of Five Representative Classes

Class	**# Attributes**				
	DBpedia	**Extrac. (DBpedia)**	**Freebase**	**Extrac. (Freebase)**	**Combine (Freebase &DBpedia)**
Book	21	48	5	19	60
Film	53	53	54	54	92
Country	191	360	22	150	489
University	21	484	9	57	518
Hotel	18	216	7	56	255

Intuitively, a sub-type should inherit all the properties of its super-type, so for every type/class in Freebase/DBpedia, we iteratively attach to the type/class all its super-types'/super-classes' properties, as well as the names, labels, and descriptions/comments for these properties.

Combining attribute extractions from DBpedia with Freebase. We combine the two KBs by attaching the attributes of every DBpedia class to its similar types in Freebase. By similar types, we mean the types with synonymous names of the class, or the types that have high overlaps (e.g., more than 50%) with the class in their contained entities. To avoid redundancy, we compare the attributes of Freebase and DBpedia in terms of name, label, and comment to determine if they are actually the same. For the attributes that have no comments/descriptions, we solely rely on their names, but leave the development of methods for annotating these attributes as our future work. Table 3.2 shows that our approach obtains more attributes for all the five representative classes in Freebase.

Comparison with Biperpedia. Our work distinguishes from the most similar approach, Biperpedia [16], in three aspects. First, we fuse existing KBs (Freebase and DBpedia), instead of using a single KB—as what Biperpedia does, for attribute extraction. Second, we define filters and more practical patterns for query stream extraction. Third, while Biperpedia

regards Web tables meaningless, the value of Web tables for attribute extraction has been proved by many works (e.g., [24]). For this reason, besides Web texts, we additionally extract attributes from DOM trees of which Web tables are regarded as a sub-type.

3.3.2 Seed Extraction from Query Stream

Query stream is useful for attribute extraction because it naturally reflects users' collective convictions on possible attributes of entities. To extract attributes from query stream, for each type, T, in Freebase, we conduct an iterative procedure consisting of five steps, which will be discussed in this section.

The first step focuses on identifying relevant query stream. To do so, we apply an *entity recognizer* to identify the queries that contain entities of T and these queries as regarded as the relevant query stream of T. Then in the second step, we identify attribute candidates. We particularly exploit a set of predefined patterns, such as "*what/how/when/who is the A of (the/a/an) E*", "*the A of (the/a/an) E*", and "*E's A*", to extract attribute candidates from the relevant query stream. For example, given relevant queries, namely "who is the director of Taken 3", "the release date of Taken 3", and "Taken 3's box office", we can identify that *director*, *release date* and *box office* as the attributes of type *Film* in Freebase according to above patterns. We denote the set of all the identified attributes as *Attri*, and the set of queries that match the above patterns as $Selected\_QUERY$.

The third step focuses on filtering out non-attributes because the *Attri* set may be noisy. For example, the query text "*The University of Adelaide*" matches one of our patterns, but "*University*" is not actually an attribute of "*Adelaide*". Similarly, for the query "*the plural of country*", "*plural*" is not an attribute of "*country*". We develop the following two rules to handle such cases.

Rule 1: We use a *blacklist* to exclude the generic attributes that appear in multiple different types. First, we select a set of dissimilar types from the query stream, and rank the attributes

by the number of dissimilar types they belong to. We then add the top attributes to the blacklist to avoid their appearance in *Attri*. In this way, the words such as "*lack*", "*rest*", "*best*", "*meaning*", "*summary*", "*definition*" and "*plural*" can be excluded from *Attri*.

Rule 2: We match each query in $Selected\_QUERY$ with the entities of T in Freebase to resolve the false negatives caused by long named entities. Specially, some queries with capitalized initials can be directly recognized as named entities and removed directly from $Selected\_QUERY$. For example, "*Toyota*" is not an attribute of "*Glendale*" in "*the Toyota of Glendale*". For this case, we directly remove this query from $Selected\_QUERY$.

The fourth step is to identify entity-attribute pairs. For the relevant query stream of T, if a query contains both entity, denoted as E, of T in Freebase and attribute (denoted as A, $A \in Attri$), we keep the corresponding (A,E) pair in a set denoted as *AEpair*. For now, we only consider simple queries containing single entity and attribute. We will deal with complex queries with multiple entities and multiple attributes in our future work. The final step focuses on identifying credible attributes. To further improve the quality of the extractions, we first define the following three functions that will be used:

- *EntityNumber(T)*: The number of entities contained by T in Freebase.

- *EntityDiversity(T, A)*: The number of distinct entities co-appears with an attribute A in *AEpair* . We employ the standard co-reference resolution algorithm [170] to identify all the (A,E) ($\in AEpair$) pairs where A and E co-refer to the same entity, and delete all such pairs to reduce the redundancy.

- *EntityFrequency(E)*: The number of queries that contain E in the relevant query stream of T.

We develop two more rules to further clean the extractions as the following:

Rule 3: Given a type T containing entities $\{E_1, E_2, \ldots, E_k\}$ and an attribute A, A will not be attached to T, if $\exists E_j \in \{E_1, E_2, \ldots, E_k\}$, *EntityFrequency(*$E_j$*)* $\geq \max_{(A,E*)\in AEpair}$ (*EntityFrequency (E*)*) and $(A, E_j) \notin AEpair$.

Rule 4: For each *EntityDiversity(T, A)* $\neq 0$, we remove the (T, A) pair, if $\frac{EntityDiversity(T,A)}{EntityNumber(T)} \leq \alpha$ (α a pre-defined threshold).

After the filtering, we finally obtain the credible attributes of type T in the form of $(A, EntityDiversity(T, A))$.

3.3.3 Extraction from DOM Trees

Web pages are typically semi-structured and described by nested HTML tags. The tree-like structures can be commonly found in Web pages that contain Web lists, Web tables, as well as deep-Web sources, and are referred to as DOM trees[3]. Traditional extractors simply remove the tags and extract data from the plain texts. Thus, they fail to exploit the knowledge contained in the DOM trees. In this work, we introduce a two-step approach to extract attributes from DOM trees for ontology augmentation. We first extract additional attributes from the DOM trees seeded by the attributes extracted from query stream and existing KBs (denoted by *SEED_SET(T)*). We then define a set of filters to refine and differentiate the attributes.

Original Extraction from DOM Trees

Different from attribute extraction from Web texts, where lexical and parse patterns can be learned and used all over the Web, extracting attributes from DOM trees is more tricky because different Websites have different styles and formats, and the tag path patterns extracted from one Web page can hardly be applied to another page. To resolve this challenge,

[3]http://www.w3.org/DOM

our approach alternatively extract attributes and learns tag path patterns through an iterative process. The detailed procedure is described in Algorithm 1.

Briefly, given a type T, the algorithm first identifies the Websites related to T (e.g., http://www.imdb.com/ for type *Film*). For each Web page, the algorithm analyzes the DOM structure and classifies the text nodes into *entity node* (the texts represent the name of an entity E of T) and *non-entity node*. The tag paths between each *entity node* and their corresponding *non-entity node* are then extracted, removed of noisy tags, and kept in a *tag path set*. For each Website, the algorithm iteratively finds out Web pages that contain at least one (A,E) pair, where E is an *entity node*, A is the content of a *non-entity node* and $A \in$ *SEED_SET(T)*. For each Web page, the algorithm traverses the *tag path set* for this Web page to obtain the tag paths between the seed A and E, and transfers these tag paths from the *tag path set* to an *induced tag path pattern set* for this Web page. We next compare all the tag paths in the *tag path set* with the patterns in the *induced tag path pattern set*. Those *non-entity nodes* with tag paths that are similar with the induced patterns are finally recognized as new attributes, and are added to *SEED_SET(T)*, with the corresponding tag paths removed from the *tag path set*.

The algorithm turns to another Website when the number of attributes in *SEED_SET(T)* reaches a certain threshold. Since the number of Web pages and text nodes in a Web page are limited, the algorithm can always terminate with an output.

Extraction Filtering

Similar to extractions from query stream, the attributes obtained by the first step may contain noises due to the open nature of Web content. We therefore employ the following three types of features to refine the extracted attributes:

- The inherent features of attribute: A node that denotes an attribute in a DOM tree always follows some inherent rules, e.g., the text node always contains a colon as the end of the string, and the length of the text is always limited to a certain number.

Algorithm 1: Algorithm for DOM Tree Extraction

Input: *Type* T_k in Freebase; a set of Websites regarding to T_k, S=$\{S_1, S_2, \ldots, S_n\}$, for each Website $S_j \in S$, it contains a set of Web pages, P_j=$\{P_{j_1}, P_{j_2}, \ldots, P_{j_m}\}$,$j_m$ is the number of Web pages belong to S_j; the entity set Set_E of T_k in Freebase; the seed attribute set A_{T_k} extracted from query stream and existing KBs for T_k

Output: Original attributes for Type T_k in Freebase (i.e., enriched A_{T_k}).

1 **Initialization**: identify all the *entity node* and *non-entity node* in every Web pages, and obtain *tag path set*(denote as Tagpath) for each Web page, e.g., for $P_{j_l} \in P_j$, we keep a set of tag paths Tagpath(P_{j_l}).

2 **for** *each* $S_j \in S, j = 1, 2, \ldots, n$ **do**

3 **for** *each* $P_{j_i} \in P_j, i = 1, 2, \ldots, j_m$, *and* P_{j_i} *contains at least an entity* $E \in Set_E$ *and an attribute* $A \in A_{T_k}$ **do**

`/* if` $|A_{T_k}|$ `is increased, the algorithm continues the loop for this Website; else the algorithm begins to traverse another Website */`

4 extract the tag path(s) between E and A, and transfer them to the *induced tag path pattern set*;

5 compare all the other tag paths $\in$ Tagpath(P_{j_i}) with the induced tag path(s) in *induced tag path pattern set*;

6 **if** *(a tag path is similar to the induced tag path(s))* **then**

7 add the text of that *non-entity node* to A_{T_k};

8 remove the tag path from $Tagpath(P_{j_i})$;

- The intra-site features of attribute: If a Website contains an attribute, the attribute tends to appear frequently in a considerable number of pages of this Website.
- The inter-site features of attribute: Attributes tend to appear in multiple Websites instead of very few Websites.

We can simply remove the attributes that mismatch these features, but this may result in some loss of recall. For example, the number of movies that win an Oscar award are quite limited. Thus, the attribute "*winner in Oscar*" would not appear frequently in the Web pages of a movie Website, which could surely dissatisfy the second feature. We sequentially use three filters to deliver three attribute sets in turn, i.e., potential attributes, attribute candidates and credible attributes. Each set represents a different balance between the precision and recall of the extraction results and can be used by knowledge-driven applications based on

Table 3.3 Inherent Features of DOM Trees

Feature	**Description**
Word count	The number of words of an attribute should be no more than 10 (we discover that almost all attributes in Freebase is described by less than 10 words)
End with colon	The text ends with a colon should be the name of an attribute
The first letter of every word is in the upper-case	Web page always capitalizes the first letter for each word of the name of an attribute

their own requirements. We exploit the *inherent feature filter* to obtain the set of potential attributes by using the specific rules followed in the DOM trees, see Table 3.3 for some examples.

For the *intra-site feature filter*, we remove all attributes with intra-site frequency lower than a predefined threshold β to obtain the attribute candidate set. We calculate the frequency of an attribute A_i in a Website $S_j(i = 1,2,\ldots, j = 1,2,\ldots)$ by $f_j(A_i) = \frac{N(A_i)}{N(S_j)}$, where $N(S_j)$ is the number of Web pages of each Website, and $N(A_i)$ is the number of Web pages that contain attribute A_i.

Finally, the intra-site feature filter may incorrectly take some Web site-specific terms as attributes. For example, "*edit*" appears frequently in IMDb (a famous movie Website), but seldom contained by other Web sites. We remove such terms by examining the inter-site frequency feature of each attribute. Based on above discussion, we can obtain the credible attribute set by keeping only the attributes that appear evenly and frequently in many different websites. Specifically, we calculate the inter-site frequency of an attribute, and use a predefined threshold γ to exclude the attributes with low inter-site frequency.

Table 3.4 Entities in Representative Classes

Class	# Representative Entities	Examples of Entities
Book	1200	Asia Grace, Cool Tools
Film	1000	A Christmas Story, A Chump at Oxford
Country	727	Germany, Australia, Iran
University	1000	Brandeis University, Maynooth University
Hotel	1000	Hotel Sacher, Hotel Georgia

3.4 Experimental Evaluation

We implemented the proposed approach in Java and conducted some preliminary experimental studies using an ASUS P550C computer with a 2.5 GHz i7 processor and 8 GB RAM. In this section, we report our experimental results on attribute extraction from query stream and DOM trees.

3.4.1 Experiments on Query Stream Extraction

We conducted experiments on five representative types in Freebase, namely *Book*, *Film*, *Country*, *University*, and *Hotel*, to validate the capability of our approach for extracting attributes from query streams. For the entity recognition, each class is specified as a set of representative entities of Freebase (Table 3.4). Since our goal is to extract attributes rather than attribute values and the entities of the same class should share the same attributes, pre-specifying the target class by a set of entities will not be a limiting factor. Based on a query stream of 29,283,918 query records (which is the combination of two real-world datasets, Google[4] and AOL[5]), we finally obtained the extraction results as shown in Table 3.5).

We took the voting of three volunteers to determine the precision of the results. Volunteers manually gave their opinions on whether each attribute is reasonable for a class. The precision was calculated as the fraction of attributes that were labeled as reasonable. To measure the

[4]https://code.google.com/p/hypertable/downloads/detail?name=query-log.tsv.gz
[5]http://www.cim.mcgill.ca/~dudek/206/Logs/AOL-user-ct-collection/

Table 3.5 Query Stream Extraction Results

Class	Relevant Query Records	Credible Attributes	Precision (%)			
			Top-10	Top-20	Top-50	Top-100
Book	259,556	96	80	65	62	N/A
Film	403,672	59	100	75	66	N/A
Country	393,244	182	100	96	95	93
University	24,633	20	100	100	N/A	N/A
Hotel	15,544	N/A	N/A	N/A	N/A	N/A

precision, we ranked the attributes of a class by *EntityDiversity(T, A)* attached to each attribute. Specifically, we evaluated the top-k (k=10, 20, 50, 100) attributes for each class. The evaluation results (see Table 3.5) indicate that more relevant query records lead to more reasonable attributes. The precision of the top-k attributes peaks at k=10, but decreases as k increases. This is consistent with our assumption, that the attributes appear with various entities would be more credible. Although the results have shown good precision (60%$\sim$100%), the query stream used for the experiments is still relatively small. For this reason, we can hardly obtain good attributes for some classes. For example, the class of *Hotel* has only 15,544 relevant query records, and no reasonable attribute could be found for it. On the other hand, as relevant query dataset of *Country* was much larger, where we successfully obtained 182 credible attributes. It is reasonable to anticipate that more attributes can be extracted if larger datasets are available.

3.4.2 Experiments on DOM Tree Extraction

We also conducted experiments for the same five classes to study the extraction performance of our approach for DOM trees. As inputs, we used the merged extractions from existing KBs (Table 3.2) and query stream (Table 3.5). For the entity recognition, we also used the same entity dataset for each class (Table 3.4). We exploited crawler4j[6] to craw the Web and jsoup 1.8.1 to reformat the collected Web pages. We filtered out all the nodes with long text

[6]http://code.google.com/p/crawler4j/

Table 3.6 DOM Tree Extraction Results

Class	# Attributes				Precision (%)
	Query Stream	Existing KBs	Seed Attributes	DOM Trees	
Book	96	60	118	168	81.5
Film	59	92	121	329	88.6
Country	182	489	621	725	92.7
University	20	518	536	539	93.3
Hotel	N/A	255	255	312	79.8

(more than ten words in our case) to avoid tackling too many non-attribute nodes. Similarly, we used voting of three volunteers to determine the quality of resulting attributes (Table 3.6).

From Table 3.6, we can see that more attributes were extracted from DOM trees than from either query stream or existing KBs. More seeds tend to lead to more attributes extracted from DOM trees. The results also demonstrate a high precision achieved by our approach. This is reasonable because the information contained in DOM trees are often more structured and cleaner than that of Web texts. Clearly, DOM trees are a high-quality source for attribute extraction, which unfortunately are not considered in many recent research efforts in knowledge base construction such as Biperpedia [16].

3.5 Summary

In this chapter, we have proposed a framework for ontology augmentation by extracting attributes from four types of sources (existing KBs, query stream, Web texts, and DOM trees). We combine the attribute extractions from existing KBs (Freebase and DBpedia) and improve the existing query stream extraction methods by introducing new extraction patterns and filtering rules. We then apply these attribute extractions as seeds to induce extractions from the open Web (Web texts and DOM trees). While Web texts extraction has been widely studied, we focus on attribute extraction from DOM trees. To the best of our knowledge, this is the first approach to extracting attributes from DOM trees using open

information extraction techniques. Experimental results show that our system achieves more comprehensive yet still accurate ontology augmentation.

We have discussed the new predicate extraction of the knowledge extraction phase included by GrandBase construction. From the next chapter, we will discuss our methods on solving several truth discovery issues: in Chapter 4 and Chapter 5, we will present our graph-based approaches for MTD, as the solutions to truth discovery phase of GrandBase construction. After a comprehensive review of the existing truth discovery methods, we agree that a "one-fits-all" method is not achievable due to the limitations of the existing methods. We will propose an ensemble approach for better truth discovery in Chapter 6. In Chapter 7, we will introduce the bias introduced by sparse ground truth in evaluating the truth discovery methods. For the case where ground truth is missing, we make the attempt towards conducting evaluation without using ground truth.

Chapter 4

Multi-Valued Truth Discovery via Inter-Source Agreements

In this chapter, we propose a novel approach, *SourceVote*, to estimate value veracity for multi-valued objects. SourceVote models the endorsement relations among sources by quantifying their two-sided inter-source agreements. In particular, two graphs are constructed to model inter-source relations. Then two aspects of source reliability are derived from these graphs and are used for estimating value veracity and initializing existing truth discovery methods. Empirical studies on two large real-world datasets demonstrate the effectiveness of our approach. This chapter is based on our research reported in [171, 172].

4.1 Overview

In today's digital and connected world, we are experiencing the ever more freely created and published data on open sources every day. Those massive data on the Web hold the potential to revolutionize many aspects of our modern society, for example, enterprises can leverage these data to analyze the market and promote their products; government agencies can analyze these data for decision in security issues; researchers can study these data for

effective knowledge discovery. However, it is easy to observe that multiple sources often provide *conflicting* descriptions on the same objects of interest, due to typos, out-of-date data, missing records, or erroneous entries [28, 29, 31, 20, 21, 34], making it difficult to determine which data source should be trusted. Being misled by those conflicting data could lead to considerable damages and financial loss in many applications such as drug recommendation in healthcare systems or price prediction in the stock markets [35]. Moreover, due to the large-scale of data, it is unrealistic to expect a human to be able to manually determine which data is true. Therefore, a fundamental research topic named *truth discovery* has emerged as a fundamental research topic.

Considerable research efforts have been conducted to solve the truth discovery problem [36, 43, 44, 40, 45, 46]. Though these methods apply different formulas and models while incorporating different additional factors, they commonly assume that each object has exactly one true value (i.e., *single-valued* assumption). However, in real world, multi-valued objects—such as the children of a person, the authors of a book—widely exist. One may argue that previous methods under single-valued assumption (i.e., *single-valued method*s) can deal with multi-valued objects by simply regarding a value set, which may contain several values, claimed by each source as a joint single value, and determining the most confident value set as the truth. However, the value sets provided by different sources are generally correlated. There may be some overlaps between two sources' claimed value sets, indicating that they are not totally voting against each other. Neglecting this implication could degrade the accuracy of truth discovery. Moreover, single-valued methods overlook the important distinction between two aspects of quality, namely, false negatives and false positives, by measuring source quality using a single parameter, such as precision or accuracy. For multi-valued objects, some sources may provide erroneous values, making false positives, while some other sources may provide partial true values without erroneous values, making false negatives. Regarding these two types of errors as equivalent, the previous single-valued truth

discovery methods cannot distinguish the quality of those two types of sources. However, measuring source reliability by considering these two different types of errors is crucial to identify the complete true values for multi-valued objects.

To the best of our knowledge, few research efforts have been devoted to the multi-valued issue in the field of truth discovery. We identify the challenges of multi-valued truth discovery and the disadvantages of existing approaches as follows. Firstly, all existing methods require initializing source reliability, and for many of them, source reliability initialization impacts their performance in terms of convergence rate and accuracy. Secondly, sources providing some values in common indicates sources endorse one another. Intuitively, a source endorsed by more sources is regarded more authoritative and its provided values can be more trusted. This implication can be utilized to infer source reliability. Thirdly, while false positives and false negatives are equivalent for single-valued objects, for multi-valued objects, differentiating these errors is crucial for identifying the complete true value set. In a nutshell, our work makes three main contributions: i) we propose a graph-based model, called *SourceVote*, as a solution to the multi-valued truth discovery problem. It uses two graphs, i.e., $\pm$*Agreement Graph*, to model the two-sided endorsement relations among sources. Random walk computations are applied on both graphs to derive two-sided vote counts of sources and to finally estimate value veracity; ii) we further derive two-sided source reliability from the two graphs to better estimate sources' quality and initialize existing truth discovery methods; iii) we conduct extensive experiments on two large real-world datasets. The results show that SourceVote consistently outperforms the baselines.

The remainder of this chapter is organized as follows. Section 4.2 provides an overview of the related work. We describe the data model and formalize our research problem in Section 4.3. Section 4.4 presents in detail our SourceVote approach. We report our experimental results in Section 4.5, and provide some concluding remarks in Section 4.6.

4.2 Related Work

Since source reliability is the key to determining value veracity and existing truth discovery methods generally require source reliability initialization to launch their algorithm, more precise source reliability initialization is much in demand. Recent work adopts an external trustful source [15], a subset of labeled data [65, 132, 115], or the similarity among sources [103] as prior knowledge to initialize or help initialize the source reliability. To the best of our knowledge, SourceVote is the very first few to help source reliability initialization nearly without any prior knowledge.

The Web-link based truth discovery methods [91, 43] are the closest to our method. They compute the trustworthiness of sources and the truthfulness of values by using PageRank, where each link between a source and a value represents the source provides that value. However, they make single-valued assumption. To the best of our knowledge, *multi-valued truth discovery* is rarely studied by the previous work. LTM (Latent Truth Model) [46] and the method proposed by Wang et al. [125] are two probabilistic models that take multi-valued objects into consideration. Waguih et al. [48] conclude with extensive experiments that this type of models make strong assumptions on the prior distributions of latent variables, which render the modeled problem intractable and inhibitive to incorporating various considerations, and cannot scale well. Wang et al. [116] analyze the unique features of MTD and propose an MBM (Multi-truth Bayesian Model). However, they make strong assumptions on the copying of false information among sources and the independent provisioning of correct information by sources. It also requires initialization of several parameters including source reliability and copy probabilities of copiers. Recently, Wang et al. [127] design three models for enhancing existing truth discovery methods. Their experiments show that those models are effective in improving the accuracy of multi-valued truth discovery using existing truth discovery methods. However, LTM and MBM still performed better than those enhanced methods. None of the above methods takes the endorsement relations among sources into

consideration. Different from them, our approach assumes no prior distribution or source dependency and requires no initialization of source reliability. Therefore, it is robust to various problem scenarios and insensitive to initial parameters.

Table 4.1 Notations used in Chapter 4

Notation	Explanation
o, $\mathscr{O}$	An object (resp., set of all objects)
s, $\mathscr{S}$	A source (resp., set of all sources)
v, $\mathscr{V}$	A claimed value (resp., set of all claimed values)
$\mathscr{V}_o^*$	Identified truth for o
$\mathscr{V}_o^g$	Ground truth for o
$\mathscr{S}_o$	Set of sources provide values on o
$\mathscr{S}_v$, $\mathscr{S}_{\tilde{v}}$	Set of sources claim/disclaim v on o
$\mathscr{O}_s$	Set of objects covered by s
$\mathscr{V}_{s_o}$, $\tilde{\mathscr{V}}_{s_o}$	Set of positive/negative claims provided by s on o
$\mathscr{U}_o$	Set of all claimed values on o
$\mathscr{A}(s_1,s_2)$, $\tilde{\mathscr{A}}(s_1,s_2)$	Endorsement degree from s_1 to s_2 on positive/negative claims
$A_o(s_1,s_2)$, $\tilde{A}_o(s_1,s_2)$	Agreement between the positive/negative claims of s_1 and s_2 on o
$\omega(s_1 \to s_2)$, $\tilde{\omega}(s_1 \to s_2)$	The weight of edge from s_1 to s_2 in $\pm$agreement graph
α	source confidence factor
$V(s)$, $\tilde{V}(s)$	the vote for s's positive/negative claims being true/false
$\tau(s)$, $\tilde{\tau}(s)$	$\pm$SourceVote, evaluation of positive/negative precision of s

4.3 Problem Formulation

In Section 4.3.1 we formally define the multi-valued truth discovery problem. We validate the intuition that motivates us to model source reliability by quantifying the two-sided inter-source agreements in Section 4.3.2.

4.3.1 Problem Definition

A multi-valued truth discovery problem (i.e., MTD) generally involves five components (Table 4.1 summarizes the notations used in this chapter) during its life cycle:

Explicit inputs include i) a set of multi-valued *objects*, $\mathscr{O}$, each of which may have more than one true value to be discovered. The numbers of true value(s) can vary from object to object; ii) a set of *sources*, $\mathscr{S}$. Each $s \in \mathscr{S}$ provides potential true values on a subset of objects in $\mathscr{O}$; iii) *claimed values*, the values provided by any source of $\mathscr{S}$ on any objects of $\mathscr{O}$. Given a source s, we regard the set of values provided by s on an object o as *positive claim*s, denoted as $\mathscr{V}_{s_o}$.

Implicit inputs are derived from the explicit inputs and include: i) the complete set of values provided by all sources on any object o, denoted as $\mathscr{U}_o$; ii) by incorporating the *mutual exclusion assumption*, given an object o, a source s that makes positive claims $\mathscr{V}_{s_o}$ is believed to implicitly disclaim all the other values on o. We denote the set of values disclaimed by s as $\tilde{\mathscr{V}}_{s_o}$ (i.e., *negative claims* provided by s on o), which is calculated by $\mathscr{U}_o - \mathscr{V}_{s_o}$.

Intermediate variables are generated and updated during the iterative truth discovery procedure. They include: i) *source reliability*, which reflects the capability of each source providing true values; ii) *confidence score*, which reflects the confidence on a value's being true or false. In this chapter, we differentiate the false positives and false negatives made by sources by modeling two aspects of source reliability, namely positive precision (i.e., the probability of the positive claims of a source being true), and negative precision (i.e., the probability of the negative claims of a source being false). In the following section, we will derive $\pm$*SourceVote*, denoted by $\tau(s)$ and $\tilde{\tau}(s)$ for a source s, as the evaluations of source reliability, by simply capturing source authority features based on two-sided inter-source agreements. We will improve the source reliability evaluations by incorporating four important implications in next chapter.

Outputs are the *identified truth* for each object $o \in \mathscr{O}$, denoted as $\mathscr{V}_o^*$.

Ground truth is the factual truth for each object $o \in \mathscr{O}$, denoted as $\mathscr{V}_o^g$, which is used to measure the effectiveness of the truth discovery methods.

Table 4.2 An illustrative example: four sources provide author names of two books

	9780072830613	9780072231236
Ground Truth	Stephen;James	Michael
s_1	Stephen;James;Merrill	Michael;Lloyd
s_2	Stephen;James	Michael
s_3	Stephen;Kate	Michael;Susan
s_4	Stephen;Kate	Michael;Susan

Example 4.3.1. *Table 4.2 shows a sample Book-Author dataset. In this particular example, four sources (i.e., s_1, s_2, s_3, and s_4) claim values on two objects (i.e., the authors of two books id* : 9780072830613 *and id* : 9780072231236*, denoted as o_1 and o_2). Each cell in the table demonstrates the positive claims of a specific source on a specific object. For example, s_1 provides {Stephen;James;Merrill} as positive claims, i.e., $\mathscr{V}_{s_{o_1}}$, on object o_1. There are conflicts among these four sources as they provide different positive claims on the same objects. Table 4.2 also shows the ground truth of the two objects. Given the conflicting data, our goal is to identify the true authors for these two books. We can derive from the dataset that $\mathscr{U}_{o_1}$ = {Stephen;James;Merrill;Kate}, $|\mathscr{U}_{o_1}| = 4$, and $\mathscr{U}_{o_2}$ = {Michael;Lloyd;Susan}, $|\mathscr{U}_{o_2}| = 3$, based on which we can further extract implicit inputs regarding the two objects from the raw dataset as shown in Table 4.3 and Table 4.4. Comparing with the ground truth, s_2 provides all the true values, which deserves a higher positive precision and negative precision. Sources s_3 and s_4 provide the same values and there may be supportive relations or copying relations between them. We will discuss these two types of relations in next chapter. Source s_1 is audacious, which claims all the true values and additionally a false value for each object, while s_3 and s_4 are error-prone, both of which claim a false value for each object.*

We formally define the multi-valued truth discovery problem as follows:

Definition 4.3.1. ***Multi-Valued Truth Discovery Problem (MTD)*** *Given a set of multi-valued objects ($\mathscr{O}$) and a set of sources ($\mathscr{S}$) that provide conflicting values $\mathscr{V}$. The goal of*

Table 4.3 Truth discovery inputs regarding the first book

9780072830613				
	positive claims (pc)	# pc	negative claims (nc)	# nc
s_1	Stephen;James;Merrill	3	Kate	1
s_2	Stephen;James	2	Merrill;Kate	2
s_3	Stephen;Kate	2	James;Merrill	2
s_4	Stephen;Kate	2	James;Merrill	2

Table 4.4 Truth discovery inputs regarding the second book

9780072231236				
	positive claims (pc)	# pc	negative claims (nc)	# nc
s_1	Michael;Lloyd	2	Susan	1
s_2	Michael	1	Lloyd;Susan	2
s_3	Michael;Susan	2	Lloyd	1
s_4	Michael;Susan	2	Lloyd	1

MTD is to identify a set of true values ($\mathcal{V}_o^*$) from $\mathcal{V}$ for each object o, satisfying that $\mathcal{V}_o^*$ is as close to the ground truth $\mathcal{V}_o^g$ as possible. A truth discovery process often proceeds along with the estimation of the reliability of sources, i.e., positive precision and negative precision. The perfect truth discovery results satisfy $\mathcal{V}_o^* = \mathcal{V}_o^g$. □

4.3.2 Agreement as Hint

For multi-valued objects, sources may provide totally different, the same, or overlapping sets of values from one another. Given an object, we define the common values claimed by two sources on the object as *inter-source agreement*. Based on the mutual exclusion, we consider two-sided inter-source agreements. Specifically, +*agreement* (resp., –*agreement*) is the agreement between two sources on positive (resp., negative) claims, indicating that they agree with each other on their claimed (resp., disclaimed) common values being true (resp., false). Intuitively, the agreement among sources indicate endorsement. If the positive (resp., negative) claims of a source are agreed/endorsed by the majority of other sources, this source may have a high positive (resp., negative) precision and is called an *authoritative source*.

Suppose $\mathscr{V}_o^g$ is the ground truth of an object o, $\mathscr{U}_o$ is the set of all claimed values of o, we denote by $\mathscr{U}_o - \mathscr{V}_o^g$ the set of false values of o. For the simplicity of presentation, we use T, U, and F to represent $\mathscr{V}_o^g$, $\mathscr{U}_o$, and $\mathscr{U}_o - \mathscr{V}_o^g$ in this section. For any two sources s_1 and s_2, the +agreement between them on an object o is calculated as:

$$A_o(s_1, s_2) = \mathscr{V}_{s_{1o}} \cap \mathscr{V}_{s_{2o}} \tag{4.1}$$

Suppose s_1 and s_2, each selects a true value from T independently. We denote their selected values as t_1 and t_2, respectively. The probability of $t_1 = t_2$, denoted as $P_{A_o}(t_1, t_2)$, can be calculated as follows[1]:

$$P_{A_o}(t_1, t_2) = \frac{1}{|T|} \tag{4.2}$$

Similarly, let f_1 and f_2 be the two values independently selected by s_1 and s_2 from F, and $P_{A_o}(f_1, f_2)$ be the probability of s_1 and s_2 providing the same false value (i.e., $f_1 = f_2$), $P_{A_o}(f_1, f_2)$ can be calculated using:

$$P_{A_o}(f_1, f_2) = \frac{1}{|F|} \tag{4.3}$$

In reality, an object usually has a small truth set and random false values, i.e., $|T| \ll |U|$. Applying this to Equation (4.2) and Equation (4.3), we get:

$$P_{A_o}(f_1, f_2) \ll P_{A_o}(t_1, t_2) \tag{4.4}$$

Typically, the values claimed by sources would contain a fraction of true values from T and a faction of false values from F. By applying Equation (4.4), positive claims from T are more likely to agree with each other than those from F. This implies that the more

[1] Note that this probability is based on a prior knowledge that s_1 and s_2 each provides a true value, which is different from the probability of two sources s_1 and s_2 independently provide the same true value.

true values a source claims, the more likely the other sources agree with its claimed values. Inversely, if a source shows a high degree of agreement with the other sources regarding its claimed values, the values claimed by this source would have higher probability to be true, and this source would have a bigger positive precision. The positive precision of a source is endorsed by the +agreements between this source and the other sources.

Similarly, the –agreement between any two sources s_1 and s_2 on an object o is calculated as:

$$\tilde{A}_o(s_1, s_2) = \tilde{\mathscr{V}}_{s_{1o}} \cap \tilde{\mathscr{V}}_{s_{2o}} = U - (\mathscr{V}_{s_{1o}} \cup \mathscr{V}_{s_{2o}}) \tag{4.5}$$

Let $\tilde{A}_o(s_1, s_2) \cap T$ be the true values in the –agreement between s_1 and s_2, and $\tilde{A}_o(s_1, s_2) \cap F$ be the false values in the –agreement between s_1 and s_2, satisfying $|\mathscr{V}_{s_{1o}}| \ll |U|$, $|\mathscr{V}_{s_{2o}}| \ll |U|$, $|T| \ll |U|$. It can be proved that $|\tilde{A}_o(s_1, s_2) \cap T| \ll |\tilde{A}_o(s_1, s_2) \cap F|$. Therefore, it is more likely for sources to agree with each other on false values than true values with respect to their negative claims. This implies that the more false values a source disclaims, the more likely that other sources agree with its negative claims. Inversely, if a source shows a high degree of agreement with the other sources on its negative claims, the values disclaimed by this sources would have higher probabilities to be false, and this source would have a bigger negative precision. The negative precision of a source is endorsed by the –agreements between this source and the other sources.

4.4 The SourceVote Approach

In this section, we present a graph-based approach, called *SourceVote*, as a solution to multi-valued truth discovery, which is a two-step process: i) creating two graphs based on agreements among sources (Section 4.4.1), and ii) assessing two-sided source quality based

on the graphs and further use the assessment results to estimate value veracity or initialize truth discovery methods (Section 4.4.2).

4.4.1 Creating Agreement Graphs

By quantifying the two-sided inter-source agreements, we can construct two fully connected weighted graphs, namely ±*agreement graph*s. In each graph, vertices represent sources, each directed edge depicts that one source agrees with/endorses another source, and the weight on each edge depicts to what extent one source endorses the other source. In particular, +agreement (resp., –agreement) graph models the +agreement (resp., –agreement) among the sources. We define $\mathscr{A}(s_1,s_2)$ (resp., $\tilde{\mathscr{A}}(s_1,s_2)$) as the *endorsement degree* from s_1 to s_2 on positive (resp., negative) claims, representing the rate, at which s_2 is endorsed by s_1 on the values being true (resp., false).

+Agreement Graph. To construct the +agreement graph, we first formalize the endorsement from one source to another (e.g., $s_1 \rightarrow s_2$) on their common positive claims. Specifically, for each object that they both cover, we calculate the endorsement based on the +agreement between the two sources. Then, we sum up the endorsement on all their overlapping objects as follows,

$$\mathscr{A}(s_1,s_2) = \sum_{o \in \mathscr{O}_{s_1} \cap \mathscr{O}_{s_2}} \frac{|A_o(s_1,s_2)|}{|\mathscr{V}_{s_{2o}}|} \tag{4.6}$$

where $\mathscr{O}_s$ denotes the set of objects covered by s. Then, we calculate the weight on the edge from s_1 to s_2 as:

$$\omega(s_1 \rightarrow s_2) = \beta + (1-\beta) \cdot \frac{\mathscr{A}(s_1,s_2)}{|\mathscr{O}_{s_1} \cap \mathscr{O}_{s_2}|} \tag{4.7}$$

In Equation (4.7), we add a "*smoothing link*" by assigning a small weight to every pair of vertices, where β is the smoothing factor. This measure guarantees that the graph is always

connected and source reliability calculation can converge. For our experiments, we simply set $\beta = 0.1$ (empirical studies such as the work done by Gleich et al. [173] demonstrate more accurate estimation). Finally, we normalize the weights of out-going links from every vertex by dividing the edge weights by the sum of the out-going edge weights from the vertex. This normalization allows us to interpret the edge weights as the transition probabilities for the random walk computation.

–Agreement Graph. We construct the –agreement graph in a similar way by applying the following two equations:

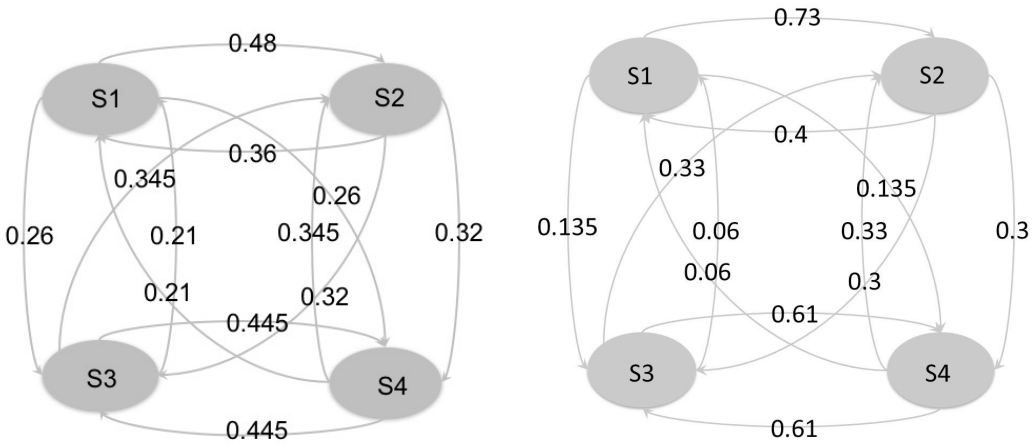
$$\tilde{\mathscr{A}}(s_1, s_2) = \sum_{o \in \mathscr{O}_{s_1} \cap \mathscr{O}_{s_2}} \frac{|\tilde{A}_o(s_1, s_2)|}{|\tilde{\mathscr{V}}_{s_{2o}}|} \tag{4.8}$$

$$\tilde{\omega}(s_1 \to s_2) = \beta + (1-\beta) \cdot \frac{\tilde{\mathscr{A}}(s_1, s_2)}{|\mathscr{O}_{s_1} \cap \mathscr{O}_{s_2}|} \tag{4.9}$$

Example 4.4.1. *Fig. 4.1 shows the sample ±agreement graphs for the dataset described in Example 4.3.1. Take the link from s_1 to s_2 in the sample +agreement graph as an example, by applying Equation* (4.1)*, we get $|A_{o_1}(s_1, s_2)| = 2$, and $|A_{o_2}(s_1, s_2)| = 1$. By substituting this result into Equation* (4.6)*, $\mathscr{A}(s_1, s_2) = \frac{2}{2} + \frac{1}{1} = 2$, and by further substituting this result into Equation* (4.7)*, $\omega(s_1 \to s_2) = 0.1 + (1 - 0.1) \times \frac{2}{2} = 1$. In the same way, we obtain $\omega(s_1 \to s_3) = 0.55$, and $\omega(s_1 \to s_4) = 0.55$. Finally, the normalized weights of edges $s_1 \to s_2$, $s_1 \to s_3$, and $s_1 \to s_4$ of +agreement graph are $\frac{1}{1+0.55+0.55} = 0.48$, $\frac{0.55}{1+0.55+0.55} = 0.26$, $\frac{0.55}{1+0.55+0.55} = 0.26$, respectively.*

4.4.2 Estimating Value Veracity and Source Reliability

To derive two-sided source reliability (positive and negative precision) from the two graphs, the measurements should capture two features: i) vertices with more input edges are assigned higher precision because those sources are endorsed by a large number of sources and should



(a) +Agreement Graph (b) –Agreement Graph

Fig. 4.1 Sample ±agreement graphs of four sources in Table 4.2

be more trustworthy[2]; ii) endorsement from a source with more input edges should be more trusted because both the authoritative sources and the sources endorsed by authoritative sources are more likely to be trustworthy. We adopt *Fixed Point Computation Model* (FPC) to capture the transitive propagation of source trustworthiness through agreement links based on the ±agreement graphs [89].

By applying FPC, we obtain the ranking scores of the two-sided precision of each source among all the sources. Specifically, we refer to each agreement graph as a Markov chain, where vertices serve as the states and the weights on edges as transition probabilities between the states. We calculate the asymptotic stationary visiting probabilities of the Markov random walk, where for each graph, all visiting probabilities sum up to 1.

Although, in this way, the visiting probabilities may not reflect the sources' real positive and negative precision, such feature renders the visiting probabilities of each source in the two graphs comparable. For this reason, we can count the visiting probability of each source in the +agreement (resp., –agreement) graph as the vote for its positive (resp., negative)

[2]Here we neglect the smoothing links, i.e., no link would be there between two sources in the graphs if no common value exists between the two sources.

Algorithm 2: SourceVote Algorithm

Input: a set of objects ($\mathscr{O}$), and the conflicting claimed values ($\mathscr{V}$) collected from a set of sources ($\mathscr{S}$)
Output: $\mathscr{V}_o^*$ identified truth for each $o \in \mathscr{O}$.

Initialization: smoothing factor β, source confidence factor α.
for *each* $s_i \in \mathscr{S}$ **do**
  for *each* $s_j \in \mathscr{S}, j \neq i$ **do**
    `/* Construct ±Agreement Graphs. */`
    calculate the weight of each edge in +agreement graph by Equation 4.6, 4.7;
    calculate the weight of each edge in –agreement graph by Equation 4.8, 4.9;
apply *FPC* to calculate $V(s)$ and $\tilde{V}(s)$ for each source;
for *each* $v \in \mathscr{V}$, $o \in \mathscr{O}$ **do**
  determine the veracity by Equation 4.10, and add the true values into $\mathscr{V}_o^*$;

claims being true (resp., false). We denote the corresponding vote count of each source as $V(s)$ (resp., $\tilde{V}(s)$) and further estimate the veracity of each claimed value as follows:

$$Veracity(v) = \begin{cases} True; & \text{if } \sum_{s \in \mathscr{S}_v} V(s) > \alpha \cdot \sum_{s \in \mathscr{S}_{\tilde{v}}} \tilde{V}(s) \\ False; & otherwise \end{cases} \tag{4.10}$$

where α is the source confidence factor, $\mathscr{S}_v$ (resp., $\mathscr{S}_{\tilde{v}}$) represents the set of sources that claim (resp., disclaim) v regarding o. Given a single-valued object, if a source claims a value, the source certainly disclaims all the other potential values. However, sources may not know the number of true values on the objects and thus do not necessarily reject negative claims on multi-valued objects. Therefore, we adopt a new mutual exclusion definition [116] and further add a source confidence factor, $\alpha \in (0,1)$, to differentiate the confidence of each source on its positive claims and negative claims. We will study the impact of α on the performance of SourceVote in Section 4.5.3. The detailed procedure of SourceVote is shown in Algorithm 2. The time complexity of the algorithm is $O(|\mathscr{S}|^2 + |\mathscr{V}|)$.

To further quantify the two-sided source reliability based on the calculated visiting probabilities, we apply a two-step normalization process: i) we set the positive precision

(resp., negative precision) of the source with the highest visit probability in the +agreement graph (resp., –agreement graph) as pp_{max} (resp., np_{max}), and calculate the *normalization rate* by dividing the precision by the corresponding visit probability; ii) normalizing the visiting probabilities of all sources as positive precision or negative precision (±SourceVote), denoted as $\tau(s)$ and $\tilde{\tau}(s)$, by multiplying the corresponding normalization rates.

Example 4.4.2. *After random walk computations for the agreement graphs in Example 4.4.1, we obtain the visit probabilities of the sources in the +agreement graph as* $\{s_1 : 0.21, s_2 : 0.28, s_3 : 0.26, s_4 : 0.26\}$*, and those in the –agreement graph as* $\{s_1 : 0.17, s_2 : 0.29, s_3 : 0.27, s_4 : 0.27\}$*. Suppose the real positive precision of* s_2 *is* 1*, and the real negative precision of* s_2 *is also* 1*, we finally obtain the +SourceVote of the sources as* $\{s_1 : 0.75, s_2 : 1, s_3 : 0.93, s_4 : 0.93\}$*, and –SourceVote of them as* $\{s_1 : 0.59, s_2 : 1, s_3 : 0.93, s_4 : 0.93\}$*. We can see that the above results capture the authority features of those four sources. For example, the positive claims and negative claims of* s_2 *are both endorsed by more sources than those of the other sources. Thus,* s_2 *is assigned with the highest ±SourceVote. According to the ground truth provided by Table 4.2,* s_3 *and* s_4 *agree with each other on false values, which depicts that there should be malicious agreement between them, leading to the result that the positive precision and negative precision of* s_3 *and* s_4 *are over-estimated (will be described in detail in next chapter).*

Note that most existing methods start with initializing source reliability as a default value, e.g., set source reliability as 0.8 [48]. LTM introduces two aspects of source quality, namely sensitivity and specificity, and thus requires initializing two parameters. Such initialization may fundamentally impact the convergence rate and precision of methods. According to Li et al. [47], "*knowing the precise trustworthiness of sources can fix nearly half of the mistakes in the best fusion results*". As constructing and computing our agreement graphs can be easily realized and require no initialization of source reliability, our approach can be applied to

existing methods for more precise source reliability initialization. Specifically, +*SourceVote* (resp. –*SourceVote*) is for positive (resp., negative) precision initialization.

4.5 Experimental Evaluation

4.5.1 Experimental Setup

We used two real-world datasets, which are comparable in size to those used in previous work, for our experiments. In particular, the *Parent-Children Dataset* [43], contains 11,099,730 records about individuals' birth dates, death dates, and the names of their parents/children and spouses. These data records have been edited by different users (i.e., data sources) on Wikipedia. For experimental purposes, we used the latest editing records as the ground truth. We specially extracted the records on the parent-children relations from this dataset for our experiments. After duplication removals, we obtained 55,259 sources claiming children for 2,579 people, each having 2.45 children on average. The *Book-Author Dataset* [44] is crawled from `www.abebooks.com` and contains 33,971 data records. The records are contributed by numerous book stores (i.e., sources), where each record represents the claimed values provided by a source on the author list of a book. We used the ground truth provided within the original dataset as the gold standard. We post-processed the data set to remove duplicate records, and obtained 12,623 distinct claims, where 649 sources (i.e., websites) provide author names on 664 books. Each book has 3.2 authors on average.

To compare our method with traditional truth discovery algorithms, we investigated the existing approaches that can be modified to tackle the multi-valued truth discovery problem. In contrast, most of the existing methods are inapplicable. For example, the approach in [43] requires normalizing the veracity of values, which is infeasible for the problem; the algorithms in [174], which cannot be applied to our problem because they all assume the number of false values as prior knowledge; the method in [118] focuses on handling

numerical data, while our approach is proposed specially for categorical data. As a result, we identified the following six methods as baselines:

- Voting. This method counts one vote for a claimed value when it is provided by a source. The claimed values will be regarded as true if the proportion of the sources (i.e., the vote count of a claimed value over the number of related sources) that claim the values exceeds a certain threshold.

- Sums (Hubs and Authorities) [91] and Average-Log [43]. They compute total trustworthiness of all sources that claim and disclaim a value separately, and recognize the value as true if the former is larger than the latter. In particular, the Sums method evaluates sources and values alternately from each other, while Avg-Log uses a non-linear function (a combination of logarithm and average functions) to assess sources, with the aim of avoiding overestimation of the trustworthiness of those sources that make more claims.

- TruthFinder [44], 2-Estimates [36], LTM [46], and MBM [116]. The four methods can be directly applied without modification, all of which recognize a value as true if its veracity score exceeds 0.5. In particular, TruthFinder alternately computes two measures, the confidence of fact (here, facts refer to values) and the trustworthiness of sources, from each other through an iterative procedure. It also considers the inter-value influence to improve accuracy. 2-Estimates incorporates the mutual exclusion between categorical values. LTM applies generative models to estimate truth. MBM is an integrated multi-truth Bayesian model. LTM and MBM are designed by relaxing the single-valued assumption.

To ensure the fair comparison, we ran a series of experiments to determine optimal parameter settings for each baseline method and used the same stop criterion for all the

Table 4.5 Comparison of Different Methods: The Best and Second Best Performance Values are in Bold.

Method	Book-Author Dataset				Parent-Children Dataset			
	Precision	Recall	F_1 score	Time(s)	Precision	Recall	F_1 score	Time(s)
Voting	**0.84**	0.63	0.72	**0.07**	**0.90**	0.74	0.81	**0.56**
Sums	**0.84**	0.64	0.73	0.85	**0.90**	0.88	0.89	1.13
Avg-Log	**0.83**	0.60	0.70	0.61	**0.90**	0.88	0.89	**0.75**
TruthFinder	**0.84**	0.60	0.70	0.74	**0.90**	0.88	0.89	1.24
2-Estimates	0.81	0.70	0.75	**0.38**	**0.91**	0.88	0.89	1.34
LTM	0.82	0.65	0.73	0.98	0.88	**0.90**	0.89	0.99
MBM	**0.83**	**0.74**	**0.78**	0.67	**0.91**	0.89	**0.90**	2.17
SourceVote	0.81	**0.77**	**0.79**	0.63	**0.90**	**0.92**	**0.91**	0.91

iterative methods. For our approach, we set $\alpha = 0.6$ (we will present our studies on the impact of α in Section 4.5.3).

We conducted experiments on a 64-bit Windows 7 PC with an octa-core 3.4GHz CPU and 16GB RAM. All algorithms of our approach were implemented in Python 3.4.0. We ran each method ten times and used four evaluation metrics (*precision*, *recall*, F_1 *score*, and *execution time*) to evaluate the average performance of each method, where F_1 score serves as an overall metric because neither precision nor recall could represent the method accuracy independently.

4.5.2 Comparison of Truth Discovery Methods

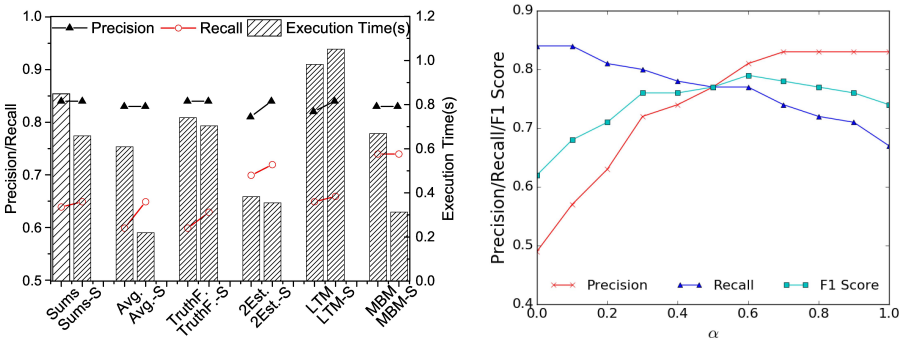
Table 4.5 shows the performance of different approaches on the two datasets in terms of precision, recall, F_1 score, and execution time. The results show that our approach consistently achieved the best recall and F_1 score among the methods. Compared with the two existing multi-valued truth discovery methods (LTM and MBM), SourceVote had the lowest execution time. This is because LTM conducted complicated Bayesian inference over a probabilistic graphical model, and MBM includes time-consuming copy detection. Moreover, Both LTM and MBM are iterative approaches; in contrast, our approach is based on a simpler graph-based model. All the algorithms achieved lower precision on the Book-Author dataset. The possible reasons include the small scale of this dataset, poor quality of sources, and insufficient evidence to support all true values (e.g., a true value might be

provided by no data source). The majority of methods showed higher precision than recall, reflecting the relatively high positive precision than negative precision of most real-world sources.

Specifically, Voting achieved relatively lower recall on both datasets, demonstrated by being the second worst on Book-Author dataset and the worst on Parent-Children dataset. This is because Voting ignores the differences of source quality but simply determines the truth of data by tuning the predefined threshold. To obtain the nearly perfect precision, the threshold of Voting is set as a high value bigger than 0.5. This result implies that instead of applying for solving multi-valued truth discovery problem, Voting can be most suitable to be used for generating the ground truth for semi-supervised truth-finding approaches. Besides SourceVote, 2-Estimates and MBM performed better than other methods. This can be attributed to their consideration of mutual exclusion. Though LTM also takes this implication into consideration, it makes strong assumptions on the prior distributions of latent variables. Once the dataset does not comply with the assumed distributions, it performs poorly. Although our approach achieved no significantly superior precision, the recall was improved drastically. For F_1 score, SourceVote consistently achieved the highest values for both datasets. The results reveal that our approach performs the best overall among all these baseline methods, which is consistent with our expectation because it makes no prior assumption and considers the endorsement relations among sources by combining with the graph-based method.

4.5.3 Empirical Studies of Different Concerns

To validate the feasibility of modeling source reliability by quantifying two-sided inter-source agreements and the feasibility of using SourceVote to initialize the existing truth discovery methods, we derive source positive precision and negative precision from the $\pm$agreement graphs by additionally conducting the two-step normalization process on the two real-world



(a) (b)

Fig. 4.2 Empirical Studies of Different Concerns of SourceVote: (a) Comparison between the original versions of representative existing truth discovery methods and the versions that apply SourceVote for precise source reliability initialization. The latter versions are marked by suffix "-s". (b) Performance of SourceVote under varying source confidence factor, i.e., α.

datasets. We used these results to initialize the parameters regarding source reliability of the aforementioned baseline methods (*Sums, Average-Log, TruthFinder, 2-Estimates, LTM*, and *MBM*). Note that we did not apply *SourceVote* to *Voting*, because *Voting* assumes all sources are equally reliable.

Figure 4.2a describes the performance comparison of the SourceVote initialized methods with their original versions in terms of precision, recall, and execution time on the Book-Author dataset. We omit the results on Parent-Children dataset as it led to similar conclusions. The results show that initializing source reliability by applying *SourceVote* almost led to better performance of all methods, indicated by higher precision and recall, and lower execution time. This reflects that the source reliability evaluated by *SourceVote* is more accurate than the widely applied default value of 0.8. With precise initialization, all methods achieved faster convergence speed. Specially, the precision and recall of *MBM* stayed stable, indicateing its insensitivity to the initial assumptions of source quality. However, the execution time of *MBM* was reduced dramatically under our new precise source quality inputs. The execution time of *LTM* increased because the number of iterations was fixed to 1001 to ensure algorithm

convergence and avoid performance fluctuations (as suggested in [48]). The increased time execution cost is the period spent on *SourceVote*.

We also investigated the performance of SourceVote by tuning the values of the source confidence factor α from 0 to 1 on both datasets. Figure 4.2b shows the impact of α on the performance of SourceVote in terms of precision, recall and F_1 score on the Book-Author dataset. When α equaled 0, indicating that the negative claims were not trusted at all, all the positive claims were labeled as true. In this case, the precision was undoubtedly very low (0.49), as there should be a large amount of low-quality sources providing false values; meanwhile, the recall was with no surprisingly very high (0.84), as all claimed values were regarded as true. The recall was less than 1, implying that some true values were missed and not claimed by any sources. As α grew, the precision of SourceVote dramatically increased (from 0.49 to 0.83) while the recall of SourceVote slightly decreased (from 0.84 to 0.67), implying that by putting more confidence on source negative precision, SourceVote was inclined to reject more false values than true values. The overall performance of SourceVote peaked at the point of $\alpha = 0.6$ with an F_1 score of 0.79, which is consistent with our intuition that source confidence on positive claims should be more respected. For $\alpha \in [0.3, 0.9]$, the lowest F_1 score of SourceVote is 0.76, which is still higher than the other baseline methods. The experimental results on Parent-Children dataset showed the similar results.

4.6 Summary

In this chapter, we have proposed a novel approach, *SourceVote*, to address the multi-valued truth discovery problem, which has been rarely studied in the literature. Our approach models the endorsement relations among sources by quantifying the agreements among sources on their positive and negative claims. Two graphs, namely $\pm$*Agreement Graph*, are constructed by incorporating such relations among sources. Based on these graphs, the two-sided vote counts of each source on their positive and negative claims and two aspects

of source reliability (positive precision and negative precision) are derived to differentiate the false positives and false negatives made by sources. Due to the compact feature of SourceVote, it can be leveraged to initialize and improve the existing truth discovery methods. Experimental results on two large real-world datasets show that our approach outperforms the state-of-the-art truth discovery methods.

The next chapter will focus on improving the proposed graph-based model for multi-valued truth discovery by exploring and incorporating more implications such as copying relations among sources, object popularity, and fine-grained source confidence on both positive claims and negative claims. We will also conduct more experimental studies to further validate the performance of our graph-based truth discovery model.

Chapter 5

A Full-Fledged Graph-Based Model for Multi-Valued Truth Discovery

In this chapter, we propose a full-fledged graph-based model, *SmartVote*, to conduct better multi-valued truth discovery. SmartVote models two types of source relations with additional quantification to precisely estimate source reliability for effective multi-valued truth discovery. Two graphs are constructed and further used to derive two aspects of source reliability (i.e., *positive precision* and *negative precision*) via random walk computations. Our model incorporates four important implications, including *two types of source relations*, *object popularity*, *loose mutual exclusion* and *long-tail phenomenon on source coverage*, to pursue accurate and complete results. Empirical studies on two large real-world datasets demonstrate the effectiveness of our approach. This chapter is based on our research reported in [175, 176, 172, 177].

5.1 Overview

So far, several *multi-valued method*s [46, 116, 125, 127, 128] have been proposed to tackle the multi-valued objects. However, based on the analysis in Chapter 4, we can additionally

identify the following disadvantages of those methods, which render the problem of truth discovery for multi-valued objects, a.k.a., the multi-valued truth discovery (MTD) problem, are still far from being solved. Firstly, most methods, except the method proposed in [116] make the source independent assumption. However, there are *supportive relation*s among sources, implying sources implicitly agree with/endorse one another sources by providing the same true values. Intuitively, a source endorsed by more sources is regarded as more authoritative and can be more trusted regarding its provided values. Sources may also maliciously copy false values from others, indicating *copying relation*s, which has only been studied by Wang et al. [116] in the cases where multi-valued objects exist. By relaxing the source independent assumption and identifying two types of source relations, namely *supportive relations* and *copying relations*, the general inter-source agreements quantified by SourceVote will be divided into two-sided agreements. Secondly, while *object difficulty* [36] (i.e., the difficulty of getting true values for varied objects) and *relations among objects* [43, 103] (i.e., objects may have relations that affect each other) have been studied by previous research efforts, to the best of our knowledge, no previous work differentiate the popularity of different objects. However, in reality, the impact of knowing the true values of various objects might be totally different. For example, between the email addresses and the children of a famous researcher, the email addresses are apparently more popular and have bigger impacts as other researchers or students doing research in the same areas often need to contact him/her. Taking object popularity into consideration could better model the real-world truth discovery and therefore lead to more accurate result. Thirdly, the long-tail phenomenon on source coverage of multi-sourced data has been empirically investigated on four real-world datasets by Li et al. in [123] for single-valued scenarios. However, no previous work has considered this issue in multi-valued cases. To conduct more precise multi-valued truth discovery, in a nutshell, our work makes the following main contributions:

- We propose a graph-based model, called *SmartVote*, as an overall solution to MTD. This model incorporates four important implications, including *two types of source relations*, *object popularity*, *loose mutual exclusion*, and *long-tail phenomenon on source coverage*, for better truth discovery.

- By relaxing the assumption that sources are independent of each other, we globally model two types of source relations, namely *supportive relation*s and *copying relation*s. Graphs capturing source features are constructed based on those relations. Specifically, source authority features and two-sided source precision are captured by $\pm$*supportive agreement graph*s, while source dependence scores are quantified by $\pm$*malicious agreement graph*s. Random walk computations are applied on both types of graphs to estimate source reliability and dependence scores.

- We propose to differentiate the popularity of different objects by leveraging object occurrences and source coverage, to minimize the number of people misguided by false values. The long-tail phenomenon on source coverage is not rare in the real-world. Our model solves MTD while additionally being aware of this phenomenon, to avoid the quality of sources with very few claims from being under- or over-estimated.

- We conduct extensive experiments to demonstrate the effectiveness of our approach via comparison with the state-of-the-art baseline methods on two real-world datasets. The impact of different implications on our model are also empirically studied and discussed.

The rest of the chapter is organized as follows. We discuss the observations that motivate our work in Section 5.2. Section 5.3 presents the model of SmartVote and the incorporated implications. We report our experiments and results in Section 5.4, and review the related work in Section 5.5. Finally, Section 5.6 provides some concluding remarks.

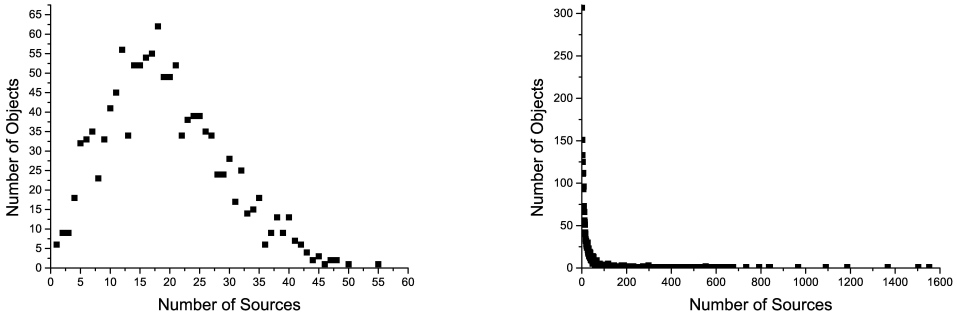
5.2 Preliminaries

The long-tail phenomenon on source coverage of multi-sourced data has been empirically investigated on four real-world datasets by Li et al. in [123] for single-valued scenarios, and the observations of sources' authority features and sources' copying relations have been presented in [91] and [131, 45, 129, 38] for single-valued scenarios. Unfortunately, no previous work has investigated the role of objects in the truth discovery community. To describe the significance of this implication, we present the statistical observations of objects on real-world datasets, and analyze the motivation of incorporating object popularity into our model in this section.

We have investigated the distributions of objects over sources in various real-world datasets. As an example, Figure 5.1a and Figure 5.1b show the results on the *Book-Author* [44] and *Biography*[1] [43] datasets, respectively. Each point (x,y) in the figure depicts y objects are covered by x sources in the corresponding dataset. We observe an apparent long-tail phenomenon from the distributions of Biography dataset (contains $2,579$ objects), which indicates that very few objects are referenced by large number of sources in the dataset, and many objects are covered by very few sources. For the Book-Author dataset with much fewer objects (contains $1,262$ objects), the long-tail phenomenon is less evident, but objects are claimed by significantly varying numbers of sources, indicating objects are of different occurrences. For example, there are 624 sources in total in the Book-Author dataset. The author list of book (id : 1558606041) is claimed by 55 sources, while the lists of book (id : 0201608359) and book (id : 020189551X) are only claimed by one source each.

Intuitively, sources tend to publish more popular information to gain more attention from the public, and the objects with more occurrences in the sources' claims indicate that they are more popular. Since the number of potential audiences of popular objects is usually bigger than that of less popular objects, if a source provides false values on a popular object, it

[1] In this chapter we focus on the parent-children relation in the dataset, because this is a multi-valued object.



(a) Book-Author Dataset (b) Biography Dataset

Fig. 5.1 The number of sources that provide values on objects: different objects are covered by varying numbers of sources.

will mislead more people than on a less popular object. With this consideration, we believe that there is different impact on the public for knowing the true values of different objects. We therefore propose to distinguish source reliability by differentiating the *popularity* of objects, to minimize the number of people misguided by false values. Sources providing false values for popular objects should be penalized more and assigned with lower reliability, to discourage them in misguiding the public. Meanwhile, sources providing false values for less popular objects should not be aggressively penalized. Moreover, from the data sufficiency's point of view, popular objects are generally claimed by more sources than the less popular objects, and more evidences can be used for estimating value veracity regarding those objects, leading to more reliable truth estimation. This supports the rationale of assigning more weights to popular objects in the calculation of source reliability, which indirectly leads to more accurate estimation.

Table 5.1 Notations Used in Chapter 5

Notation	Explanation
$Cov(s)$	The coverage of s
$\tau'(s)$, $\tilde{\tau}'(s)$	$\pm$SmartVote, improved evaluation of positive/negative precision of s
$\mathscr{C}_v$, $\mathscr{C}_{\tilde{v}}$	The confidence score of v being true/false
$\mathscr{P}_o$	The popularity degree of o
$\mathscr{D}(s,o)$, $\tilde{\mathscr{D}}(s,o)$	The dependence score of s providing positive/negative claims on o
$\mu(s,o)$, $\tilde{\mu}(s,o)$	The confidence score of s providing positive/negative claims on o
$\mathscr{L}(s_1,s_2)$	The long-tail phenomenon compensation for edge from s_1 to s_2
$\omega'(s_1 \to s_2)$, $\tilde{\omega}'(s_1 \to s_2)$	The weight of edge from s_1 to s_2 in $\pm$supportive agreement graph
$\omega_{c_o}(s_1 \to s_2)$, $\tilde{\omega}_{c_o}(s_1 \to s_2)$	The weight of edge from s_1 to s_2 in $\pm$malicious agreement graph of o

5.3 The SmartVote Approach

Based on SourceVote introduced in last chapter, we further propose a full-fledged graph-based model, called SmartVote, to pursue more accurate and complete results. We continue with the problem and notations formulated in last chapter. Table 5.1 summarizes the addtional notations used in this chapter. The improved evaluations of source positive and negative precision are named as $\pm$*SmartVote*, denoted by $\tau'(s)$ and $\tilde{\tau}'(s)$ for a source s. Accordingly, we estimate both the confidence scores of a value v being true (i.e., $\mathscr{C}_v$) and false ($\mathscr{C}_{\tilde{v}}$).

In reality, sources might not only support one another by providing the same true claims, but also may maliciously copy from others to provide the same false claims, which sometimes mislead the audiences. Therefore, we further identify two types of source relations to conduct more accurate source reliability estimation. Specifically, sharing the same true values means one source supports/endorses the other source, indicating a *supportive relation* between the two sources. We define the common values between these two sources as *supportive agreement*. Based on the analysis in last chapter, we can measure source reliability by quantifying inter-source supportive agreements. Even though one source can copy from the other in this case, we consider this type of copying relations as benignant. On the contrary, sharing the same false values is typically a rare event when the sources are fully independent. If two sources share a significant amount of false values, they are likely to copy from each other, indicating a *copying relation* between them. We define these common false values as *malicious agreement*. Neglecting the existence of deliberate copying of false values would impair the accuracy of source reliability estimation.

Besides source relations, several additional heuristics can also be considered to precisely estimate source reliability in reality. To this end, we propose a full-fledged graph-based model, called *SmartVote*, to solve the MTD problem, which incorporates four implications.

5.3.1 The Graph-Based Model

Figure 5.2 shows our SmartVote framework. Given the large-scale noisy multi-sourced Web data, it is difficult for a human to determine what the truth is. The goal of our model is to automatically predict the truth from the conflicting multi-sourced data. Our model incorporates four implications, including two types of source relations, object popularity, loose mutual exclusion, and long-tail phenomenon on source coverage by integrating four optimization components into one graph-based core component. We classify and briefly describe the components as follows:

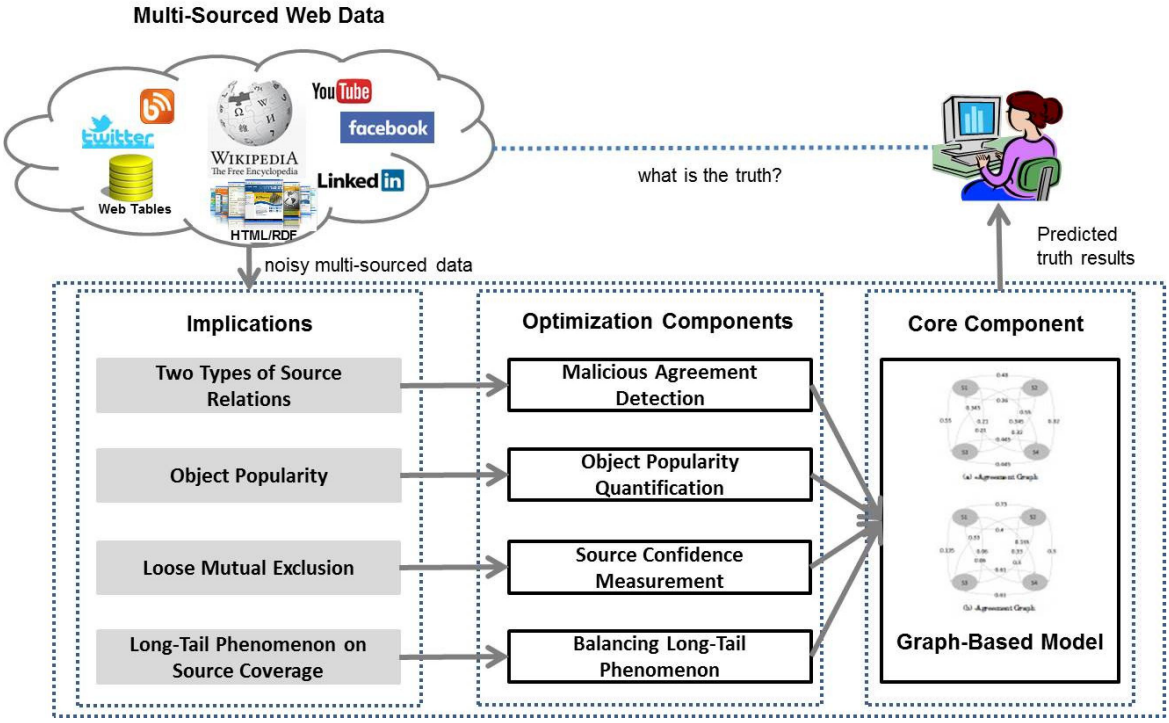


Fig. 5.2 The Framework of SmartVote

Core Component. It applies the following principle for truth discovery [49]: sources providing more true values are assigned with higher reliability; meanwhile, values provided by higher-quality sources are more likely to be true. Value confidence scores and source reliability are iteratively calculated from each other until convergence. By relaxing the source independent assumption and identifying two types of source relations, namely *supportive*

relations and *copying relations*, the general inter-source agreements quantified by SourceVote are divided into *supportive agreement*s and *malicious agreement*s. The SmartVote core component derives the improved evaluations of source positive and negative precision, i.e., $\pm$*SmartVote*, from two constructed $\pm$*supportive agreement graph*s. The constructions of $\pm$supportive agreement graphs incorporate the outputs of the four optimization components. Note that the supportive relations among sources are modeled by supportive agreement graphs constructed by the core component.

Optimization Components. These four optimization components compute the parameters regarding the four implications required by the core component. The *malicious agreement detection component* models the copying relations among sources and derives the dependence score of each source providing claims on each object (Section 5.3.2). The *object popularity quantification component* differentiates the popularity of objects based on the consideration that knowing the truths of different objects impacts differently on source reliability estimation (Section 5.3.3). For a multi-valued object, since sources may cautiously provide partial true values and omit the values they are not sure about, or audaciously provide all potential values, even if the veracity of the claimed values is uncertain, the mutual exclusion among values is not as strict as that of the single-valued object, i.e., the loose mutual exclusion. For this reason, SmartVote uses the *source confidence measurement component* to calculate the source confidence scores of providing positive (resp., negative) claims on each object, and reconcile sources' belief in their positive and negative claims (Section 5.3.4). Finally, the *balancing long-tail phenomenon on source coverage component* calculates the compensation of long-tail phenomenon on source coverage for each link in the $\pm$supportive agreement graphs to avoid small sources from being assigned with extreme reliability (Section 5.3.5).

In the core component, the constructions of $\pm$supportive agreement graphs are similar to those of $\pm$agreement graphs. In particular, we calculate the endorsement degree from s_1 to s_2 on positive claims by modifying Equation 4.6 in last chapter as follows:

$$\mathscr{A}(s_1,s_2)=\mathscr{L}(s_1,s_2)+\sum_{o\in\mathscr{O}_{s_1}\cap\mathscr{O}_{s_2}}\frac{|A_o(s_1,s_2)|}{|\mathscr{V}_{s_2o}|}\cdot(1-\prod_{v\in A_o(s_1,s_2)}\mathscr{C}_{\tilde{v}})\cdot\mathscr{P}_o\cdot(1-\mathscr{D}(s_1,o))\cdot\mu(s_1,o) \tag{5.1}$$

where $\mathscr{D}(s,o)$ is the dependence score of s providing positive claims on o (defined in Section 5.3.2), $\mathscr{P}_o$ is the popularity degree of o (defined in Section 5.3.3), $\mu(s,o)$ is the confidence score of s providing positive claims on o (defined in Section 5.3.4), and $\mathscr{L}(s_1,s_2)$ is the long-tail phenomenon compensation of edge from s_1 to s_2 (defined in Section 5.3.5).

We calculate the weight on each edge of +supportive agreement graph using:

$$\omega'(s_1\rightarrow s_2)=\beta+(1-\beta)\cdot\frac{\mathscr{A}(s_1,s_2)}{|\mathscr{O}_{s_2}|} \tag{5.2}$$

Similarly, we define the calculation of edge weights of –supportive agreement graph as:

$$\tilde{\mathscr{A}}(s_1,s_2)=\mathscr{L}(s_1,s_2)+\sum_{o\in\mathscr{O}_{s_1}\cap\mathscr{O}_{s_2}}\frac{|\tilde{A}_o(s_1,s_2)|}{|\tilde{\mathscr{V}}_{s_2o}|}\cdot(1-\prod_{v\in\tilde{A}_o(s_1,s_2)}\mathscr{C}_v)\cdot\mathscr{P}_o\cdot(1-\tilde{\mathscr{D}}(s_1,o))\cdot\tilde{\mu}(s_1,o) \tag{5.3}$$

$$\tilde{\omega}'(s_1\rightarrow s_2)=\beta+(1-\beta)\cdot\frac{\tilde{\mathscr{A}}(s_1,s_2)}{|\mathscr{O}_{s_2}|} \tag{5.4}$$

We apply FPC random walk to those two graphs, then obtain $\tau'(s)$ and $\tilde{\tau}'(s)$ as ±SmartVote for each source by conducting the same normalization process as with SourceVote. Besides the features captured by ±SourceVote, ±SmartVote additionally capture the following characteristics:

- The endorsement from a source on values with higher probability to be true (resp., false) in +supportive agreement graph, should be more (resp., less) respected. Meanwhile,

the endorsement from a source on values with higher probability to be false (resp., true) in –supportive agreement graph, should be more (resp., less) respected.

- The endorsement independently provided by a source should be more trustworthy, since the endorsement provided by copiers can be malicious and might have little wisdom in it. Also, the endorsement from a source on popular objects should be highlighted, since popular objects are more valued by the public, false values of which can lead to bad consequences.

- If a source shows bigger confidence on the claims (positive or negative) of an object, the endorsement from this source on the object should be highlighted. Also, sources covering few objects should not be assigned with extreme big or small values of ±SmartVote, since the evidences for estimating their reliability are limited.

To jointly determine value veracity from source reliability, we consider each source that belongs to $\mathscr{S}_o$ casts a smart vote to each potential value of o. In particular, if a source provides v as a positive claim, then it casts a vote proportional to $\tau'(s)$ for it; in contrast, if a source disclaims v, then it casts a vote proportional to $(1-\tau'(s))$ for it. Therefore, we compute the confidence scores of each value v being true and false by applying the following equations:

$$\mathscr{C}_v = \frac{\sum_{s\in\mathscr{S}_v}\tau'(s)+\sum_{s\in\mathscr{S}_{\tilde{v}}}(1-\tilde{\tau}'(s))}{|\mathscr{S}_o|} \tag{5.5}$$

$$\mathscr{C}_{\tilde{v}} = \frac{\sum_{s\in\mathscr{S}_v}(1-\tau'(s))+\sum_{s\in\mathscr{S}_{\tilde{v}}}\tilde{\tau}'(s)}{|\mathscr{S}_o|} \tag{5.6}$$

5.3.2 Malicious Agreement Detection

Copying relations among sources in real world are complex. For example, a copier may copy all values or partial values from a source; a source may transitively copy from another source

or collect information from several sources; and multiple sources may copy one source. To better model the malicious agreement among sources globally, we construct ± malicious agreement graphs for sources that provide values on an object o, i.e., $\mathscr{S}_o$, for each object $o \in \mathscr{O}$. Similar to the graphs constructed above, each edge of +malicious (resp., –malicious) agreement graph represents one source maliciously endorses the other on the positive (resp., negative) claims of an object with a quantified endorsement degree, denoted as $\omega_{c_o}(s_1 \to s_2)$ (resp., $\tilde{\omega}_{c_o}(s_1 \to s_2)$), calculated by:

$$\omega_{c_o}(s_1 \to s_2) = \beta + (1-\beta) \cdot \frac{|A_o(s_1,s_2)|}{|\mathscr{V}_{s_{2o}}|} \cdot (1 - \prod_{v \in A_o(s_1,s_2)} \mathscr{C}_v) \cdot \mu(s_1,o) \tag{5.7}$$

$$\tilde{\omega}_{c_o}(s_1 \to s_2) = \beta + (1-\beta) \cdot \frac{|\tilde{A}_o(s_1,s_2)|}{|\tilde{\mathscr{V}}_{s_{2o}}|} \cdot (1 - \prod_{v \in \tilde{A}_o(s_1,s_2)} \mathscr{C}_{\tilde{v}}) \cdot \tilde{\mu}(s_1,o) \tag{5.8}$$

Both FPC random walk computation and normalization are conducted on each graph to obtain the dependence score for each source that provides positive (resp., negative) claims on an object o, denoted as $\mathscr{D}(s,o)$ (resp., $\tilde{\mathscr{D}}(s,o)$). We set the dependent score of the source with the highest visit probability in the +malicious agreement graph (resp., –malicious agreement graph) as pc_{max} (resp., nc_{max}). The computed dependence scores capture the following characteristics, all of which are consistent with our intuition:

- Vertices with more input edges should have a higher value of dependence score, since those sources are maliciously endorsed by a larger number of sources. Such sources act as collectors that copy values from several sources.

- The malicious endorsement from a source on values with lower probability to be true (resp., false) in +malicious agreement graph, should be more (resp., less) respected. Meanwhile, the malicious endorsement from a source on values with lower probability to be false (resp., true) in –malicious agreement graph, should be more (resp., less) respected.

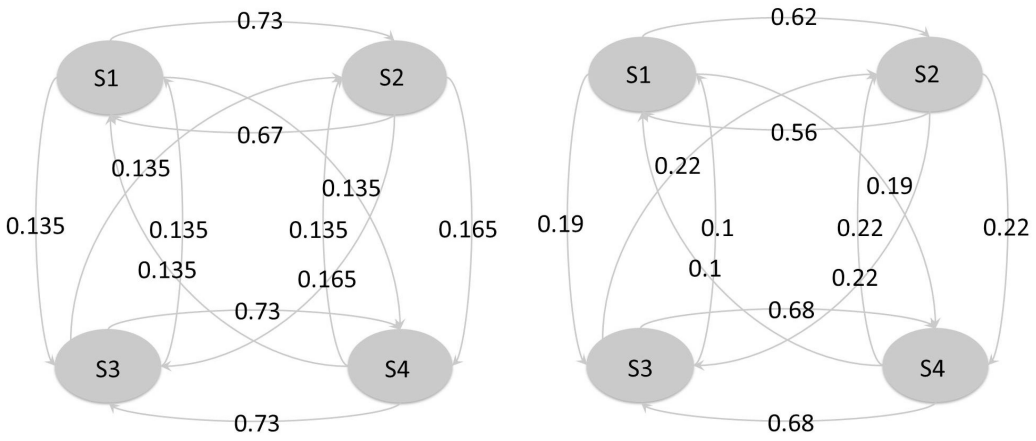
- If a source shows bigger confidence on the claims (positive or negative) of the object, the endorsement from this source should be highlighted.

Example 5.3.1. *Figure 5.3 shows sample ±malicious agreement graphs for book id* : 9780072830613 *(simply denoted as o) in the dataset described in Example 4.3.1*[2]*. Take the link from* s_1 *to* s_2 *in the sample +malicious agreement graph as an example, by applying Equation 4.1 in last chapter, we get* $A_o(s_1,s_2) = \{Stephen;James\}$*, then* $|A_o(s_1,s_2)| = 2$*. By applying majority voting, we get the votes for values in* $\mathscr{U}_o$ *as {Stephen: 4, James: 2, Kate: 2, Merrill: 1}. Therefore, we initialize the confidence scores for the values as* $\frac{4}{4} = 1$, $\frac{2}{4} = 0.5$, $\frac{2}{4} = 0.5$, $\frac{1}{4} = 0.25$. *By substituting this results in Equation* (5.7)*, we obtain* $\omega_{c_o}(s_1 \to s_2) = 0.1 + (1-0.1) \times \frac{2}{2} \times (1 - 1 \times 0.5) = 0.55$. *In the same way, we obtain* $\omega_{c_o}(s_1 \to s_3) = 0.1$*, and* $\omega_{c_o}(s_1 \to s_4) = 0.1$. *Finally, the normalized weights of edges* $s_1 \to s_2$*,* $s_1 \to s_3$*, and* $s_1 \to s_4$ *of +malicious agreement graph are* $\frac{0.55}{0.55+0.1+0.1} = 0.73$, $\frac{0.1}{0.55+0.1+0.1} = 0.135$, $\frac{0.1}{0.55+0.1+0.1} = 0.135$*, respectively. After applying random walk computations, we obtain* $\{\mathscr{D}(s_1,o) : 0.235, \mathscr{D}(s_2,o) : 0.245, \mathscr{D}(s_3,o) : 0.26, \mathscr{D}(s_4,o) : 0.26\}$*,* $\{\tilde{\mathscr{D}}(s_1,o) : 0.20, \tilde{\mathscr{D}}(s_2,o) : 0.25, \tilde{\mathscr{D}}(s_3,o) : 0.275, \tilde{\mathscr{D}}(s_4,o) : 0.275\}$*. The results capture the relation patterns in the sample dataset:* s_3 *and* s_4 *are more likely to be copiers than other sources on either positive claims or negative claims for the specific book id* : 9780072830613.

5.3.3 Object Popularity Quantification

Intuitively, popular objects tend to be covered by more sources, as sources tend to publish popular information to attract more audiences. Therefore, we quantify the popularity of each object, i.e., $\mathscr{P}_o$, in terms of occurrence. Specifically, we consider each source casts a vote for the popularity of each object it covers, and each object collect votes for its popularity from all the sources that claim values on it. We define the coverage of a source s, i.e., $Cov(s)$,

[2]We neglect the confidence scores of each source and omit the dependence score normalization step in this example.



(a) +Malicious Agreement Graph (b) –Malicious Agreement Graph

Fig. 5.3 Sample ±malicious agreement graphs of four sources on book *id* : 9780072830613 in Table 4.2 in last chapter.

as the percentage of its provided objects over $\mathscr{O}$. Formally, inspired by the idea from *term frequency–inverse document frequency* (i.e., tf-idf) in information retrieval, we measure the popularity of each object by applying the following equations, which comprehensively incorporate the occurrence of the object and the coverage of each source that covers the object:

$$\mathscr{P}_o^u = \sum_{s \in \mathscr{S}_o} \frac{1}{Cov(s)} \tag{5.9}$$

where $\mathscr{P}_o^u$ is the unnormalized popularity of object o. The $\mathscr{P}_o^u$ of all objects are then normalized as $\mathscr{P}_o$ to sum to 1.

The normalized popularity of each object captures the following two features, both of which are consistent with our intuition:

- The objects covered by more sources are more popular than those covered by fewer sources.

- The votes for the popularity of object from the sources with lower coverage should be more respected than those from the sources with higher coverage, as popular objects will be more conspicuous in small sources.

5.3.4 Source Confidence Measurement

For a single-valued object, if a source claims a value for it, then the source certainly disclaims all the other potential values of it. However, a straight application of this type of mutual exclusion for multi-valued object is unreasonable, because sources may not know the number of true values on the objects, and do not necessarily reject negative claims. To differentiate and quantify sources' confidence on their positive and negative claims, we incorporate loose mutual exclusion [116] into our model for source reliability calculation. The measurement approach of source confidence score is similar to the *Kappa coefficient* [178], the main idea is to exclude the effect of random guess in determining the extent. In particular, the confidence score of s providing positive claims on o is calculated as:

$$\mu(s,o) = \frac{1}{|\mathscr{V}_{s_o}|} \cdot (1 - \frac{1}{|\mathscr{U}_o|}) \tag{5.10}$$

Meanwhile, the confidence score of s providing negative claims on o is calculated as:

$$\tilde{\mu}(s,o) = \frac{1}{|\tilde{\mathscr{V}}_{s_o}|} \cdot \frac{1}{|\mathscr{U}_o|} \tag{5.11}$$

The computed source confidence scores capture the following two features, which are consistent with our intuition:

- A cautious source, which only provides values that it is sure to be true and omits uncertain values, may claim partial true values of an object. Thus its confidence score on positive claims is relatively higher than that of the other sources, while its confidence score on negative claims is relatively lower than that of the other sources.

- For an audacious source, which tends to provide all potential values of an object, it may cover as many as possible values of an object, including false values. Thus its confidence score on positive claims is relatively lower than that of the other sources, while its confidence score on negative claims is relatively higher than that of the other sources.

Algorithm 3: The Algorithm of SmartVote.

Input: objects of interest $\mathscr{O}$, a set of sources $\mathscr{S}$, and $\mathscr{V}_{s_o}$ the set of positive claims provided by each $s \in \mathscr{S}$ on each $o \in \mathscr{O}$.
Output: $\mathscr{V}_o^*$ identified truth for each $o \in \mathscr{O}$.

```
// Initialization phase
```
Initialize δ, β, $\beta_{\mathscr{L}}$, pp_{max}, np_{max}, pc_{max}, nc_{max}
Initialize $\mathscr{C}_v$, $\mathscr{C}_{\tilde{v}}$ for each $v \in \mathscr{V}$, $o \in \mathscr{O}$
```
// Object popularity quantification
```
foreach $o \in \mathscr{O}$ **do**
    compute $\mathscr{P}_o$ by Equation (5.9)
```
// Source confidence measurement
```
foreach $s \in \mathscr{S}$ **do**
    foreach $o \in \mathscr{O}$ **do**
        compute $\mu(s,o)$, $\tilde{\mu}(s,o)$ by Equation (5.10), (5.11)
```
// Balancing long-tail phenomenon on source coverage
```
compute $\mathscr{L}(s_1,s_2)$ by Equation (5.12)
```
// Iteration phase
```
repeat
    ```
    // Malicious agreement detection
    ```
    foreach $o \in \mathscr{O}$ **do**
        construct $\pm$malicious agreement graphs by quantifying the weights of each edge by Equation (5.7), (5.8)
        derive $\mathscr{D}(s,o)$, $\tilde{\mathscr{D}}(s,o)$ by applying random walk and normalization steps
    ```
    // ±SmartVote computation
    ```
    construct $\pm$supportive agreement graphs by quantifying the weights of each edge by Equation (5.1), (5.2),(5.3), (5.4)
    derive $\tau'(s)$, $\tilde{\tau}'(s)$ by applying random walk and normalization steps
    ```
    // Value confidence score computation
    ```
    foreach $v \in \mathscr{V}$, $o \in \mathscr{O}$ **do**
        compute $\mathscr{C}_v$, $\mathscr{C}_{\tilde{v}}$ by Equation (5.5), (5.6)
until *convergence*;
return $\{(o,v)|v \in \mathscr{V} \wedge \mathscr{C}_v > \mathscr{C}_{\tilde{v}} \wedge v \in \mathscr{U}_o, o \in \mathscr{O}\}$

5.3.5 Balancing Long-Tail Phenomenon on Source Coverage

In reality, various datasets show the long-tail phenomenon on source coverage, which refers to the fact that very few sources provide extensive coverage for the objects of interest and most of the source only provide values for very few objects. Since identifying reliable sources is the key to determining value veracity and source reliability is typically estimated by the empirical probability of making correct claims, the accuracy of truth discovery and source reliability estimation depend on the coverage of the evaluated source. When sources cover numerous objects, we can conduct more accurate estimation of source reliability based on these sufficient evidences, leading to better truth discovery. However, due to the existence of long-tail phenomenon, the majority of sources are "*small*" sources with very few claims. Source reliability estimation based on these limited evidences could be totally random. For example, consider the extreme case when most sources only cover one object. If the claimed values of one of these sources is correct and complete, the positive precision and negative precision of this source would both be one. On the other hand, if the claim is totally wrong, the positive precision would be zero. To smooth the estimation for small sources, given an object, we consider three cases for the agreement between two sources: i) both sources sharing several common values; ii) both sources providing totally different values; iii) one source covering this object while the other source ignoring this object. To deal with the long-tail phenomenon, we assert that the agreement in the third case should not be zero. If a source does not cover an object, it does not represent that this source vote against the values claimed by the other sources. In this section, our goal is to distinguish the third case from the second case. Formally, we use $\mathscr{L}(s_1, s_2)$ to represent a compensation for a link $s_1 \rightarrow s_2$, to rebalance the long-tail phenomenon on source coverage. In particular, for each object that covered by s_2 but not covered by s_1, we aproximately estimate the endorsement degree from s_1 to s_2 on this object according to Equation 5.1 and 5.3. Each factor on the right side of $\sum$ in Equation 5.12 corresponds to the factor in the same position of Equation 5.1 and 5.3.

$$\mathscr{L}(s_1,s_2)=\begin{cases}\beta_{\mathscr{L}}\cdot\sum_{o\in\mathscr{O}_{s_2}-(\mathscr{O}_{s_1}\cap\mathscr{O}_{s_2})}\frac{1}{2}\cdot\frac{1}{2}\cdot\mathscr{P}_o\cdot\frac{1}{2}\cdot(\frac{1}{|\mathscr{U}_o|}(1-\frac{1}{|\mathscr{U}_o|})); & \text{for } \mathscr{A}(s_1,s_2)\\ \beta_{\mathscr{L}}\cdot\sum_{o\in\mathscr{O}_{s_2}-(\mathscr{O}_{s_1}\cap\mathscr{O}_{s_2})}\frac{1}{2}\cdot\frac{1}{2}\cdot\mathscr{P}_o\cdot\frac{1}{2}\cdot\frac{1}{|\mathscr{U}_o|}^2; & \text{for } \tilde{\mathscr{A}}(s_1,s_2)\end{cases} \tag{5.12}$$

where $\beta_{\mathscr{L}}$ is an uncertainty factor of the compensation.

5.3.6 The Algorithm

Algorithm 3 shows the whole procedure of SmartVote. In the initialization phase, the parameters, such as iteration convergence threshold δ, smoothing factor β, uncertainty factor $\beta_{\mathscr{L}}$, positive precision pp_{max}, negative precision np_{max}, the two-sided dependence scores (pc_{max} and nc_{max}) of sources with the highest visit probabilities in $\pm$supportive agreement graphs and $\pm$malicious agreement graphs, are initialized with their a priori values (line 1). The confidence scores of each value v being true (denoted as $\mathscr{C}_v$) or being false (denoted as $\mathscr{C}_{\tilde{v}}$) are both initialized by adopting the majority voting in our experiments (in fact, other truth discovery methods can also be applied for this initialization). To start, we count the votes of each individual value of each object $o\in\mathscr{O}$, then normalize those vote counts by dividing them by $|\mathscr{S}_o|$ to represent $\mathscr{C}_v$ for each value. $\mathscr{C}_{\tilde{v}}$ is initialized as $1-\mathscr{C}_v$ (line 2). The *object popularity quantification* (lines 3-4), *source confidence measurement* (lines 5-7), and long-tail phenomenon on source coverage balancing (line 8) are calculated directly from the multi-sourced data outside the iteration. For each cycle of iteration, the algorithm recalculates the two-sided *source dependence scores* (lines 10-12), it continues to calculate $\pm$*SmartVote* (lines 13-14) of sources based on the two-sided *value confidence score*s, and compute value confidence scores (lines 15-16) based on $\pm$SmartVote of sources. The algorithm uses the convergence test where the difference of *cosine similarity* of $\pm$SmartVote between two successive iterations should be less than or equal to a given threshold, δ (line 17).

The time complexity of the algorithm is $O(|\mathscr{O}||\mathscr{S}|^2 + |\mathscr{S}|^2 + |\mathscr{V}|)$. There are many mature distributed computing tools that can be used for random walk computation to reduce the time complexity. For example, Apache Hama[3] is a framework for big data analytics, which uses the *Bulk Synchronous Parallel* (BSP) computing model. It includes the *Graph* package for vertex-centric graph computations. Note that we can easily extend the *Vertex* class to create a class for realizing parallel random walk computation.

5.4 Experiments

5.4.1 Experimental Setup

Real-World Datasets

We used two real-world datasets in our experiments. Each object in both datasets may contain multiple true values: i) *Book-Author dataset* [44] contains 33,971 book-author records crawled from *www.abebooks.com*. These records were collected from numerous book websites (i.e., sources). Each record represents a store's positive claims on the author name(s) of a book (i.e., objects). We refined the dataset by removing the invalid and duplicated records, and excluding the records with only minor conflicts to make the problem more challenging — otherwise, even a straightforward method could yield competitive results. We finally obtained 13,659 distinctive claims, 624 websites providing values about author name(s) of 655 books, each book has on average 3.1 authors. The ground truth provided by the original dataset was used as the gold standard. ii) *Parent-Children dataset* was extracted by focusing on parent-children relation from the *Biography dataset* [43], which contains 11,099,730 records edited by different users about people's birth and death dates, their parents', children', and spouses' names on Wikipedia. We obtained 227,583 claims about 2,579 people's children information (i.e., objects) edited by 54,764 users (i.e., sources). We

[3] https://hama.apache.org/

also further removed the duplicated and minor conflicting records for this dataset for more effective comparison. In the resulting dataset, each person has on average 2.48 children. We used the latest editing records as the gold standard.

Baseline Methods

We compared SmartVote with three types of truth discovery methods:

Existing MTD methods. Based on our thorough analysis of the existing MTD methods (in Section 5.5), we chose the following three state-of-the-art MTD methods as our baselines:

- *LTM (Latent Truth Model)* [46]: it applies a probabilistic graphical model to infer source reliability and value veracity.
- *MBM (Multi-truth Bayesian Model)* [116]: it incorporates source confidence and a finer-grained copy detection technique into a Bayesian model.
- *MTD-hrd* [125]: a model designed for *Multi-Truth Discovery*, which incorporates two implications, namely the calibration of imbalanced positive/negative claim distributions and the consideration of the implication of values' co-occurrence in the same claims, to improve the probabilistic approach.

STD methods. Wang et al. [127] validated the statement that by determining value veracity for multi-valued objects by jointly regarding a value set claimed by a source as a single value, traditional STD methods all result in low accuracy for MTD scenarios. To adapt existing STD methods to MTD scenarios, we pre-processed the input datasets by conducting claim value separation. For example, for the record "s_2, 9780072830613, Stephen;James" in Table 4.2 in Chapter 4, we reformatted it as "s_2, 9780072830613, Stephen" and "s_2, 9780072830613, James". We modified the STD methods to treat each value in a source's claimed value set on a given object individually, and determined the veracity of each individual value separately to accept multiple true values. We chose several typical and competitive methods

for comparison, and excluded the methods that are inapplicable to the MTD scenario. For example, the methods proposed in [38] use the *number of false values* as prior knowledge, which is not possible to be obtained in advance in the MTD scenario, because we do not know the number of true values for each object. IATD (Influence-Aware Truth Discovery) [120] makes a number of variable distribution assumptions, which are also infeasible to adapt to the MTD scenarios. The method in [43] requires the normalization of the veracity scores of values, which is infeasible for the MTD problem. The methods in [118, 40] focus on handling heterogeneous data, and the method in [123] is designed for continuous data, while our approach is designed for categorical data.

- *Voting*: this method regards a value as true if the proportion of the sources that claim the value exceeds a certain threshold.

- *Sums [91], Average-Log [43]*: these two methods are modified to incorporate mutual exclusion. They compute the total reliability of all sources that claim and disclaim a value separately. If the former is bigger than the later, then the value will be regarded as true.

- *TruthFinder* [44]: this method iteratively estimates *trustworthiness of source* and *confidence of fact* from each other by additionally considering the *influences between facts*.

- *2-Estimates* [36]: this method adopts mutual exclusion, recognizes a value as true if its truth probability exceeds 0.5.

Improved STD methods. We improved the above STD methods by incorporating truth number prediction. In particular, for each method, we treated the values in each claimed value set of each source individually, and ran the original method to output source reliability and value confidence scores. Then, we computed $|\mathscr{V}_{s_o}|$ for each source on each object, based

on which we predicted the number of true values for each object by applying the following equation:

$$P_o(n) = \sqrt[|\mathscr{S}_o|]{\prod_{|\mathscr{V}_{s_o}|=n, s\in\mathscr{S}_o} A(s) \cdot \prod_{|\mathscr{V}_{s_o}|\neq n, s\in\mathscr{S}_o} (1-A(s))} \tag{5.13}$$

where $P_o(n)$ is the unnormalized probability[4] of the number of values of an object o to be n, and $A(s)$ is the reliability of s calculated by each method.

For each object, we chose the number with the highest probability (denoted as N) as the number of true values and output the top-N values instead of choosing the value set with the biggest confidence score as the outputs. Finally, we obtained five new methods, namely *Voting**[5], *Sums**, *Average-Log**, *TruthFinder**, and *2-Estimates**.

Parameter Configuration

To ensure the fair comparison, we ran a series of experiments to determine the optimal parameter settings for each baseline method. We used the same stop criterion for all the iterative methods for convergence. For our approach, we simply used the default parameter settings for both datasets. In particular, we set $\beta_{\mathscr{L}}$ as 0.1, we study the impact of $\beta_{\mathscr{L}}$ on the performance of our approach by tuning this parameter in Section 5.4.3. Intuitively, sources tend to provide values that they are sure to be true and omit uncertain values, while copiers are likely to copy those explicitly claimed values from other sources. Therefore, we set pp_{max} as 1, np_{max} as 0.9, pc_{max} as 1, and nc_{max} as 0.8. We also studied the impact of those parameters on the performance of our approach, but omitted these experimental results due to space limitation.

[4] Such values are then normalized to represent probabilities.

[5] For Voting*, we predict the number of true values as the number with the highest vote counts.

Evaluation Metrics

We implemented all the above methods in Python 3.4.0 and ran experiments on a 64-bit Windows 10 Pro. PC with an Intel Core i7-5600 processor and 16GB RAM. We ran each method multiple times (denoted as K, for our experiments, we set K as 10) to evaluate their average performance. In particular, we used two groups of evaluation metrics.

Traditional performance measurements. *Precision* and *recall* are two traditional and commonly used performance measurements for evaluating the accuracy of truth discovery methods. We additionally used F_1 score as an overall metric as neither precision nor recall could represent the accuracy independently. Execution time was also measured for efficiency comparison.

Object popularity weighted performance measurements. As we introduced the new concept of object popularity, the traditional precision and recall on average cannot capture this implication. To measure the performance of truth discovery methods more precisely, we used the following three object popularity weighted performance metrics: i) *Weighted precision*, calculated as $\frac{1}{K}\sum_{k=1}^{K}\sum_{n=1}^{|\mathscr{O}|}\frac{|\mathscr{V}_o^{*(k)}\cap\mathscr{V}_o^{g}|}{|\mathscr{V}_o^{*(k)}|}\cdot\mathscr{P}_o$; ii) *Weighted recall*, calculated as $\frac{1}{K}\sum_{k=1}^{K}\sum_{n=1}^{|\mathscr{O}|}\frac{|\mathscr{V}_o^{*(k)}\cap\mathscr{V}_o^{g}|}{|\mathscr{V}_o^{g}|}\cdot\mathscr{P}_o$; iii) *Weighted F_1 score*, which is the harmonious mean of weighted precision and weighted recall.

5.4.2 Comparative Studies

Table 5.2 shows the performance of different methods on the two real-world datasets in terms of accuracy and efficiency. For all the accuracy evaluation metrics except precision, *SmartVote* consistently achieved the highest value. Even in terms of precision, SmartVote still achieved the second best performance on Parent-Children dataset and the third best performance on Book-Author dataset. Among the four methods specially designed for the MTD problem, our approach is the most efficient as demonstrated by its lowest execution time.

Table 5.2 Comparison of Different Methods: The Best and Second Best Performance Values are In Bold.

Method	Book-Author Dataset							Parent-Children Dataset						
	P	R	F1	WP	WR	WF1	T(s)	P	R	F1	WP	WR	WF1	T(s)
Voting	**0.84**	0.63	0.72	**0.83**	0.64	0.72	**0.07**	0.88	0.85	0.87	0.69	0.68	0.69	**0.56**
Sums	**0.84**	0.64	0.73	**0.83**	0.64	0.72	0.85	**0.90**	0.89	0.90	**0.88**	0.86	0.87	1.13
Avg-Log	**0.83**	0.60	0.70	**0.83**	0.64	0.72	0.61	**0.90**	0.89	0.89	**0.88**	0.86	0.87	**0.75**
TruthFinder	**0.84**	0.60	0.70	**0.83**	0.60	0.70	0.74	**0.90**	0.89	0.90	**0.88**	0.85	0.86	1.24
2-Estimates	0.81	0.70	0.75	0.80	0.68	0.74	0.38	**0.91**	0.89	0.90	**0.88**	0.86	0.87	1.34
Voting*	0.77	0.42	0.54	0.80	0.39	0.53	0.13	0.87	0.85	0.86	0.71	0.68	0.69	0.89
Sums*	**0.83**	0.24	0.38	**0.85**	0.21	0.34	0.99	0.86	0.88	0.87	0.67	0.84	0.75	1.45
Avg-Log*	0.74	0.49	0.59	0.80	0.53	0.64	**0.08**	0.89	0.87	0.88	0.77	0.82	0.79	0.92
TruthFinder*	0.70	0.71	0.70	0.75	**0.72**	0.73	0.99	0.85	0.91	0.88	0.69	0.88	0.77	1.16
2-Estimates*	**0.83**	0.24	0.38	0.81	0.21	0.34	0.79	0.86	0.89	0.87	0.66	0.83	0.74	1.47
LTM	0.82	0.65	0.73	0.82	0.62	0.71	0.98	0.87	0.90	0.88	0.86	0.89	0.87	0.99
MBM	**0.83**	**0.74**	**0.78**	0.82	0.71	**0.76**	0.67	**0.90**	**0.92**	**0.91**	0.87	**0.90**	**0.88**	2.17
MTD-hrd	**0.83**	0.58	0.68	**0.82**	0.59	0.69	0.72	**0.90**	0.90	**0.90**	0.87	0.89	**0.88**	1.37
SmartVote	0.81	**0.79**	**0.80**	**0.83**	**0.81**	**0.82**	0.45	**0.90**	**0.94**	**0.92**	**0.92**	**0.95**	**0.93**	0.92

This is because *LTM* and *MTD-hrd* include complicated Bayesian inference over the complex probabilistic graphical model, and *MBM* conducts time-consuming copy detection, while our approach is based on a relatively simple graph model. All the algorithms performed better on Parent-Children dataset than on Book-Author dataset. The possible reasons include the small scale, the poor quality of sources, and missing values (i.e., true values that are missed by all the data sources) of Book-Author dataset, leading to insufficient evidences to support all correct values. The majority of methods showed higher precision than recall, reflecting relatively high positive precision than negative precision of most real-world sources.

Specifically, since *Voting* conducts truth discovery without iteration and the consideration of the quality of sources, it has relatively low accuracy, but on the other hand, it consumes the shortest execution time. The improved STD methods performed even worse than their original versions. This depicts that in reality the majority of the sources tend to be cautious and only provide values they are sure to be true, thus the predicted numbers of true values were generally smaller than the real ones, leading to lower precision and recall of the improved STD methods. Besides our approach, *2-Estimates* and *MBM* also performed better than the other methods in terms of both the traditional and weighted measurements. This is attributed to their consideration of mutual exclusion. Though *LTM* and *MBM-hrd* also take

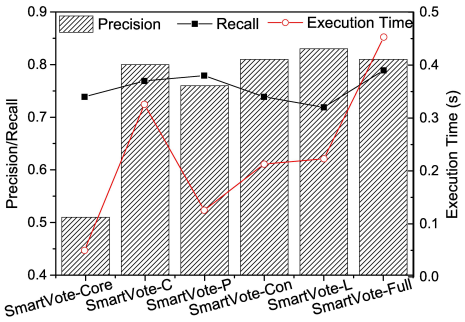
this implication into consideration, they make strong assumptions on the prior distributions of latent variables. For this reason, once the dataset does not comply with the assumed distributions, it performs poorly. Without incorporating object popularity, *2-Estimates* and *MBM* showed lower quality in terms of weighted metrics than traditional metrics. Compared with *MBM*, which showed second best performance, *SmartVote* not only includes object popularity, but also globally models two types of source relations, pays attention to the ubiquitous long-tail phenomenon on source coverage. Overall, *SmartVote* showed the best accuracy performance.

5.4.3 Impact of Different Concerns

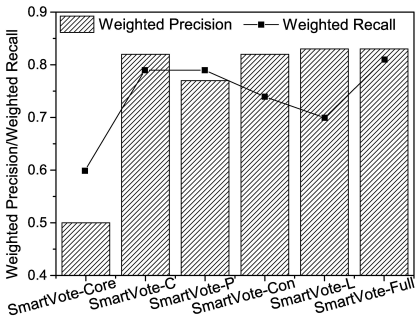
The compound effect of different technical components contributes to the performance of SmartVote. To evaluate the impact of different concerns, we implemented five variants of SmartVote:

- *SmartVote-Core*: A variant of SmartVote without incorporating the four implications.
- *SmartVote-C*: A version of SmartVote-Core that adopts the malicious agreement detection.
- *SmartVote-P*: A version of SmartVote-Core that adopts the object popularity quantification.
- *SmartVote-Con*: A version of SmartVote-Core that incorporates the loose mutual exclusion.
- *SmartVote-L*: A version of SmartVote-Core that considers the long-tail phenomenon on source coverage.

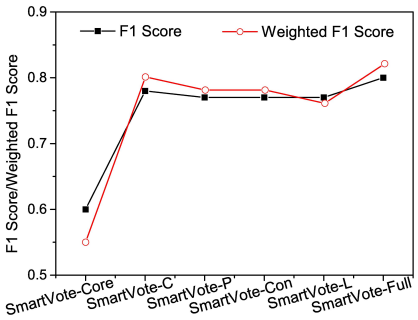
Figure 5.4 reports the performance comparison of different variants of SmartVote on Book-Author dataset. The experimental results on Parent-Children dataset show the similar



(a) Precision, Recall & T(s)



(b) Weighted Precision & Recall



(c) F1 Score & Weighted F1 Score

Fig. 5.4 Performance Comparison of Different Variants of SmartVote

insights. By incorporating each individual component into SmartVote-Core, our approach showed increasingly better performance in terms of accuracy while increased execution time slightly. The full version of SmartVote consisting of all the components led to the best result. We studied the impact of each technical component on our approach and report the findings one by one in the following sections.

Malicious Agreement

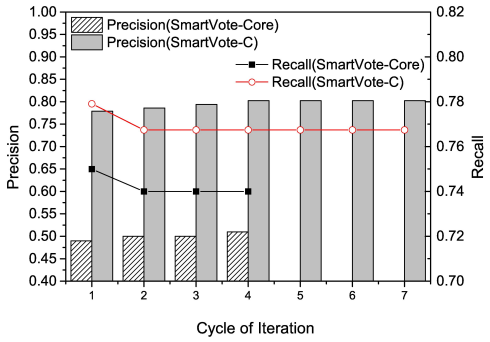
To validate the fact that two types of source relations, namely source supportive relations and copying relations (or two types of agreements, source supportive agreements and malicious agreements), widely exist in real-world datasets, we conducted data analysis on the Book-Author dataset. The experimental results on the Parent-Children dataset show the same

features. We found that among all the objects covered by the ground truth, only 11.76% sources, on which no source shows malicious agreement with others, make unique false claims. Meanwhile, there is only one object, on which no source claims the true values, indicating that no source shows supportive relation with others.

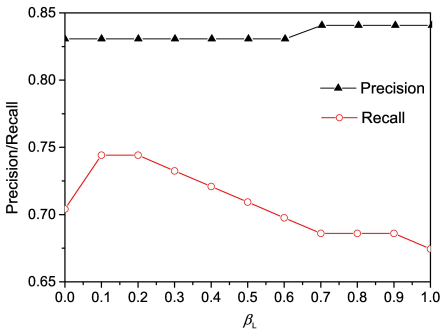
By incorporating malicious agreement detection component, both precision, recall and F_1 Score of SmartVote-C are higher than SmartVote-Core, as shown in Figure 5.4. Interestingly, when we leveraged weighted metrics to evaluate SmartVote-C, the algorithm even showed better results than by using traditional metrics. The results report the wide existence of copying relations in the real-world datasets. Neglecting these relations would lead to the result of overestimating the reliability of copiers and impair the performance of truth discovery methods. Among other components, malicious agreement detection is the most time-consuming, as we need to compute the dependence score of each source on each object iteratively from the confidence scores of the claimed values of each object, and calculate the reliability of each source iteratively from the independence score of each source and the confidence score of each value. However, when compared with the performance improvement introduced by incorporating this component, this additional amount of time can be justified. To study the effect of this component in depth, we compared the performance of SmartVote-Core and SmartVote-C in terms of precision and recall for each cycle of iteration, as shown in Figure 5.5a. Although SmartVote-C took a long time to converge, i.e., 7 rounds of iteration (SmartVote-Core only required 4 rounds of iteration), it consistently achieved higher performance in each round of iteration.

Object Popularity

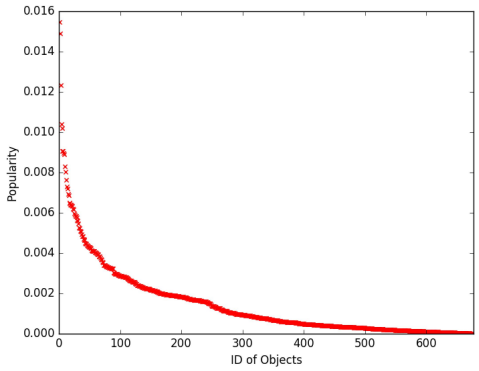
By considering the different popularity of objects, SmartVote-P performed better in terms of accuracy with nearly no extra execution time cost. This is because more sources provide claims on popular objects, and more evidences can be obtained to model the endorsement



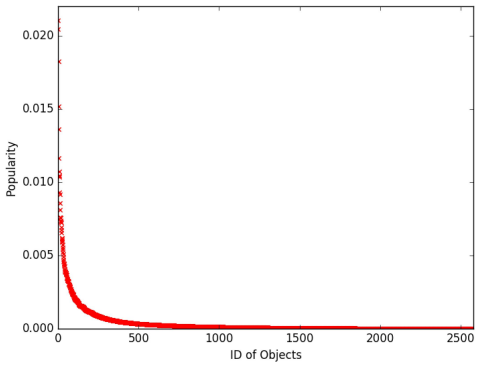
(a) Performance comparison between SmartVote-Core and SmartVote-C during the iteration.



(b) Performance of SmartVote-L under varying uncertainty factor $\beta_{\mathscr{L}}$.



(c) Distribution of object popularity in Book-Author dataset.



(d) Distribution of object popularity in Parent-Children dataset.

Fig. 5.5 Impact of Different Concerns

among sources. Therefore, when computing source reliability, assigning more weights to the popular objects would lead to better truth discovery. In addition, object popularity is calculated directly from the multi-sourced data. Since this calculation is outside of the iteration, it can be conducted effectively under linear execution time. Another observation was that SmartVote-P achieved higher weighted accuracy than traditional accuracy. This is consistent with our expectation that source reliability evaluation relies on the claims provided on popular objects. By differentiating the popularity of objects, our approach obtained more precise results. By ranking objects in the Book-Author dataset and the Parent-Children dataset, respectively, in a descending order of their popularity degrees, we draw scatter diagrams as shown in Figure 5.5c and 5.5d, where each point depicts an object with the corresponding popularity degree (totally, there are 677 objects in the Book-Author dataset and 2579 objects in the Parent-Children dataset). We observed that in both scatter diagrams, the points with very high popularity degrees are quite sparse, indicating that only very few objects are more popular than the majority.

To further validate SmartMTD, we compared SmartMTD with MBM (the best baseline method) on the top-20 popular objects in the ground truth of the Book-Author dataset. SmartMTD returned false values on 2 objects (Book *id* : 9780072499544 and Book *id* : 9780071362856) while MBM made mistakes on 4 objects (Book *id* : 9780028056005, Book *id* : 9780072499544, Book *id* : 9780071362856, and Book *id* : 9780072843996), demonstrating that SmartMTD had better accuracy on the more popular objects. SmartMTD and MBM both returned false values on Book *id* : 9780072499544 and Book *id* : 9780071362856 because some authors are neglected by all the sources.

Source Confidence

We can see from Figure 5.4 that SmartVote-Con performed better than SmartVote-Core in terms of precision and F_1 score while keeping the recall unchanged. The reason for the better

performance of SmartVote-Con is that in real-world datasets, the total number of distinct values of an object provided by all the sources is generally much larger than the number of positive claims of a specific source. Thus, sources normally make more negative claims than positive claims, and show different confidence for these two types of claims. Neglecting this type of differences and strictly conducting the mutual exclusion would certainly increase the false negatives of the truth discovery methods. On the other hand, rebalancing and quantifying the confidence scores of sources for their positive claims and negative claims according to their distributions in the datasets make our approach closer to the reality.

Long-Tail Phenomenon on Source Coverage

Incorporating the balancing long-tail phenomenon on source coverage component dramatically increased the precision of SmartVote-Core with only a slight decrease in recall, resulting in a higher value of F_1 score. In reality, different sources often cover different objects. For the case that a source covers an object while the other source ignores it, the source reliability will be under-estimated if we directly consider there is no agreement between the two sources. On the other hand, the source reliability may be over-estimated, if we use the average endorsement degree between the two sources on the commonly covered objects to measure the overall endorsement degree between the sources. Our approach tackles this issue by modeling the claim distributions, and experimental results validate the effectiveness of our approach. We also investigated the performance of our approach by tuning the values of the uncertainty factor $\beta_{\mathscr{L}}$ from 0 to 1 (as shown in Fig. 5.5b). We found that the precision stayed stable for varying values of $\beta_{\mathscr{L}}$, while the recall peaked at the points when $\beta_{\mathscr{L}}$ equals to 0.1 and 0.2. This implies that over-estimating the endorsement degree between sources in the aforementioned case would impair the recall of our approach due to the over-estimation of negative precision of the sources.

5.5 Related Work

Due to the significance of the veracity to *Big Data*, truth discovery has been a hot topic and studied actively for years in the database community [179, 180]. Aiming at resolving the conflicts among the multi-sourced data, and determining the underlying true values, significant research efforts have been conducted and many methods have been proposed for truth discovery in various application scenarios (see [47–49] for surveys). Recently, Popat et. al [181] propose an approach for early detection of emerging claims, which copes with textual claims. This is a very interesting direction of truth discovery, but out of the scope of this work.

Despite active research in the field, multi-valued truth discovery (MTD) is rarely studied by the previous work. LTM (Latent Truth Model) [46], a probabilistic graphical model based method, is the first solution to the MTD. In this work, Zhao et al. measure two types of errors (false positive and false negative) by modeling two different aspects of source reliability (*specificity* and *sensitivity*) in a generative process. The disadvantage is, LTM makes strong assumptions about prior distributions for nine latent variables, rendering the model inhibitive and intractable to incorporating various implications to improve its performance. Pochampally et al. [41] study various correlations among sources by taking information extractors into consideration, the application scenario is different from ours. Their experiments show that their basic model without considering source correlations sometimes performs worse than LTM, while in our experiments, SmartMTD constantly achieves considerably better results than LTM. To rebalance the distributions of positive claims and negative claims and to incorporate the implication of values' co-occurrence in the same claims, Wang et al. [125] propose a probabilistic model that takes multi-valued objects into consideration. However, this method also requires initialization of multiple parameters, such as prior true or false count of each object, and prior false positive or true negative count of each source. Waguih et al. [48] conclude with extensive experiments that

these probabilistic graphical model-based methods cannot scale well. Zhi et al. [126] also consider the mutual exclusion between sources' positive claims and negative claims, but they model the silence rate of sources to tackle the possible non-truth objects rather than multi-valued objects. To relax unnecessary assumptions, Wang et al. [116] analyze the unique features of MTD and propose an MBM (Multi-truth Bayesian Model), which incorporates source confidence and finer-grained copy detection techniques in a Bayesian framework. However, they assume that false information is copied among sources and correct information is provided independently by sources. Recently, Wang et al. [127] design three models (i.e., the *byproduct* model, the *joint* model and the *synthesis* model) for enhancing existing truth discovery methods. Their experiments show that those models are effective in improving the accuracy of multi-valued truth discovery using existing truth discovery methods. However, LTM and MBM still performed better than those enhanced methods. Wan et al. [128] propose an uncertainty-aware approach for the real-world cases where the number of true values is unknown. However, they cope with continuous data rather than categorical data.

Different from those methods, SmartVote is a graph-based method [175], which incorporates four implications. In particular, SmartVote has four features: i) SmartVote is the first to take the impact of object popularity on source reliability into consideration; ii) instead of assuming independence of sources (in LTM) or independent copying relations among sources (in MBM), SmartMTD globally models copying relations by constructing graphs of all sources that provide values on a specific object; iii) different from the previous copy detection approaches, including MBM and other methods [45, 38, 131, 41], which only consider the copying relations among sources, SmartMTD not only punishes *malicious copiers* that make the same faults with the sources from which they copy, but also defines a new source relation, named *supportive relation*, implying that sources support each other by providing the same true values; iv) as long-tail phenomenon on source coverage is not rare in

reality, SmartVote additionally deals with this significant issue by avoiding sources with few claims from being assigned extreme reliability.

5.6 Summary

In this chapter, we focus on the problem of discovering true values for multi-valued objects (or MTD), which has rarely been studied in the truth discovery community. Based on Chapter 4, to further improve the accuracy of source reliability estimation and predict truth for multi-valued objects, we propose a full-fledged graph-based model, *SmartVote*, by incorporating four implications including *two types of source relations* (i.e., *supportive relations* and *copying relations*), *object popularity*, *loose mutual exclusion*, and *long-tail phenomenon on source coverage*. In particular, we construct $\pm$*supportive agreement graph*s to model the endorsement among sources on their positive and negative claims, from which the improved evaluations of two-sided source reliability are derived. Copying relations among sources are captured by constructing the $\pm$*malicious agreement graph*s based on the consideration that sources sharing the same false values are more likely to be dependent. We consider the popularity of objects and develop techniques to quantify object popularity based on object occurrences and source coverage. We apply source confidence scores to differentiate the extent to what a source believes its positive claims and negative claims. For the ubiquitous long-tail phenomenon on source coverage, we also add smoothing weights to the $\pm$supportive agreement graphs to avoid the reliability of small sources from being over- or under-estimated. Experimental results show that our approach outperforms the art-of-the-state truth discovery methods on two large real-world datasets.

In Chapter 4 and Chapter 5, we have proposed approaches for MTD problems. To fully leverage the advantages of the existing truth discovery methods and achieve more robust and better truth discovery, we will try to extract truth from the results of those methods in Chapter 6.

Chapter 6

An Ensemble Approach For Better Truth Discovery

Previously, many methods have been proposed to tackle truth discovery. However, none of the existing methods is a clear winner that consistently outperforms the others due to the varied characteristics of different methods. In addition, in some cases, an improved method may not even beat its original version as a result of the bias introduced by limited ground truths or different features of the applied datasets. To realize an approach that achieves better and robust overall performance, we propose to fully leverage the advantages of existing methods by extracting truth from the prediction results of these existing truth discovery methods. In particular, we first distinguish between the *single-truth* and *multi-truth* discovery problems and formally define the ensemble truth discovery problem. Then, we analyze the feasibility of the ensemble approach, and derive two models, i.e., *serial model* and *parallel model*, to implement the approach, and to further tackle the above two types of truth discovery problems. Extensive experiments over three large real-world datasets and various synthetic datasets demonstrate the effectiveness of our approach. This chapter is based on our research reported in [182].

6.1 Overview

Despite the various truth discovery methods, such as those handling different data types (e.g., categorical and continuous data), and *source dependency* (e.g., copying relation among sources), those considering *source quality* (e.g., source accuracy/recall, specificity, sensitive, and freshness of data) and *object properties* (e.g., the difficulty of and relation between data objects), and those taking into account *value implications* (e.g., *complementary vote*[1]) and *truth properties* (e.g., *multiple truths* and "unknown" truths), no single method can fit or constantly outperform the others in all application scenarios [49] (our experiments on three real-world datasets and various synthetic datasets validate this conclusion). In addition, a recent investigation [47] shows that even an improved method does not always beat its original version.

Although an appropriate truth discovery method can be selected for each specific scenario [49, 48], it is challenging to find a method that achieves generally good performance due to the technical limitations and biases of each specific method. As the ensemble approach has been proven to be effective for enhancing the robustness and overall performance of algorithms in many disciplines [183], in this work, we study on the feasibility of ensembling existing methods for better truth discovery. Realizing such an ensemble truth discovery approach is a tricky task due to the complexity and diversity of existing truth discovery methods. In a nutshell, we make the following contributions in this work:

- We distinguish between two types of truth discovery problems, i.e., the *single-truth* and *multi-truth* discovery problems, and formally define the ensemble truth discovery problem.

- We analyze the feasibility of the ensemble truth discovery approach, and propose two models, i.e., *serial* and *parallel model*, to implement the approach.

[1] If a source claims value(s) for a certain object, it implicitly votes against other candidate values of this object.

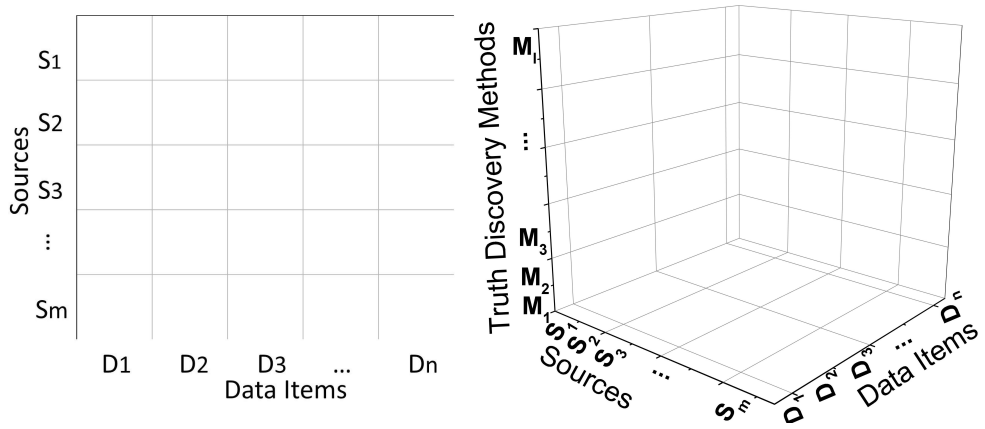
- We empirically evaluate our ensemble approach. Extensive experimental results show that our approach outperforms traditional methods on both real-world and synthetic datasets. In particular, the synthetic datasets with complete ground truths show the improved performance of the ensemble approaches without being biased by the sparsity of limited ground truths.

The rest of the chapter is structured as follows. Section 6.2 reviews the related work. Section 6.3 defines the ensemble truth discovery problem. Section 6.4 analyzes the feasibility of the ensemble approach and presents two implementation models, namely the serial and parallel models. Finally, we report the experimental results in Section 6.5, and provide some concluding remarks in Section 6.6.

6.2 Related Work

Truth discovery has been actively studied by the data integration community in the last few years. Early methods for tackling this issue consist of taking the mean, median for continuous data, and majority voting for categorical data. These methods commonly neglect sources' quality differences, treat every source equally, and are therefore inaccurate in cases where the majority of sources provide false values. Based on this consideration, various methods incorporate source quality by applying a general principle: a source is more trustworthy if it provides more truths; meanwhile, a value has a bigger possibility of being selected as truth if it is claimed by more high-quality sources.

A recent survey [47] tests the performance of several methods on two real-world datasets, which shows that no single method always outperforms the others, and nearly half of the mistakes in the best truth discovery results can be avoided if the trustworthiness of sources is known in apriori. More surveys and experimental studies in [48] and [49] show the potential of improving the usability and repeatability of existing truth discovery methods via



(a) Truth Discovery (b) Ensemble Approach

Fig. 6.1 Input Dimension Comparison of the Original and Ensemble Truth Discovery

an ensemble approach. To the best of our knowledge, [136] is the only work that applies an ensemble approach in truth discovery. It proposes two ensemble methods, i.e., *Uniform Weight Ensemble* (*UWE*) and *Adjusted Weight Ensemble* (*AWE*), and proves that the ensemble approach can generally mitigate the biases introduced by sparse ground truth and outperform the traditional methods. Our work is the first to formally define the ensemble truth discovery problem and to provide in-depth comparisons of different ensemble methods over both single-truth and multi-truth scenarios.

6.3 Problem Formulation

For the input of truth discovery, suppose M data sources (e.g., "*Wikipedia*"), $\mathbf{S}=\{S_1, S_2, \ldots, S_m\}$, provide values on N data items (e.g., "*the cast of Harry Potter*"), $\mathbf{D}=\{D_1, D_2, \ldots, D_n\}$. This input data can be visualized as an $M \times N$ data matrix (Fig. 6.1a). Each cell represents a *claim* that describes the value(s) claimed by a source on a data item (e.g., a claim "*July 9, 1956*" for the data item "*the birthday of Tom Hanks*" provided by source "*Wikipedia*"). The values in the cells of the same columns may conflict due to the different reliability of sources. The objective of the truth discovery problem is to predict the truth(s) for each data item

(corresponding to a column), given the noisy data matrix, while estimating the reliability of each source (corresponding to a row). Since the numbers of true values may vary among data items in practice, e.g., "*the birthday of Tom Hanks*" contains only one date, but "*the cast of Harry Potter*" includes a team of actors, the truth discovery problem can be classified into two categories: i) if we make the single-truth assumption by treating the values in each cell (claim) of the matrix as a joint single value, we have the *single-truth discovery problem*; and ii) if we relax the assumption by treating each distinct value individually, meaning either each cell or the truths may involve several values, we have the *multi-truth discovery problem*. LTM [46] and MBM [116] are the only two methods that are applicable for multi-truth discovery, while all the rest belongs to single-truth discovery methods.

The input of the ensemble truth discovery problem can be formulated as adding a third dimension to the aforementioned data matrix, resulting in a cube (see Fig. 6.1b). The third dimension represents different truth discovery methods, which is denoted as $\mathbf{M}=\{M_1, M_2, \ldots, M_l\}$. Each cell of the cube contains values and their corresponding labels (true or false) provided by the corresponding method. For the single-truth discovery methods, they provide the same label to the value(s) in the same cell, while the multi-truth discovery methods label the value(s) individually. As the methods may have differed performance given a specific application scenario, their results may be conflicting and of varied quality. We formally define the ensemble truth discovery problem as follows:

Definition 6.3.1. ***Ensemble Truth Discovery Problem*** *Given a 3-dimensional matrix (or cube),* $\mathbf{L}$ *truth discovery methods provide boolean labels on values claimed by* $\mathbf{M}$ *sources on* $\mathbf{N}$ *data items, the objective is to predict the truth of the* $\mathbf{N}$ *data items, while estimating the quality of different methods and sources.* □

6.4 Ensemble Approaches

6.4.1 Feasibility Analysis

Berti-Equille implements four approaches including *Simple Bayesian Ensemble* (SBE) [184], *Majority Voting* (MVE), *Uniform Weight* (UWE) and *Adjust Weight* (AWE) ensembles for combining twelve single-truth discovery methods. These approaches are straightforward, which simply unify the outputs of existing methods to the format of a triple {data item, true value(s), veracity score} and combine them directly. Although they are applicable for most of the existing methods, they neglect the useful intermediate results, such as source reliability obtained by the truth discovery methods, thus resulting in limited performance. Moreover, as one of the twelve combined methods, LTM is a special method which incorporates the enriched meaning of source reliability and can tackle multi-truth discovery problem. Naively combining LTM with other single-truth discovery methods and neglecting the two categories of truth discovery problems may further deteriorate the effectiveness of ensemble approaches. In this section, we analyze the feasibility of the ensemble approach and present the possible ways of ensembling the existing methods as follows.

Parallel Model. Although the output formats of existing truth discovery methods vary from one another, they can be transformed into the same format. Therefore, a possible way to ensemble the existing methods is to combine their outputs in a different manner, i.e., *parallel model* (to be detailed in Section 6.4.2).

Serial Model. As aforementioned, the existing methods realize truth discovery following the same general principle. Despite their different ways of implementations, they are generally mutually convertible in their ways of implementations. In particular, both the parameter inference in probabilistic graphical model based methods and the coordinate descent in optimization based methods require updating rules iteratively, which show their potential to be converted into iterative methods; meanwhile, some iterative methods can be formulated

as parameter inference tasks or optimization problems. Thus, we can consider using one method's output as another method's input for initializing on the priors, forming the *serial model* (Section 6.4.3).

For either of the above models, we introduce two methods for the two categories of truth discovery problems, i.e., *single-truth discovery ensemble* (S-ensemble) and *multi-truth discovery ensemble* (M-ensemble).

6.4.2 Parallel Model

The parallel model unifies the format of and combines their outputs to ensemble existing methods. The ensemble truth discovery problem differs from the traditional truth discovery problem in that it takes 3-dimensional rather than 2-dimensional matrix data as inputs. To realize the parallel ensemble model, we first reduce the dimension of the ensemble problem by regarding each distinctive $(Source, Method)$ pair as a virtual data source. Therefore, a value associated with a large number of $(Source, Method)$ pairs indicates that it is either supported by many sources or predicted as truth by various truth discovery methods. As each method only provides Boolean values to the values provided by sources, we can further remove the values labeled as false to reduce the solution space. After such reduction, the ensemble problem is converted into a traditional truth discovery problem and can be handled using existing methods.

Parallel S-ensemble. This approach first runs all the existing methods and formulates their outputs into a 3-dimensional matrix. Then, it trims the matrix by applying the above-mentioned reduction operations. Finally, it applies one of the existing truth discovery methods on this trimmed matrix to deliver the final results. We call these parallel S-ensemble methods "*PS-Method*" (e.g., PS-Accu). Specially, though there is no copying relation among the original methods, there might be complex latent relations among the sources. In such cases, the source dependence-aware methods, e.g., AccuCopy, are applicable for implementing

the ensemble. This is another difference between our work and UWE/AWE, as they simply ensemble the outputs of the methods, and consider the methods to be combined as virtual sources without considering data sources. Thus, they neglect the copying relations among sources.

Parallel M-ensemble. This approach first revises the existing methods under the single-truth assumption so that they can be applied to the multi-truth discovery scenario[2]. In particular, it treats the values in each cell of the matrix individually, and run the original methods to output source reliability. Then, it counts the number of values provided by each source on each data item, and calculates the truth probability of each number as follows:

$$P^*_{D_i}(n) = \sqrt[|S_{D_i}|]{\prod_{n_s=n, s\in S_{D_i}} A(s) \prod_{n_s\neq n, s\in S_{D_i}} (1-A(s))} \tag{6.1}$$

where $P^*_{D_i}(n)$ is the unnormalized probability[3] of truth number n of data item D_i, S_{D_i} is the set of sources which provides values on D_i, n_s is the number provided by source s, and $A(s)$ is the reliability of s. For each data item, it chooses the number with the biggest probability as the number of true values (denoted as N) and output the top-N values instead of choosing the value with the biggest confidence score as the outputs. It revises, if necessary, and runs all the truth discovery methods, formulates and trims their outputs as a 3-dimensional matrix. Finally, both the existing multi-truth discovery methods (LTM or MBM) and the revised single-truth discovery methods can be applied to this matrix to address the ensemble problem. We call these parallel M-ensemble methods "*PM-Method*" (e.g., PM-Accu).

6.4.3 Serial Model

As an alternative, we can sequentially combine the existing methods, i.e., using one method's outputs as another method's a priori inputs to implement the ensemble approach leading to

[2]Hereafter we call the revised methods the modified single-truth discovery methods.

[3]Such values are then normalized to represent probabilities.

the serial ensemble model. Here, we simply omit the consideration of the impact of different orders of the single-truth discovery methods on the performance of the ensemble approach, but leave further research on this issue to our future work.

Most existing methods initialize source reliability by assigning uniform weights among the sources. There are some potential disadvantages of the uniform initialization: firstly, with uniform initialization of source reliability, the performance of methods may rely on the majority. This strategy works well for the case that the majority of sources are good. However, the real scenarios usually are not the case, as sources may copy from each other or provide out-of-date information. Moreover, when we apply truth discovery on challenging tasks, such as information extraction and knowledge graph construction, most of the sources are unreliable. For example [103] describes that in their task that "62% of the true responses are produced only by one or two of the 18 systems (sources)"; secondly, for the scenario where tie cases (i.e., each source claims a unique value on a data item) exists, the results of the methods using uniform initialization are generally unrepeatable. This is because, for the tie cases, the methods would perform voting or averaging like operations and choose a random value as the truth at the beginning of the iteration, leading to randomized source reliability estimation. In contrast, "knowing the precise trustworthiness of sources can fix nearly half of the mistakes in the best fusion results" [47]. Both the above observations motivate us to ensemble existing methods based on a serial ensemble model, which utilizes the source reliability predicted by one method as the prior for initializing another method.

Serial S-ensemble. The sequence of combining the existing methods is a permutation problem. In this work, we randomly choose the methods one by one, and use the source reliability predicted by a method to initialize its direct successor method. We call the serial S-ensemble methods "*SS-#*" (e.g., SS-3).

Serial M-ensemble. We adapt the single methods, when necessary, by using the same operations designed for parallel M-ensemble. Then, we run the revised methods in the same

order as applied for serial S-ensemble. Similarly, we call the serial M-ensemble methods "*SM-#*" (e.g., SM-3).

6.5 Experimental Evaluation

In this section, we compare our ensemble approaches with existing truth discovery methods by conducting extensive experiments on both real-world and synthetic datasets. Specifically, we first introduce the experimental setup including baselines and performance measures (in Section 6.5.1). Then we present the experimental results on real-world datasets in Section 6.5.2 and that on synthetic datasets in Section 6.5.3. Additionally, we study the impact of combining different numbers of methods by applying the serial ensemble model in Section 6.5.4.

6.5.1 Experimental Setup

We compared our approaches with three groups of truth discovery methods.

Original Single-Truth Discovery Methods (STD). We chose five typical and competitive algorithms from this category for the comparison. Note that Sums was revised by incorporating complementary vote.

- *Voting.* For each item, it predicts the most frequently provided claim as the estimated truth(s) without iteration.
- *Sums, Avg-Log, TruthFinder, 2-Estimates.* All these methods iteratively evaluate source reliability and claims alternately from each other using different calculation methods.

Multi-Truth Discovery Methods (MTD). There are two existing multi-truth discovery methods:

- *LTM*. Based on a probabilistic graphical model, it recognizes a value as true if its veracity score exceeds 0.5.

- *MBM*. This method incorporates a new mutual exclusion definition for multi-truth discovery from the reformatted claims.

Modified Single-Truth Discovery Methods (MMTD). We adapted four representative single-truth discovery methods for the multi-truth scenario by applying the operations described in Section 6.4.2, resulting in four new methods, namely *Voting**, *Sums**, *Average-Log**, *TruthFinder**, and *2-Estimates**.

Based on the above representative methods, we derived methods following our ensemble approaches as follows:

- *Parallel S-Ensemble Group*. It contains five methods, i.e., *PS-Voting*, *PS-Sums*, *PS-AvgLog*, *PS-TruthFinder*, and *PS-Estimates*.

- *Parallel M-Ensemble Group*. It consists of seven methods, i.e., *PM-LTM*, *PM-MBM*, *PM-Voting**, *PM-Sums**, *PM-AvgLog**, *PM-TruthFinder**, and *PM-2Estimates**.

- *Serial S-Ensemble Group.* As Voting does not consider source reliability, we combined the other four single-truth discovery methods and implemented *SS-4*. We combined the four methods in the following order: Sums, Avg-Log, TruthFinder, and 2-Estimates[4], and compared *SS-1* through *SS-4* by gradually adding one method each time in Section 6.5.4.

- *Serial M-Ensemble Group.* We combined six methods in the following order: Sum*, Avg-Log*, TruthFinder*, 2-Estimates*, LTM, and MBM, to implement *SM-6*. We chose this order for the same reason as SS-4). We compared *SM-1* through *SM-6* in Section 6.5.4.

Table 6.1 Characteristics of Three Real-World Datasets

Book Dataset	Biography Dataset	Movie Dataset
# sources (Websites): 649	# sources (users): 55,259	# sources (Websites): 16
# claims: 13,659	# claims: 227,584	# claims: 33,194
attribute: author names	attribute: children	attribute: director names
# objects (books): 664	# objects (person): 2,579	# objects (movies): 6,402
ground truths count (GT):	ground truths count (GT):	ground truths count (GT):
86 books (12.95%)	2,578 person (99.9%)	200 movies(3.12%)
Avg Coverage per source: 0.0317	Avg Coverage per source:0.0016	Avg Coverage per source: 0.0625
Avg # distinct values per data item	Avg # distinct values per data item	Avg # distinct values per data item
(conf): 3.2	(conf): 2.45	(conf): 1.2
Avg # claims per source: 21.05	Avg # claims per source: 4.12	Avg # claims per source: 2074.62

We implemented all the above methods in Java 7 and ran experiments on 3 PCs with Intel Core i7-5600 processor (3.20GH$\times$8) and 16GB RAM. The methods were evaluated in terms of three metrics, including *precision*, which is the average percentage of the true positives returned by the methods in the set of all predicted true values on all values of all data items, *recall*, which is the average percentage of the true positives returned by the methods in the set of ground truths on all values of all data items, and F_1 *score*, which is the harmonic mean of precision and recall, from which we can see the comprehensive performance of all the compared methods.

6.5.2 Experiments on Real-World Datasets

In this section, we present the evaluation of our ensemble approaches with respect to the existing methods on three real-world datasets (namely *Book dataset* [44], *Biography dataset* [43], and *Movie dataset* [116], described in Table 6.1), where we have removed the duplicated and invalid records to clean the original datasets.

Table 6.2 shows the evaluation results. For each single method group (i.e., single-truth discovery method group and multi-truth discovery method group, including the modified single-truth methods), no methods consistently outperformed the others on all the real-world datasets, which is consistent with the previous survey studies [49]. Among those single

[4]We chose this order because it is the increasing order of precision of these four methods performed on three real-world datasets in [116].

Table 6.2 Method comparison on real-world datasets and synthetic datasets (The best performance values in each method group are in bold. We consider multi-truth discovery methods and modified single-truth methods as one group. The best performance values among our ensemble approaches are highlighted in the gray background).

Group	Method	Book			Biography			Movie			Syn.(R)	Syn.(80P)
		Prec.	Recall	F_1	Prec.	Recall	F_1	Prec.	Recall	F_1	Corr. Rate	Corr. Rate
STD	Voting	**0.837**	0.328	0.471	0.876	0.855	0.865	**0.91**	0.292	0.442	**0.321**	0.581
	Sums	**0.837**	0.54	0.656	0.859	0.881	0.87	0.847	0.591	0.696	0.319	0.623
	AvgL.	0.826	0.605	0.698	0.904	0.886	0.895	0.847	0.643	0.731	0.317	0.58
	TruthF.	**0.837**	0.605	0.702	0.905	0.886	0.895	0.847	**0.71**	**0.772**	**0.32**	0.62
	Est.	**0.837**	**0.621**	**0.713**	**0.908**	**0.888**	**0.898**	0.863	0.692	0.768	0.319	**0.626**
MTD	LTM	**0.826**	0.651	0.728	0.91	0.88	0.895	0.812	0.813	0.812	0.225	0.223
	MBM	**0.826**	**0.744**	**0.783**	**0.915**	**0.89**	**0.902**	0.852	**0.833**	**0.842**	**0.32**	0.533
MMTD	Voting*	0.756	0.638	0.692	0.873	0.851	0.862	0.864	0.523	0.652	0.318	0.586
	Sums*	**0.826**	0.644	0.724	0.905	0.887	0.896	0.81	0.534	0.644	0.319	0.623
	AvgL*	0.663	0.709	0.685	0.88	**0.89**	0.885	0.812	0.65	0.722	0.317	0.58
	TruthF.*	0.698	0.709	0.703	0.876	0.88	0.878	0.853	0.723	0.783	**0.32**	0.623
	Est.*	**0.826**	0.734	0.777	0.89	0.88	0.885	**0.865**	0.722	0.787	0.319	**0.626**
PS-ens.	PS-Voting	**0.837**	0.63	0.719	**0.905**	**0.886**	**0.895**	0.915	0.75	0.824	**0.323**	**0.632**
	PS-Sums	**0.837**	**0.64**	**0.725**	**0.905**	**0.886**	**0.895**	0.92	0.78	0.844	0.322	0.631
	PS-AvgL.	**0.837**	0.638	0.724	**0.905**	**0.886**	**0.895**	0.92	0.78	0.844	0.322	**0.632**
	PS-TruthF.	**0.837**	**0.64**	**0.725**	**0.905**	**0.886**	**0.895**	**0.927**	0.792	0.854	0.322	0.631
	PS-Est.	**0.837**	**0.64**	**0.725**	**0.905**	**0.886**	**0.895**	0.925	**0.816**	**0.867**	0.322	0.631
PM-ens.	PM-Voting*	**0.86**	0.754	0.804	0.91	0.9	0.905	0.899	0.821	0.858	0.321	0.627
	PM-Sums*	0.827	0.751	0.787	0.91	0.89	0.9	0.883	0.833	0.857	0.32	0.627
	PM-AvgLog*	0.829	0.763	0.795	0.915	0.897	0.906	0.886	0.833	0.859	**0.325**	0.623
	PM-TruthF.*	0.834	0.791	0.812	0.91	0.9	0.905	0.886	0.854	0.87	0.322	0.626
	PM-Est.*	0.842	0.766	0.802	0.92	0.89	0.905	0.904	0.846	0.874	0.32	0.626
	PM-LTM	0.837	0.808	0.822	0.93	0.91	0.92	0.91	**0.86**	0.884	0.322	0.623
	PM-MBM	**0.86**	**0.812**	**0.836**	**0.93**	**0.92**	**0.925**	**0.922**	0.85	**0.885**	0.32	**0.628**
SS-ens.	SS-4	0.837	0.721	0.775	0.91	0.9	0.905	0.87	0.753	0.807	0.325	0.628
SM-ens.	SM-6	0.836	0.764	0.798	0.93	0.92	0.925	0.913	0.866	0.889	0.321	0.563

methods, Voting almost always achieved the best precision. As the data items in all the three real-world datasets involve multiple true values, LTM and MBM generally achieved better performance than the original single-truth discovery methods, esp. in recall and F_1 score. The modified single-truth discovery methods also achieved relatively higher precision and recall than their original methods. The original single-truth discovery methods showed higher precision but achieved lower recall than multi-truth discovery methods. This indicates that the original single-truth discovery methods tend to underestimate the number of true values.

Both our parallel ensemble methods, i.e., PM and PS, returned better results than the element methods. The serial ensemble methods, i.e., SS-4 and SM-6, also showed relatively better performance. In particular, both PM and SM-6 (resp., PS and SS-4) outperformed the original multiple (resp., single) truth discovery methods they combined in terms of precision, recall and F_1 score on all the three real-world datasets. In our experiments, five single-truth discovery methods are combined for PS and seven multi-truth discovery methods are combined for PM. The obtained 3-dimensional matrices are not significantly different

from each other, which resulted in the outcome that all PM and PS methods show similar performance. Due to the existence of multiple true values in the datasets, PM and SM-6 methods performed better than PS and SS-4 methods. However, neither the SM-6 nor the PM methods could consistently dominate the other, and the results are different among different datasets. Similar situations occurred when we compared SS-4 with PS. Further performance studies of SS and SM will be presented in Section 6.5.4.

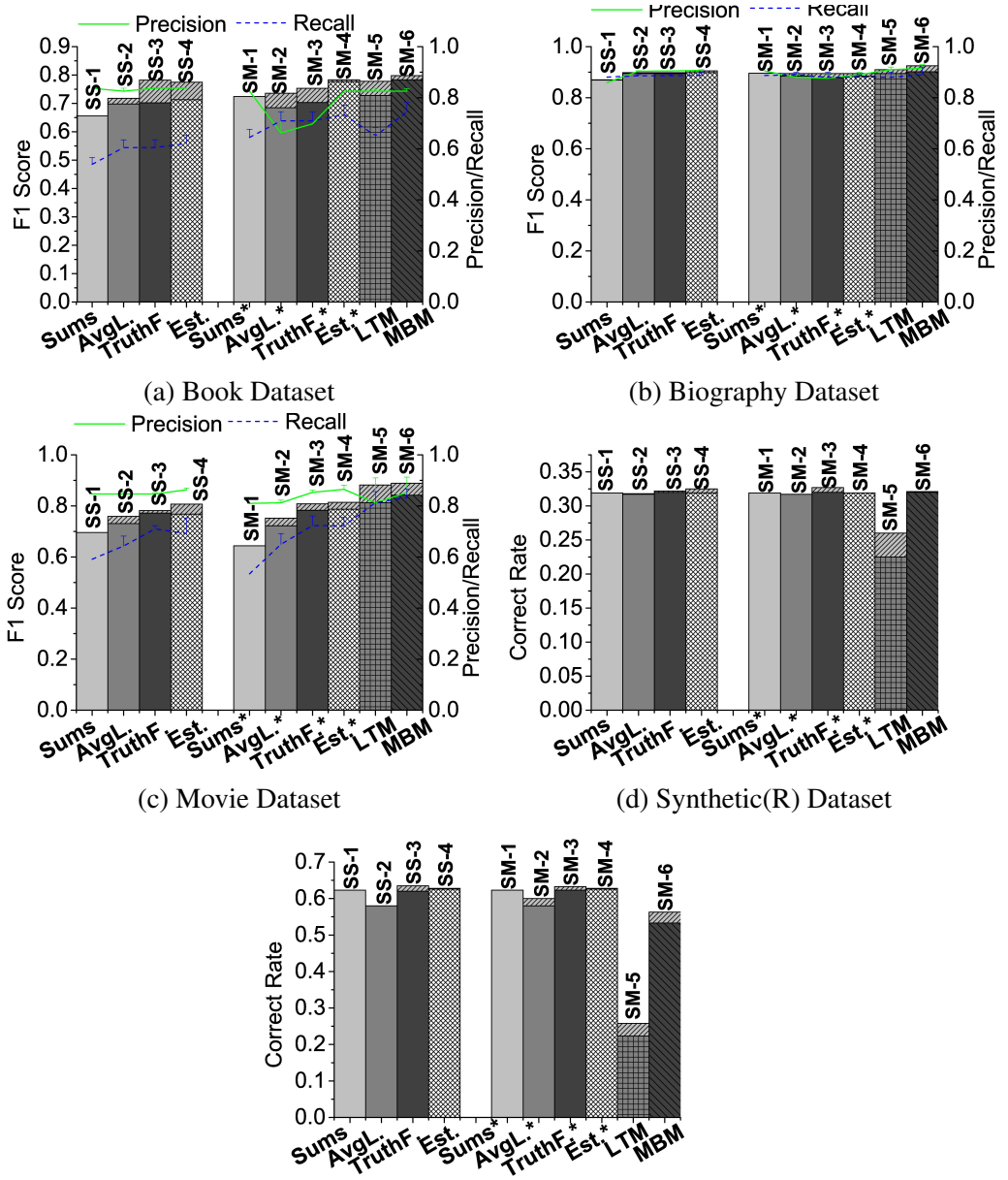
6.5.3 Experiments on Synthetic Datasets

Due to the limited ground truths of real-world datasets, the performance evaluation may be biased by the available ground truth. In this section, we present the comparison of our approaches with the element methods on synthetic datasets with a wide spectrum of distribution settings and complete ground truths. We first generated synthetic datasets by applying the dataset generator proposed by Waguih et al [48]. This generator contains six parameters that can be configured to simulate a wide spectrum of truth discovery scenarios. Three parameters, namely the number of sources (M), the number of data items (N), and the number of distinct values per data item (V), determine the scale of the generated dataset, while the other three parameters, source coverage (cov), ground truth distribution per source (GT), and distinct value distribution per data item ($conf$), determine the characteristics of the generated dataset.

We fixed the scale parameters by setting $M = 50$, $N = 1,000$, and $V = 20$, configured both cov and $conf$ to follow exponential distributions. In particular, we chose two distributions (i.e., the random[5] and 80-pessimistic[6] distributions) for GT. We chose these distributions as they are closest to the real world scenarios. Specifically, for the exponential distribution of $conf$, the majority of data items have few distinct values while few data items have

[5]Random ground truth distribution per source means the number of true positive claims per source is random.

[6]80-pessimistic ground truth distribution per source means 80% of the sources provide 20% true positive claims, while 20% of the sources provide 80% true positive claims.



(a) Book Dataset

(b) Biography Dataset

(c) Movie Dataset

(d) Synthetic(R) Dataset

(e) Synthetic(80P) Dataset

Fig. 6.2 Impact of combining different numbers of single methods on SS and SM. The offsets on the precision and recall lines are the corresponding precision and recall of the corresponding SS and SM methods, while the upper bounds of the stack columns are the corresponding F_1 score of the corresponding SS and SM methods.

many conflicts. For the case of exponential source coverage, most sources claim values for few data items whereas few sources cover the majority of data items. When we face with the challenging task of information extraction and knowledge base construction, the majority of sources are always error-prone, and truths are maintained by the minority. Therefore, random and 80-pessimistic *GT* distributions are more representative. Based on the above configurations, we obtained two types of synthetic datasets, namely *Synthetic(R)* and *Synthetic(80P)*, each containing 10 datasets. The metrics of each method were measured as the average of 10 executions over the 10 datasets included by the same dataset type.

Table 6.2 shows the performance comparison of different methods on the synthetic datasets. As each data item in the synthetic datasets has only one single true value, every method predicted values for all the data items. In this case, we specially measure the methods in terms of *correct rate* by computing the percentage of matched values between each method's output and ground truths. Specifically, the experimental results show almost the same pattern with those on the real-world datasets, which confirms that the ensemble approaches indeed lead to more accurate truth discovery. As sources in Synthetic(R) claim random numbers of true positive values, all methods returned low-quality results for this dataset with correct rate kept around 0.32. Our ensemble methods only showed slightly better performance. The multi-truth discovery methods, especially LTM, failed to return good results on both datasets, where each data item has only one single true value. This is also the reason why SM-6 and PM methods performed worse than SS-4 and PS.

6.5.4 Impact of Method Numbers on Serial Ensemble Model

To analyze the impact of the number of methods (which are used to derive the ensemble approaches) on the two serial ensemble models (i.e., SS and SM), we conducted experiments on all the above datasets. In particular, we studied the performance of the serial ensemble methods by gradually adding one method each time. We combined the existing methods in

the same order as described in Section 6.4.3, where SS-1 is the same as Sums, the source reliability output by Sums was used as the input of AverageLog to realize SS-2. Following a similar way, we further added TruthFinder and 2-Estimates to implement SS-3 and SS-4. Similarly, we gradually combined Sums*, AverageLog*, TruthFinder*, 2-Estimates*, LTM, and MBM to form SM-1 through SM-6. Through the above procedures, we finally obtained four SS methods (from SS-1 to SS-4) and six SM methods (from SM-1 to SM-6).

Fig. 6.2 shows the performance of SS, SM, and the applied existing methods. In particular, the precision, recall and F_1 score of SS and SM fluctuated on all the real-world datasets, and the correct rate of them fluctuated on all the synthetic datasets, while we gradually combined more methods. Each serial ensemble method outperformed the last combined method except the special case of SS-1 (exactly Sums) and SM-1 (exactly Sums*), where the two methods are the same. This indicates that naively and serially combining more methods does not necessarily improve the effectiveness of the serial ensemble methods in a proportional manner. However, the accuracy of a single-truth discovery method could be improved by using the source reliability predicted by other methods as inputs. This indicates parallel ensemble model is generally better than serial ensemble model in obtaining the best ensemble performance.

6.6 Summary

In this chapter, we focus on the problem of ensembling the existing truth discovery methods for more robust and consistent truth discovery. Several surveys have shown that a "one-fits-all" truth discovery method is not achievable due to the limitations of the existing methods. Therefore, combing various competing methods could be an effective alternative for conducting high-quality truth discovery. Given that very few research efforts have been conducted on this issue, we analyze the feasibility of such an ensemble approach. We propose two novel models, namely *serial model* and *parallel model*, for combining the

truth discovery methods. We further present several implementations based on the above models for both single-truth and multi-truth discovery problems. Extensive experiments over three real-world datasets and various synthetic datasets demonstrate the effectiveness of our ensemble approaches.

In this chapter, we have introduced two models for combining the existing truth discovery methods. To evaluate the truth discovery methods in the cases where ground truth is missing, we will introduce a novel approach in Chapter 7.

Chapter 7

Performance Evaluation on Truth Discovery Methods

Although many truth discovery methods have been proposed based on different considerations and intuitions, investigations show that no single method consistently outperforms the others. To select the right truth discovery method for a specific application scenario, it becomes essential to evaluate and compare the performance of different methods. A drawback of current research efforts is that they commonly assume the availability of certain ground truth for the evaluation of methods. However, the ground truth may be very limited or even impossible to obtain, rendering the evaluation biased. In this chapter, we present *CompTruthHyp*, a general approach for comparing the performance of truth discovery methods without using ground truth. In particular, our approach calculates the probability of observations in a dataset based on the output of different methods. The probability is then ranked to reflect the performance of these methods. We review and compare twelve representative truth discovery methods and consider both single-valued and multi-valued objects. Empirical studies on both real-world and synthetic datasets demonstrate the effectiveness of our approach for comparing truth discovery methods. This chapter is based on our research reported in [185, 176].

7.1 Overview

So far, various truth discovery methods [49] have been proposed based on different considerations and intuitions. However, investigations show that no methods could constantly outperform the others in all application scenarios [49, 48, 47]. Moreover, Li et al. [47] demonstrate with experiments that even an improved method does not always beat its original version, such as *Investment* and *PooledInvestment* [43], *Cosine, 2-Estimates* and *3-Estimates* [36]. Therefore, to help users select the most suitable method to fulfill their application needs, it becomes essential to evaluate and compare the performance of different methods.

To evaluate the effectiveness of truth discovery methods, current research usually measures their performance in terms of *accuracy* (or *error rate*), F_1*-score*, *recall*, *precision*, *specificity* for categorical data [48], and *Mean of Absolute Error* (MAE) and *Root of Mean Square Error* (RMSE) for continuous data [49]. All these metrics are measured and compared based on the assumption that a reasonable amount of ground truth is available. However, the fact is, the labor cost of ground truth collection is rather expensive. Ground truth is often very limited or even impossible to obtain (generally less than 10% of the size of the original dataset [48]). For example, the knowledge graph construction [15] involves a large number of objects, making it impossible to have the complete ground truth for performance validation. In addition, it requires enormous human efforts to acquire even a small set of ground truth. The lack of sufficient ground truth can, in many cases, statistically undermine the legitimacy of evaluating and comparing existing methods using the ground truth-based approach. For example, previous comparative studies [38, 45, 47, 118, 46, 40, 41, 116, 125], based on real-world datasets with sparse ground truth, could all bring bias to the performance measurement of the methods. Methods with good accuracy may, by chance, return incorrect results on the particular objects covered by the sparse ground truth, while methods with poor accuracy may, occasionally, be consistent with the sparse ground truth. Moreover, methods

that show the same accuracy on the rather limited objects covered by the sparse ground truth, may have different performance in reality. Under this circumstance, it is hard to conclude which method performs better as we cannot trust the comparison results. This also makes it difficult to select the method with the best performance to be applied to specific application scenarios. Therefore, evaluating the performance of various truth discovery methods with missing or very limited ground truth can be a significant and challenging problem for the truth discovery applications [49]. We identify the key challenges around this issue as the following:

- The only way to obtain evidence for performance evaluation without ground truth is to extract features from the given dataset for truth discovery [47–49]. However, the features of a dataset are sometimes complex, encompasing source-to-source, source-to-object, object-to-value, and value-to-value relations. In addition, it is challenging to find a method to capture those relations without creating additional biases.

- Current truth discovery methods commonly determine value veracity and calculate source trustworthiness jointly. Source trustworthiness and value confidence scores are the common intermediates of the existing methods, which are also the key elements for identifying the truth for each object [49]. Therefore, we can consider identifying the relations among sources, objects, and values by leveraging those measurements to match the relations extracted from the given dataset. However, even if we are able to obtain the features of the given dataset, different truth discovery methods may calculate the source trustworthiness and value confidence scores using different metrics, which have various meanings and require non-trivial normalization.

- Even if we are able to resolve the above two challenging issues, it is still tricky to find appropriate metrics for comparing those features, to fulfill the requirement of method comparison.

In this work, we focus on truth discovery method comparison without using ground truth. In a nutshell, we make the following contributions:

- To our knowledge, we are the first to reveal the bias introduced by sparse ground truth in evaluating the truth discovery methods, by conducting experiments on synthetic datasets with different coverages of the leveraged ground truth.

- We analyze, implement, and compare twelve specific truth discovery methods, including *majority voting, Sums, Average-Log, Investment, PooledInvestment* [43], *TruthFinder* [44], *2-Estimates, 3-Estimates* [36], *Accu* [38], *CRH* [40], *SimpleLCA*, and *GuessLCA* [37].

- We propose a novel approach, called *CompTruthHyp*, to **comp**are the performance of **truth** discovery methods without using ground truth, by considering the output of each method as a **hyp**othesis about the ground truth. CompTruthHyp takes both single-valued and multi-valued objects into consideration. It utilizes the output of all methods to quantify the probability of observation of the dataset and then determines the method with the largest probability to be the most accurate.

- We conduct extensive experiments on both synthetic and real-world datasets to demonstrate the effectiveness of our proposed approach. Our approach consistently achieves more accurate rankings of the twelve methods than traditional ground truth-based evaluation approach.

The rest of the chapter is organized as follows. We review the related work in Section 7.2. Section 7.3 introduces some background knowledge about truth discovery and the observations that motivate our work. Section 7.4 presents our approach. We report our experiments and results in Section 7.5. Finally, Section 7.6 provides some concluding remarks.

7.2 Related Work

Generally, there are two categories of previous studies on performance evaluation and comparison of truth discovery methods. The first category includes novel and advanced approaches for truth discovery in various scenarios. To validate the performance of their proposed approaches and to show how their approaches outperform the state-of-the-art truth discovery methods, those projects conduct comparative studies by running experiments on real-world datasets with manually collected ground truth. Truth discovery is first formulated by Yin et al. [44]. To show the effectiveness of their proposed *TruthFinder*, they conduct experiments on one real-world dataset, i.e., *Book-Author* dataset, which contains $1,263$ objects. The manually collected ground truth only covers 7.91% of the objects. With truth discovery becoming more and more popular, considerable methods [43, 45, 65, 46, 118, 123, 116, 125, 141, 175, 181] have been proposed to fit various scenarios. We find that there is a common limitation of those works that they all conduct experiments on real-world datasets with limited ground truth. Besides the Book-Author dataset, the frequently-used datasets, including *Flight* [47] (covers 8.33% of complete ground truth), *Weather* [45] (74.4%), *Population* [43] (0.702%), *Movie* [46] (0.663%) and *Biography* [43] (0.069%) all feature sparse or low-quality ground truth, which makes the experimental data evaluated on those datasets cannot be fully trusted.

The second category of the studies is presented in surveys [47–49] that aim at investigating and analyzing the strengths and limitations of the current state-of-the-art techniques. In particular, in 2012, Li et al. [47] study the performance of sixteen data fusion methods in terms of precision and recall, on two real-world domains, namely *Stock* and *Flight*. Based on their experiments, the authors point out that the collected ground truth tends to trust data from certain sources, which sometimes puts wrong values or coarse-grained values in the ground truth. Moreover, we find that their constructed ground truth are relatively sparse, with the one for the stock domain covering only $200/1000 = 20\%$ of the complete ground truth, and

the one for the flight domain covering only $100/1200 = 8.33\%$. The most recent survey [49] provides a comprehensive overview of truth discovery methods and summarizes them from five different aspects, but they do not conduct any comparative experiments to show the diverse performance of the methods. Waguih et al. [48] point out that the sparse ground truth is not statistically significant to be legitimately leveraged for the accuracy evaluation and comparison of methods. To the best of our knowledge, they are the first to implement a dataset generator to generate synthetic datasets with the control over ground truth distribution, for the sake of comparing existing methods. Different from their work, our approach tries to evaluate the performance of various truth discovery methods without using ground truth, which is applicable to more general real-world scenarios.

7.3 Preliminaries

Current truth discovery methods take as input some conflicting triples (i.e., a given dataset) in the form of {*source, object, value*}, where *source* ($s \in S$) denotes the location where the data originates, *object* ($o \in O$) is an attribute of an entity, and *value* ($V_{s_o} \subset V$) depicts the potential value set of an object claimed by a source. For example, a triple, {"www.imdb.com", "the director of Beauty and the Beast", "Bill Condon"}, indicates that the website "IMDb" claims that the director of the movie "Beauty and the Beast" is "Bill Condon". If o is a single-valued object, $|V_{s_o}| = 1$. For example, "the age of a person" only has one single value; on the other hand, if o is a multi-valued object, $|V_{s_o}|$ is bigger than 1. For example, a person might have more than one child. Based on the triples, the methods infer a Boolean truth label ("true"/"false") for each triple as the output. Formally, we name the factual value of an object o as the *ground truth* of o, denoted by ${V_o}^*$, and the triple involves o with the label "true" output by a truth discovery method m as the *identified truth* of o, denoted by ${V_o}^m$. After applying a group of truth discovery methods M one by one on the triples, each method $m \in M$ outputs the *identified truth* for each object $o \in O$. The closer ${V_o}^m$ is to ${V_o}^*$ for each object,

Table 7.1 Notations Used in Chapter 7

Notation	Explanation
o, O	An object (resp., Set of all objects), o may be a single-valued or a multi-valued object
s, S	A source (resp., Set of all sources)
v, V	A claimed value (resp., a set of all claimed values)
V_s	The set of all values provided by s
V_o	The set of all claimed values on o
m, M	A truth discovery method (resp., Set of truth discovery methods)
V_{s_o}	The potential value set of o claimed by s
$V_o{}^*$, V^*	The ground truth of o (resp., of the given dataset)
$V_o{}^m$, V^m	The identified truth of o (resp., the given dataset) output by m
V^i	The incomplete ground truth of the given dataset
S_v	The set of sources provide claimed value v on an object
$c_{\mathscr{V}}$	The confidence score of $\mathscr{V}$, $\mathscr{V}$ is a single joint value
τ_s	The trustworthiness of s
ϕ	The observation of which value each source in the given dataset votes for
ϕ_{s_v}	The observation of s providing a particular value v ($v \in V_o$)
ϕ_s	The observation of source s with its claimed values
$P(\phi \mid V^m)$	The probability of ϕ conditioned on V^m
$\tau_s(m)$	Given V^m, the probability that the claimed values of s is true
$P_s(v_t \mid V_o^m)$ (resp., $P_s(v_f \mid V_o^m)$)	Given V^m, the probability that s provides a particular true (resp., false) value for o
$V_s{}^t(m)$, $V_s{}^f(m)$	The set of all true (resp., false) values provided by s, given V^m
$P(\phi_{s_v} \mid V^m)$	The probability of ϕ_{s_v} conditioned on V^m
$P(\phi_s \mid V^m)$	The probability of ϕ_s conditioned on V^m
$\mathscr{C}_m$	The *confidence* of method m

the better the method m performs. We denote the *identified truth* of all objects in O output by method m as V^m ($V_o{}^m \subset V^m$), and the *ground truth* of all objects in O, i.e., the *complete ground truth* of the given dataset, as V^* ($V_o{}^* \subset V^*$). Table 7.1 summarizes the notations used in this chapter. In most cases, the ground truth provided with each frequently utilized real-world dataset, denoted by V^i, is only a subset of the complete ground truth ($V^i \subset V^*$). We define the *coverage* of the ground truth as follows:

Definition 7.3.1. ***Coverage of the Ground Truth*** *indicates the percentage of objects covered by the ground truth over all the objects in the given dataset. The coverage of the complete ground truth is* 100%. □

7.3.1 Ground Truth-Based Evaluation Approach

Given the output of each truth discovery method, i.e., V^m, $m \in M$, and the ground truth (V^i), the traditional ground truth-based evaluation approach evaluates the effectiveness of each method in terms of *precision, recall,* F_1 *score, accuracy/error Rate*, and *specificity*

Table 7.2 Confusion Matrix of Method m

		Ground Truth	
		True	False
Method	True	True Positive (TP_m)	False Positive (FP_m)
	False	False Negative (FN_m)	True Negative (TN_m)

for categorical data. For each metric of each method, the higher the value is, the better the method performs. In particular, to derive those five metrics, the ground truth-based approach first produces a confusion matrix (as shown in Table 7.2) for each method. It cumulatively counts the numbers of true positives, false positives, true negatives, and false negatives for each object o covered by V^i. Then, based on the matrix, it calculates the metrics as follows:

- *Precision* of method m represents the probability of its positive outputs being correct, computed as $\frac{TP_m}{TP_m+FP_m}$.
- *Recall* of method m indicates the probability of true values being identified as true, computed as $\frac{TP_m}{TP_m+FN_m}$. $1-recall$ is the so-called *false negative rate*.
- F_1 *score* of method m demonstrates the harmonious mean (i.e., a weighted average) of *precision* and *recall*, computed as $\frac{2 \cdot precision \cdot recall}{precision+recall}$.
- *Accuracy/Error Rate* of method m is the probability of its outputs being correct, computed as $\frac{TP_m+TN_m}{TP_m+FP_m+TN_m+FN_m}$.
- *Specificity* of method m presents the probability of false values being identified as false, computed as $\frac{TN_m}{FP_m+TN_m}$. $1-specificity$ is the so-called *false positive rate*.

However, as V^i is generally only a very small part of V^*, the distributions of true positives, false positives, true negatives, and false negatives, obtained in this small sample space cannot reflect the real distributions. Therefore, the derived metrics are not statistically significant to be legitimately leveraged for method accuracy evaluation and comparison. We will show the biases introduced by the limited ground truth in Section 7.3.2.

Additionally, most of the existing truth discovery methods assume that each object in the given dataset has only one true value [46]. When multi-valued objects (e.g., "children of a person") exist in the given dataset, they simply concatenate and regard the values provided by the same source as a single joint value. Specifically, given a multi-valued object o ($|V_{s_o}| > 1$), they regard V_{s_o} as a single joint value, denoted as $\mathscr{V}$, instead of considering each claimed value $v \in V_{s_o}$ individually. They label the values in V_{s_o} as all true (i.e., $\mathscr{V}$ is true) or all false (i.e., $\mathscr{V}$ is false) together. Thus, under such assumption, by identifying a value of an object to be true, a truth discovery method is believed to implicitly claim that all the other values of the object are false. When a method incorrectly identifies a false value of an object to be true, it certainly asserts the true value as a false value. In this case, the false positives are equivalent to false negatives, and the recall and F_1 scores equal to the precision.

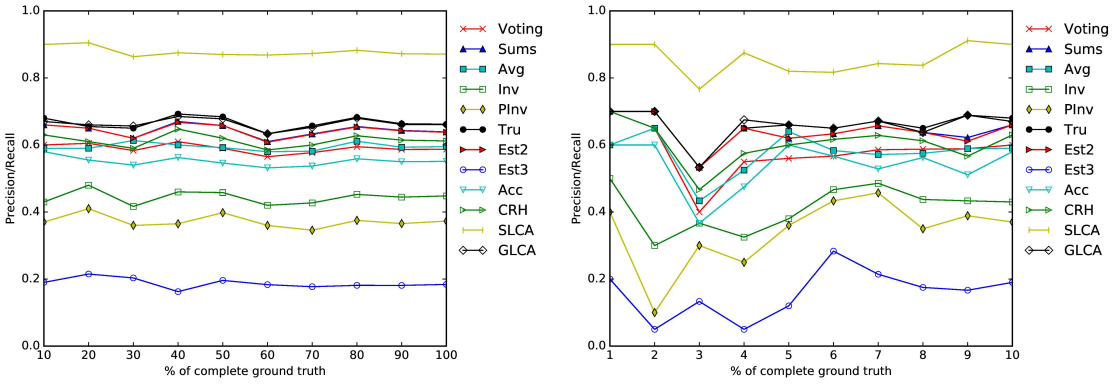
However, when it comes to the case of multi-valued objects, the identified truth of a multi-valued object may overlap with the ground truth. Simply labeling a value set as true or false according to whether it equals to the ground truth will degrade the accuracy of the performance evaluation of the method. For example, if the identified truth for "Tom's children" is {"Anna, Tim"}, and the ground truth is {"Anna, Tim, Lucas"}, the identified truth is partially true, rather than false. Therefore, we propose to treat each value in the identified value set individually. In this case, the false positives are no longer equivalent to false negatives. Neither the precision nor the recall of a method can reflect the performance of the method individually, we need to measure both the accuracy and the completeness of the methods' output. For example, given two methods m_1 and m_2, m_1 identifies {"Anna, Tim"} as "Tom's children", while m_2 identifies {"Anna"} is the only child of "Tom". The precision of both methods is 1, as their identified values are all true values, indicating their performance are the same. However, the recall of m_1 is $\frac{2}{3}$ and that of m_2 is $\frac{1}{3}$, indicating the performance of m_1 is better than m_2.

In this work, we will evaluate the performance of methods separately for *single-valued scenario*s (i.e., scenarios where only single-valued objects exist) and *multi-valued scenario*s (i.e., scenarios where multi-valued objects exist).

7.3.2 Motivation

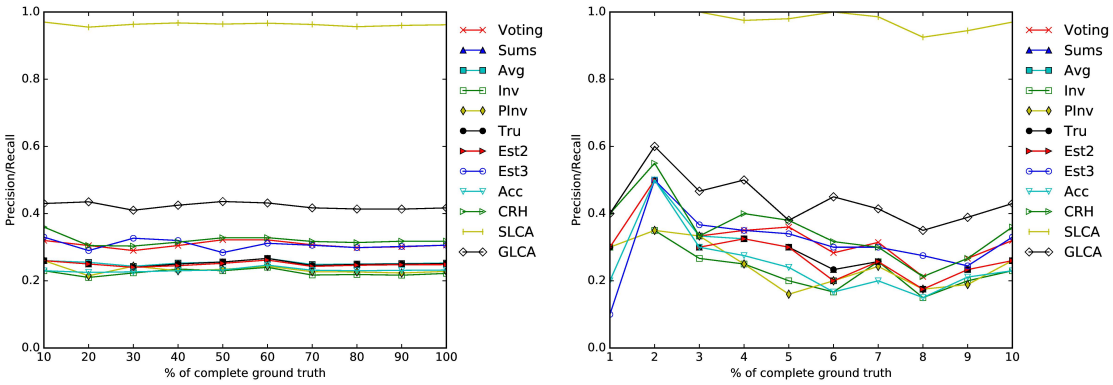
To investigate the bias introduced by the incomplete ground truth on method performance evaluation, we conducted experiments by evaluating twelve truth discovery methods (we will introduce these methods in Section 7.4.1), using synthetic datasets while tuning the *coverage* of the ground truth.

The synthetic datasets with complete ground truth are generated by the dataset generator implemented by Waguih et al. [48]. This generator involves six parameters that are required to be configured to simulate a wide spectrum of truth discovery scenarios. We will introduce the settings of those parameters in detail in Section 7.5.1. We tuned the ground truth distribution per source (GT) for all the seven possible distributions, including *uniform, Random, Full-Pessimistic, Full-Optimistic, 80-Pessimistic, 80-Optimistic*, and *Exponential*. Based on the above configurations, we obtained seven dataset groups, each group containing 10 datasets. The metrics, namely *precision, recall,* F_1 *score, accuracy* and *specificity* of each method were measured as the average of 10 executions over the 10 datasets included by the same dataset group. To calculate the metrics, for each dataset, we tuned the coverage of the ground truth from 10% to 100%, and also from 1% to 10%, by randomly picking up the specific quantity of objects from the complete ground truth. We only show the experimental results on two settings, namely *80-Pessimistic* and *Fully-Optimistic*, with the corresponding datasets depicted as *Synthetic80P* and *SyntheticFP*. The experimental results on all the other datasets show the same results. Note that all the objects in the synthetic datasets have only one true value, thus the resulting precision, recall, and F_1 score equal to each other. The accuracy and specificity show the same ranking results. Figure 7.1a and Figure 7.1c show the precision



(a) Synthetic80P Dataset

(b) Synthetic80P Dataset



(c) SyntheticFP Dataset

(d) SyntheticFP Dataset

Fig. 7.1 Precision/Recall of Twelve Truth Discovery Methods Evaluated on Different Coverages of the Leveraged Ground Truth

and recall of all the twelve methods with the coverage of the leveraged ground truth tuned from 10% to 100%, while Figure 7.1b and Figure 7.1d show those of the methods with the coverage tuned from 1% to 10%. The latter range forms the sparse ground truth, which is closer to the reality, where the coverage of the collected ground truth is always below 10%, sometimes even below 1%.

Ideally, if the performance evaluation is not biased by the incomplete ground truth, there should be no intersecting lines in the figures, demonstrating that the ranking of the metrics of the methods is consistent with the results measured on complete ground truth. Even if two or

more methods show the same performance, the precision/recall lines of those methods in the figures should totally overlap rather than intersect.

However, for both types of datasets, we cannot get the completely correct ranking for each type of datasets until the coverage of the leveraged ground truth grows up to 60%, which is generally impossible to obtain in reality. The results are even worse for the sparse ground truth. As shown in Figure 7.1b and Figure 7.1d, by tuning the coverage of the ground truth, the ranking of methods fluctuates all the time, and no correct result is returned. That means the performance evaluation is strongly biased by the sparse ground truth. In most cases, real-world datasets would not have strict mathematical distributions, such as source coverage distributions, ground truth distribution per source, and distinct value distribution per object might be random. Therefore, the ranking based on real-world datasets with sparse ground truth would be even less correct.

7.4 Our Approach

The most straightforward approach for truth discovery is to conduct majority voting for categorical data or averaging for continuous data. The largest limitation of such approach is that it assumes all the sources are equally reliable, which does not hold in most real-world scenarios. Thus, the most important feature of the existing truth discovery methods is their ability to estimate source trustworthiness [49]. While identifying the truth, current methods also return $c_{\mathscr{V}}$, the confidence score of each value $\mathscr{V}$ (or the probability of $\mathscr{V}$ being true), and τ_s, the trustworthiness of each source s (or the probability of source s providing true information), as the intermediate variables. In particular, a higher $c_{\mathscr{V}}$ indicates that value $\mathscr{V}$ is more likely to be true, and a higher τ_s indicates that source s is more reliable and the values claimed by this source are more likely to be true. Though the calculations of $c_{\mathscr{V}}$ and τ_s differ from one method to the other, current methods generally apply the same principle for truth discovery: if a source claims true values frequently, it will receive a high trustworthiness;

meanwhile, if a value is claimed by sources with high trustworthiness, it will be more likely to be identified as truth. To determine the truth, weighted aggregation of the multi-source data is performed based on the estimated source trustworthiness. Thus, value confidence score and source trustworthiness calculation are the key elements for truth discovery and can be leveraged to compare the performance of current truth discovery methods. In this section, we first review twelve existing truth discovery methods to be compared. Then we present our approach, *CompTruthHyp*, which compares those methods without using ground truth in both single-valued scenarios and multi-valued scenarios.

7.4.1 Twelve Truth Discovery Methods

In this section, we describe each algorithm briefly.

Majority voting. By regarding all the sources as equally reliable, *voting* does not estimate the trustworthiness of each source. Instead, it calculates $c_{\mathscr{V}}$ as $\frac{|S_{\mathscr{V}}|}{S_o}$, where $S_{\mathscr{V}}$ is the set of sources which provide $\mathscr{V}$ on object o, and S_o is the set of sources which provide values on o. For each object, the value with the highest confidence score will be identified as the truth.

Accu [38]. Dong et al. propose the first Bayesian truth detection model that incorporates copying detection techniques. *Accu* estimates trustworthiness of each source as the average confidence score of its provided values, while calculates value confidence scores by leveraging source trustworthiness using Bayes Rule.

TruthFinder [44]. It applies Bayesian analysis to estimate source trustworthiness and identify the truth, and additionally takes value similarity into consideration. *Truthfinder* terminates when the results from two successive iterations are less than a given threshold. The value with the highest confidence score is selected as the true value.

Sums, Average-Log, Investment, PooledInvestment [43, 114]. *Sums* employs *authority-hub analysis* [91] and computes source trustworthiness as the average confidence score

of its provided values. It has the disadvantage of overestimating the sources that make larger coverage of objects. *Average-Log*, *Investment*, and *PooledInvestment* apply different methods to assess source trustworthiness. Specifically, *Average-Log* uses a non-linear function to assess sources. Investment assumes each source uniformly invests/distributes its trustworthiness to the values it provides, while sources collect credits back from the confidence of their claimed values. *PooledInvestment* follows a similar procedure with *Investment*, except it uses a linear function to estimate the confidence of values instead of a non-linear function.

2-Estimates, 3-Estimates [36]. *2-Estimates* incorporates the mutual exclusion (i.e., while claiming a value of an object, a source is voting against all the other potential values of this object). *3-Estimates* augments *2-Estimates* by additionally taking the hardness of fact into consideration.

SimpleLCA, GuessLCA [37]. LCA (Latent Credibility Analysis) models source trustworthiness by a set of latent parameters. It enriches the meaning of source trustworthiness by tackling the difference between telling the truth and knowing the truth. In the *SimpleLCA* model, source trustworthiness is considered as the probability of a source asserting the truth, while in the *GuessLCA* model, source trustworthiness is regarded as the probability of a source both knowing and asserting the truth.

CRH [40]. It is a framework that tackles heterogeneity of data. Source trustworthiness calculation is jointly conducted across all the data types together. Different types of distance functions can be incorporated into the framework to capture the features of different data types.

7.4.2 CompTruthHyp

To compare the performance of the above methods without using the ground truth, our data model includes the following inputs: i) the input dataset for truth discovery; ii) the identified truth of each method ($m \in M$, $|M| = 12$); iii) source trustworthiness and value confidence scores output by each method. The output of our data model is a ranking of the accuracy of the twelve methods. As we do not have any ground truth, we propose to obtain the ranking by comparing the methods' ability to infer the observation of the given dataset from their outputs. We denote by ϕ the observation of which source votes for which value in the dataset, V^m the output of a method m, and $P(\phi|V^m)$ the probability of ϕ conditioned on V^m. A higher $P(\phi|V^m)$ indicates that the method m has bigger ability to capture the features of the given dataset, thus its output is more reliable.

Our computation requires several parameters, which can be derived from the inputs: $\tau_s(m)$, the probability that the claimed values of s is true, given V^m. $P_s(v_t|V_o^m)$ (resp., $P_s(v_f|V_o^m)$), the probability that a source provides a particular true (resp., false) value for object o, given V^m. As analyzed in Section 7.3.1, we compute the required parameters by applying different algorithms for single-valued and multi-valued scenarios.

Given V^m, if $v \in V_o^m$, v is a true value; if $v \in V_o - V_o^m$, v is a false value. Formally, if a source s covers an object o, we have the probability of the observation of s providing a particular value v ($v \in V_o$), conditioned on V^m, as:

$$P(\phi_{s_v}|V^m) = \begin{cases} \tau_s(m)P_s(v_t|V_o^m); & if\ v \in V_o^m \\ (1-\tau_s(m))P_s(v_f|V_o^m); & if\ v \in V_o - V_o^m \end{cases} \tag{7.1}$$

In our observation, we are interested in two sets of values: given V^m, ${V_s}^t(m)$, denoting the set of true values provided by s; ${V_s}^f(m)$, denoting the set of false values provided by s. ${V_s}^t(m) \cup {V_s}^f(m) = V_s$, V_s is the set of all values provided by s. Since we assume each source

provides each value independently, we have the probability of the observation of source s with its claimed values, i.e., ϕ_s, conditioned on V^m, as:

$$P(\phi_s|V^m) = (\prod_{v \in V_s{}^t(m)} \tau_s(m) P_s(v_t|V_o^m) \prod_{v \in V_s{}^f(m)} (1-\tau_s(m)) P_s(v_f|V_o^m)) \tag{7.2}$$

By assuming sources are independent on each other, the conditional probability of observing the given dataset ϕ is:

$$P(\phi|V^m) = \prod_{s \in S} (\prod_{v \in V_s{}^t(m)} \tau_s(m) P_s(v_t|V_o^m) \prod_{v \in V_s{}^f(m)} (1-\tau_s(m)) P_s(v_f|V_o^m)) \tag{7.3}$$

To simplify the computation, we define the *confidence* of method m, denote by $\mathscr{C}_m$, as

$$\mathscr{C}_m = \sum_{s \in S} (\sum_{v \in V_s{}^t(m)} \ln \tau_s(m) P_s(v_t|V_o^m) + \sum_{v \in V_s{}^f(m)} \ln(1-\tau_s(m)) P_s(v_f|V_o^m)) \tag{7.4}$$

Source Trustworthiness Normalization

The accuracy of truth discovery methods significantly depends on their source trustworthiness estimation. Although all methods calculate source trustworthiness as the weighted aggregation of value confidence scores, they adopt different models and equations. Therefore, the calculated source trustworthiness of each method has different meaning and is incomparable. To normalize source trustworthiness output by twelve methods, our approach, *CompTruthHyp*, regards the trustworthiness of a source as the probability of its claimed values being true (i.e., precision). We can derive a confusion matrix similar to Table 7.2 for each source based on the identified truth of each method. Then, we calculate the precision of each source output by each method ($\tau_s(m)$) as follows:

$$\tau_s(m) = \frac{TP_s{}^m}{TP_s{}^m + FP_s{}^m} \tag{7.5}$$

where $TP_s{}^m$ (resp., $FP_s{}^m$) is the number of true positives (resp., false positives) of the values claimed by source s, given V^m.

In the single-valued scenario, each source provides one value for any object of interest. Given V^m, all the values in $V_o - V_o{}^m$ are regarded as false ($|V_o - V_o{}^m| = |V_o| - 1$). We calculate $\tau_s(m)$ for each source by performing Algorithm 4. In particular, for each method $m \in M$ (Line 1), for each $s \in S$ (Line 2), for each $o \in O_S$ (Line 4, where O_s is the objects covered by s), if V_{s_o} is true (Line 5), $TP_s{}^m$ increases by one (Line 6), otherwise, $FP_s{}^m$ increases by one (Line 7, 8). For each source s, $\tau_s(m)$ is calculated by applying Equation 7.5 (Line 9).

Algorithm 4: The algorithm of source trustworthiness normalization for the single-valued scenario

Input: Given dataset $\{s, o, V_{s_o}\}$ and V^m for each $m \in M$.
Output: $\tau_s(m)$ for each $s \in S$, $m \in M$.

```
1  foreach m ∈ M do
2      foreach s ∈ S do
3          TP_s^m = 0; FP_s^m = 0;
4          foreach o ∈ O_s do
5              if V_{s_o} = V_o^m then
6                  TP_s^m++;
7              else
8                  FP_s^m++;
9          Calculate τ_s(m) by applying Equation 7.5;
10 return τ_s(m) for each s ∈ S, m ∈ M.
```

In the multi-valued scenario, each source may provide more than one value for a multi-valued object. Instead of regarding each value set provided by a source on the same object as a joint single value, we treat each value in the value set individually. Therefore, $|V_o{}^m|$ and $|V_{s_o}|$ may be bigger than 1. We calculate $\tau_s(m)$ for each source by performing Algorithm 5.

Value True-False Distributions

We analyze the true-false distribution of values for each object in a given dataset. Because each object has one single value in the single-valued scenario, we have $P_s(v_t|V_o^m)$ fixed to 1.

Algorithm 5: The algorithm of source trustworthiness normalization for the multi-valued scenario

Input: Given dataset $\{s, o, V_{s_o}\}$ and V^m for each $m \in M$.
Output: $\tau_s(m)$ for each $s \in S$, $m \in M$.

```
foreach m ∈ M do
    foreach s ∈ S do
        TP_s^m = 0; FP_s^m = 0;
        foreach o ∈ O do
            foreach v ∈ V_{s_o} do
                if v ∈ V_o^m then
                    TP_s^m++;
                else
                    FP_s^m++;
        Calculate τ_s(m) by applying Equation 7.5;
return τ_s(m) for each s ∈ S, m ∈ M.
```

As false values for an object can be random, $P_s(v_f|V_o^m)$ is different for false values, and the false value distribution of the object is also different. Given a set of false values of o ($V_o - V_o^m$), we need to analyze their distribution and calculate the probability ($P_s(v_f|V_o^m)$) for sources to pick a particular value from the distribution. In particular, we calculate this probability for each false value of each object using Algorithm 6. We define the *untrustworthiness* of a source as the probability that its claimed values are false, i.e., $(1-\tau_s(m))$. For each false value of an object, we calculate $P_s(v_f|V_o^m)$ by:

$$P_s(v_f|V_o^m) = \frac{\sum_{s \in S_{v_f}} (1-\tau_s(m))}{\sum_{v_f' \in V_o - V_o^m} \sum_{s' \in S_{v_f'}} (1-\tau_{s'}(m))} \tag{7.6}$$

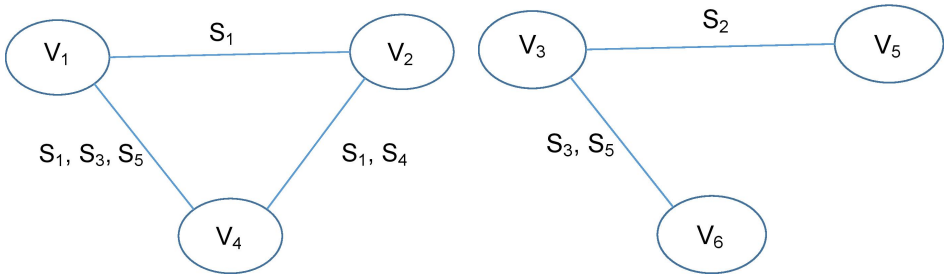
where S_{v_f} is the set of sources provide v_f on o.

In the multi-valued scenario, values in a source's claimed value set are not totally independent. Intuitively, the values occurring in the same claimed value set are believed to impact each other. The co-occurrence of values in the same claimed value set indicates that those values have potentially similar probabilities of being selected.

Algorithm 6: The algorithm of $P_s(v_f|V_o^m)$ calculation for the single-valued scenario

Input: Given dataset $\{s, o, V_{s_o}\}$ and V^m for each $m \in M$.
Output: $P_s(v_f|V_o^m)$ for each $v_f \in V_o - V_o^m$, $o \in O$, $m \in M$.

```
foreach m ∈ M do
    foreach o ∈ O do
        foreach v_f ∈ V_o − V_o^m do
            foreach s ∈ S_{v_f} do
                P_s(v_f|V_o^m) += (1 − τ_s(m));
            P_s(v_f|V_o^m) of each v_f is normalized to satisfy Σ_{v_f ∈ V_o − V_o^m} P_s(v_f|V_o^m) = 1;
return P_s(v_f|V_o^m) for each v_f ∈ V_o − V_o^m, o ∈ O, m ∈ M.
```



(a) True value graph (b) False value graph

Fig. 7.2 An Example of Value Co-Occurrences for a Multi-Valued Object

We define the weighted association among the distinctive values on the same object to represent their influence on each other, based on which to compute the probability of each value being selected. In particular, given V_o^m, we represent the bipartite mapping between true (resp., false) values on each multi-valued object and sources that claim the true (resp., false) values into a true (resp., false) value graph. In each true (reps., false) value graph, the identified true values (resp., false values) in V_o^m (resp., $V_o - V_o^m$) are the vertices, and sources that claim those values are the weights of edges which connect with the values. For example, the value co-occurrences for a multi-valued object are shown in Figure 7.2. $V_o = \{v_1, v_2, v_3, v_4, v_5, v_6\}$, $V_o^m = \{v_1, v_2, v_4\}$. True values v_2 and v_4 are claimed by both s_1 and s_4, while false values v_3 and v_5 are claimed by s_2.

The detailed procedure of $P_s(v_t|V_o^m)$ and $P_s(v_f|V_o^m)$ calculation is shown in Algorithm 7. For each true (resp., false) value graph, we further generate a corresponding square adjacent "true" (resp., "false") matrix, which should be irreducible, aperiodic, and stochastic to be guaranteed to converge to a stationary state. In particular, we first initialize each element in the matrix as the sum of the trustworthiness (resp., untrustworthiness) of all sources that claim the co-occurrence of the corresponding pair of true (resp., false) values (Line 8 and Line 17). To guarantee the three features of the matrix, we add a "*smoothing link*" by assigning a small weight to every pair of values (Line 9 and Line 18), where β is the smoothing factor. For our experiments, we set $\beta = 0.1$ (empirical studies such as the work done by Gleich et al. [173] demonstrate more accurate estimation). We then normalize the elements to ensure that every column in the matrix sums to 1 (Line 10 and Line 19). This normalization allows us to interpret the elements as the transition probabilities for the random walk computation. Finally, we adopt the *Fixed Point Computation Model* (FPC) [89] on each "true" (resp., "false") matrix to calculate $P_s(v_t|V_o^m)$ (reps., $P_s(v_f|V_o^m)$) for each true (resp., false) value of each object $o \in O$ (Line 11 and Line 20).

7.5 Experimental Evaluation

In this section, we report the experiments to evaluate our approach and discuss the results. In particular, for the single-valued scenario, we made comparative studies between our approach and the ground truth-based evaluation approach on eight groups of synthetic datasets with the complete ground truth. For the multi-valued scenario, we compared the approaches on two real-world datasets with the collected ground truth.

Algorithm 7: The algorithm of $P_s(v_t|V_o^m)$ and $P_s(v_f|V_o^m)$ calculation for the multi-valued scenario

Input: Given dataset $\{s, o, V_{s_o}\}$ and V^m for each $m \in M$.
Output: $P_s(v_t|V_o^m)$ for each $v_t \in V_o^m$, $P_s(v_f|V_o^m)$ for each $v_f \in V_o - V_o^m$, $o \in O$, $m \in M$.

$\beta = 0.1$;
foreach $m \in M$ **do**
    // "true" matrix generation
    foreach $o \in O$ **do**
        foreach $v_{t_1} \in V_o^m$ **do**
            foreach $v_{t_2} \in V_o^m$ **do**
                if $v_{t_1} \neq v_{t_2}$ **then**
                    foreach $s \in S_{v_{t_1}} \cap S_{v_{t_2}}$ **do**
                        $TrueMatrix[v_{t_1}][v_{t_2}] += \tau_s(m)$;
                    $TrueMatrix[v_{t_1}][v_{t_2}] = \beta + (1-\beta) * TrueMatrix[v_{t_1}][v_{t_2}]$;
        Normalize TrueMatrix;
        Apply FPC random walk computation to obtain $P_s(v_t|V_o^m)$ for each $v_t \in V_o^m$;
    // "false" matrix generation
    foreach $o \in O$ **do**
        foreach $v_{f_1} \in V_o - V_o^m$ **do**
            foreach $v_{f_2} \in V_o - V_o^m$ **do**
                if $v_{f_1} \neq v_{f_2}$ **then**
                    foreach $s \in S_{v_{f_1}} \cap S_{v_{f_2}}$ **do**
                        $FalseMatrix[v_{f_1}][v_{f_2}] += 1 - \tau_s(m)$;
                    $FalseMatrix[v_{f_1}][v_{f_2}] = \beta + (1-\beta) * FalseMatrix[v_{f_1}][v_{f_2}]$;
        Normalize FalseMatrix;
        Apply FPC random walk computation to obtain $P_s(v_f|V_o^m)$ for each $v_f \in V_o - V_o^m$;
return $P_s(v_t|V_o^m)$ for each $v_t \in V_o^m$, $P_s(v_f|V_o^m)$ for each $v_f \in V_o - V_o^m$, $o \in O$, $m \in M$.

7.5.1 Experimental Setup

Evaluation Metrics

We implemented all the twelve selected truth discovery methods, ground truth-based evaluation approach, and CompTruthHyp, in Python 3.4.0. All experiments were conducted on a 64-bit Windows 10 Pro. PC with an Intel Core i7-5600 processor and 16GB RAM. We

ran each truth discovery method 10 times and used the above introduced five traditional evaluation metrics, including *precision, recall, accuracy,* F_1 *score*, and *specificity*, as well as *confidence* output by CompTruthHyp, to evaluate their average performance. For the single-valued scenario, as the experimental results show that the rankings of different metrics are all equivalent, we discuss the precision of each method as an example. For the multi-valued scenario, we additionally introduce a new metric, namely *average*, to measure the overall performance of the methods, which is calculated as the average of the precision, recall, accuracy and specificity of each method.

To validate our approach, CompTruthHyp, we need to show the ranking of *confidence* of twelve selected methods, is closer than the rankings of various evaluation metrics of the methods derived from sparse/low-quality ground truth, to the real ranking of the performance of the methods derived from the complete ground truth. In this work, we adopt *Cosine similarity* (denoted as *Cos.*) and *Euclidean distance* (denoted as *Dist.*) to measure the distance of the two rankings. For Cosine similarity, a bigger value means better performance, while for Euclidean distance, a smaller value indicates better performance.

Synthetic Datasets

For the single-valued scenario, we applied the dataset generator introduced in Section 7.3.2, which can be configured to simulate a wide spectrum of truth discovery scenarios (except the multi-valued scenario). In particular, three parameters determine the scale of the generated dataset, including the number of sources ($|S|$), the number of objects ($|O|$), and the number of distinct values per object ($|V_o|$). The other three parameters determine the characteristics of the generated dataset, including source coverage (*cov*), ground truth distribution per source (GT), and distinct value distribution per object ($conf$). We fixed the scale parameters by setting $|S| = 50, |O| = 1,000$, and $|V_o| = 20$. To better simulate the real-world scenarios, we configured both *cov* and $conf$ as exponential distributions. By tuning GT as all possible

settings, including *uniform, Random, Full-Pessimistic, Full-Optimistic, 80-Pessimistic, 80-Optimistic*, and *Exponential*, we obtained eight groups of synthetic datasets (each group contains 10 datasets): i) *U25* (Uniform 25), each source provides the same number (25%) of true positive claims; ii) *U75* (Uniform 25), each source provides the same number (75%) of true positive claims; iii) *80P* (80-Pessimistic), 80% of the sources provide 20% true positive claims; 20% of the sources provide 80% true positive claims. iv) *80O* (80-Optimistic), 80% of the sources provide 80% true positive claims. 20% of the sources provide 20% true positive claims; v) *FP* (Full-Pessimistic), 80% of the sources provide always false claims and 20% of the sources provide always true positive claims; vi) *FO* (Full-Optimistic), 80% of the sources provide always true positive claims and 20% of the sources provide always false claims. vii) *R* (Random), the number of true positive claims per source is random; viii) *Exp* (Exponential), the number of true positive values provided by the sources is exponentially distributed. All synthetic datasets were generated with the complete ground truth.

Real-World Datasets

Since most existing datasets for categorical truth discovery [47, 48] are inapplicable for multi-truth scenarios, we refined two real-world datasets for our experiments, where each object may contain multiple true values.

The **Book-Author dataset** [44] contains $33,971$ book-author records crawled from *www.abebooks.com*. These records are collected from numerous book websites (i.e., sources). Each record represents a store's positive claims on the author list of a book (i.e., objects). We refined the dataset by removing the invalid and duplicated records, and excluding the records with only minor conflicts to make the problem more challenging—otherwise, even a straightforward method could yield competitive results. We finally obtained $13,659$ distinctive claims, 624 websites providing values about author name(s) of 677 books, each book has on average 3 authors. The ground truth provided by the original dataset was utilized,

which covers only 7.91% of the objects. The manually collected ground truth is sparse yet with high quality.

The **Parent-Children dataset** was prepared by extracting the parent-children relations from the *Biography dataset* [43]. We obtained 227,583 claims about the names of the children of 2,579 people (i.e., objects) edited by 54,764 users (i.e., sources). In the resulting dataset, each person has on average 2.48 children. We used the latest editing records as the ground truth, which covers all the objects. However, the quality of ground truth collected in this simple way is obviously very poor.

7.5.2 Experiments on Synthetic Datasets

In this set of experiments, we aim to compare the confidence ($\mathscr{C}_m$) and the precision of twelve methods calculated on different coverages of leveraged ground truth, denoted as P(1%) to P(100%), with their real precision calculated on the complete ground truth, denoted as P(100%), on eight groups of synthetic datasets with different settings of ground truth distributions. Table 7.3 shows the experimental results. As the results on U25 and U75 show similar features with 80P, we omit to show them in this chapter.

Table 7.3 Experimental Results for Six Types of Representative Synthetic Datasets (the Single-valued Scenario)

Dataset	Method	P(1%)	P(2%)	P(3%)	P(3%)	P(5%)	P(6%)	P(7%)	P(8%)	P(9%)	P(10%)	P(20%)	P(30%)	P(40%)	P(50%)	P(60%)	P(70%)	P(80%)	P(90%)	P(100%)	$\mathscr{C}_m$
80P	Voting	0.600	0.650	0.400	0.550	0.560	0.567	0.586	0.588	0.589	0.600	0.605	0.583	0.610	0.590	0.565	0.577	0.595	0.587	0.587	-16604
	Sums	0.700	0.700	0.533	0.650	0.620	0.633	0.657	0.638	0.622	0.660	0.650	0.620	0.670	0.658	0.610	0.633	0.655	0.643	0.639	-16514
	Avg	0.600	0.650	0.433	0.525	0.640	0.583	0.571	0.575	0.589	0.590	0.590	0.613	0.600	0.592	0.580	0.581	0.611	0.593	0.595	-16603
	Inv	0.500	0.300	0.367	0.325	0.380	0.467	0.486	0.438	0.433	0.430	0.480	0.417	0.460	0.458	0.420	0.427	0.453	0.444	0.448	-17319
	PInv	0.400	0.100	0.300	0.250	0.360	0.433	0.457	0.350	0.389	0.370	0.410	0.360	0.365	0.398	0.360	0.346	0.375	0.366	0.373	-17843
	Tru	0.700	0.700	0.533	0.650	0.660	0.650	0.671	0.650	0.689	0.680	0.655	0.650	0.693	0.684	0.633	0.657	0.683	0.663	0.662	-16489
	Est2	0.700	0.700	0.533	0.650	0.620	0.633	0.657	0.638	0.611	0.660	0.650	0.620	0.668	0.658	0.608	0.631	0.654	0.642	0.638	-16514
	Est3	0.200	0.05	0.133	0.050	0.120	0.283	0.214	0.175	0.167	0.190	0.215	0.203	0.163	0.196	0.183	0.177	0.181	0.181	0.184	-18629
	Accu	0.600	0.600	0.367	0.475	0.600	0.567	0.529	0.563	0.511	0.580	0.555	0.540	0.563	0.546	0.532	0.537	0.559	0.550	0.551	-16640
	CRH	0.700	0.650	0.467	0.575	0.600	0.617	0.629	0.613	0.567	0.630	0.610	0.590	0.648	0.620	0.585	0.600	0.628	0.614	0.612	-16558
	SLCA	**0.900**	**0.900**	**0.767**	**0.875**	**0.820**	**0.817**	**0.843**	**0.838**	**0.911**	**0.900**	**0.905**	**0.863**	**0.875**	**0.870**	**0.868**	**0.873**	**0.883**	**0.872**	**0.871**	**-15933**
	GLCA	0.700	0.700	0.533	0.675	0.660	0.650	0.671	0.638	0.689	0.670	0.660	0.657	0.685	0.678	0.633	0.653	0.680	0.661	0.661	-16480
	Dist.	5.916	4.359	3.873	3.000	4.123	1.732	2.000	2.646	3.162	1.732	2.236	2.236	1.414	1.000	1.000	0.000	0.000	0.000	0.000	1.414
	Cos.	0.975	0.987	0.989	0.993	0.987	0.998	0.997	0.995	0.992	0.998	0.996	0.996	0.998	0.999	0.999	1.000	1.000	1.000	1.000	0.998
80O	Voting	**1.000**	**0.950**	**1.000**	**1.000**	**1.000**	0.983	**1.000**	**1.000**	0.989	0.990	0.985	0.990	0.983	0.992	0.992	0.989	0.991	0.991	0.992	-10574
	Sums	**1.000**	**0.950**	**1.000**	**1.000**	**1.000**	0.983	**1.000**	**1.000**	**1.000**	**1.000**	0.985	0.990	0.985	0.992	**0.993**	0.991	0.994	0.993	0.994	-10567
	Avg	**1.000**	**0.950**	**1.000**	**1.000**	**1.000**	0.983	**1.000**	**1.000**	0.989	0.990	0.985	0.990	0.983	0.992	0.992	0.989	0.991	0.991	0.992	-10574
	Inv	0.900	0.800	**1.000**	0.875	0.800	0.817	0.871	0.900	0.889	0.900	0.815	0.833	0.813	0.834	0.837	0.834	0.859	0.844	0.847	-11944
	PInv	**1.000**	0.850	**1.000**	0.925	0.940	0.850	0.914	0.950	0.944	0.960	0.880	0.897	0.868	0.888	0.892	0.890	0.905	0.893	0.898	-11537
	Tru	**1.000**	**0.950**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.989	0.990	0.990	**0.997**	0.990	0.994	0.992	0.993	0.994	0.994	0.995	-10554
	Est2	**1.000**	**0.950**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.990	**0.997**	0.990	0.994	**0.993**	**0.994**	**0.995**	**0.996**	**0.996**	-10554
	Est3	0.200	0.150	0.200	0.125	0.200	0.150	0.257	0.138	0.133	0.250	0.155	0.210	0.195	0.192	0.212	0.184	0.206	0.193	0.196	-12107
	Accu	**1.000**	**0.950**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.990	**0.997**	0.990	0.994	**0.993**	**0.994**	**0.995**	**0.996**	**0.996**	-10554
	CRH	**1.000**	**0.950**	**1.000**	**1.000**	**1.000**	0.983	**1.000**	**1.000**	0.989	**1.000**	0.985	0.993	0.983	0.986	0.990	0.989	0.991	0.991	0.992	-10577
	SLCA	**1.000**	**0.950**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.989	0.990	**0.995**	**0.997**	**0.993**	**0.996**	**0.993**	**0.994**	**0.995**	**0.996**	**0.996**	**-10552**
	GLCA	**1.000**	**0.950**	**1.000**	**1.000**	**1.000**	0.967	**1.000**	**1.000**	0.978	0.990	0.980	0.990	0.980	0.990	0.990	0.987	0.990	0.990	0.991	-10578
	Dist.	15.427	12.530	12.530	12.530	12.530	3.610	12.530	12.530	5.477	8.426	3.162	4.796	2.646	4.359	4.123	1.000	0.000	1.000	0.000	3.317
	Cos.	0.778	0.859	0.859	0.859	0.859	0.989	0.859	0.859	0.975	0.938	0.992	0.981	0.994	0.984	0.985	0.999	1.000	0.999	1.000	0.991
	Voting	0.300	0.500	0.333	0.350	0.360	0.283	0.314	0.213	0.267	0.320	0.305	0.290	0.305	0.322	0.322	0.307	0.299	0.301	0.306	-19758
	Sums	0.300	0.500	0.300	0.325	0.300	0.233	0.257	0.175	0.233	0.260	0.250	0.240	0.250	0.256	0.267	0.246	0.249	0.250	0.251	-19647
	Avg	0.300	0.500	0.333	0.325	0.300	0.200	0.257	0.175	0.233	0.260	0.255	0.243	0.253	0.256	0.267	0.249	0.250	0.251	0.253	-19660
	Inv	0.300	0.350	0.267	0.250	0.200	0.167	0.257	0.150	0.200	0.230	0.210	0.223	0.235	0.230	0.240	0.217	0.219	0.217	0.222	-19841
	PInv	0.300	0.350	0.333	0.250	0.160	0.200	0.243	0.175	0.189	0.260	0.215	0.243	0.230	0.234	0.243	0.227	0.226	0.223	0.228	-19931
	Tru	0.300	0.500	0.300	0.325	0.300	0.233	0.257	0.175	0.233	0.260	0.250	0.240	0.250	0.256	0.267	0.246	0.249	0.250	0.251	-19639

FP	Est2	0.300	0.500	0.300	0.325	0.300	0.200	0.257	0.175	0.233	0.260	0.250	0.240	0.245	0.252	0.262	0.243	0.246	0.248	0.248	-19634
	Est3	0.100	0.500	0.367	0.350	0.340	0.300	0.300	0.275	0.244	0.330	0.290	0.327	0.320	0.284	0.312	0.306	0.299	0.302	0.306	-19526
	Accu	0.200	0.500	0.300	0.275	0.240	0.167	0.200	0.150	0.211	0.230	0.225	0.227	0.230	0.232	0.247	0.231	0.230	0.231	0.232	-19576
	CRH	0.400	0.550	0.333	0.400	0.380	0.317	0.300	0.213	0.267	0.360	0.305	0.303	0.315	0.328	0.328	0.317	0.314	0.318	0.318	-19703
	SLCA	**1.000**	**1.000**	**1.000**	**0.975**	**0.980**	**1.000**	**0.986**	**0.925**	**0.944**	**0.970**	**0.955**	**0.963**	**0.968**	**0.964**	**0.967**	**0.963**	**0.956**	**0.960**	**0.962**	**-14860**
	GLCA	0.400	0.600	0.467	0.500	0.380	0.450	0.414	0.350	0.389	0.430	0.435	0.410	0.425	0.436	0.432	0.417	0.414	0.413	0.417	-19686
	Dist.	15.033	9.165	7.874	3.464	3.873	4.36	7.280	6.325	3.873	6.245	2.449	5.657	2.828	2.236	1.732	1.000	0.000	1.000	0.000	13.928
	Cos.	0.799	0.948	0.952	0.993	0.989	0.985	0.960	0.974	0.989	0.974	0.995	0.974	0.994	0.996	0.998	0.999	1.000	0.999	1.000	0.848
FO	Voting	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**-1978**
	Sums	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**-1978**
	Avg	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**-1978**
	Inv	0.500	0.350	0.500	0.650	0.380	0.400	0.571	0.550	0.578	0.490	0.490	0.450	0.458	0.460	0.478	0.447	0.456	0.477	0.464	-10458
	PInv	0.500	0.350	0.400	0.575	0.340	0.317	0.529	0.525	0.500	0.420	0.420	0.407	0.410	0.412	0.425	0.397	0.409	0.422	0.412	-11755
	Tru	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**-1978**
	Est2	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**-1978**
	Est3	0.200	0.100	0.300	0.225	0.100	0.117	0.214	0.225	0.244	0.140	0.200	0.200	0.183	0.194	0.192	0.183	0.193	0.189	0.184	-11885
	Accu	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**-1978**
	CRH	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**-1978**
	SLCA	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**-1978**
	GLCA	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**-1978**
	Dist.	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Cos.	0.999	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
R	Voting	**0.400**	0.200	0.233	0.250	0.320	**0.350**	0.300	0.288	0.267	0.340	0.340	0.313	0.290	0.306	0.283	0.301	0.314	0.310	0.303	-19581
	Sums	**0.400**	0.150	0.267	0.250	0.320	0.333	0.286	0.288	0.267	0.360	0.335	0.317	0.270	0.306	0.293	0.299	0.310	0.307	0.303	-19549
	Avg	**0.400**	0.150	0.300	0.250	0.320	**0.350**	0.343	0.263	0.278	0.350	0.315	0.323	0.300	0.310	0.293	0.307	0.320	0.311	0.307	-19570
	Inv	**0.400**	0.200	0.300	0.275	0.300	0.333	**0.371**	0.263	0.289	0.370	0.325	0.337	0.303	0.292	0.300	0.313	0.321	0.319	0.312	-19800
	PInv	**0.400**	0.200	0.267	**0.325**	0.300	0.333	0.329	0.238	**0.311**	0.340	0.340	0.330	0.305	0.292	0.308	**0.320**	0.324	0.323	0.317	-19839
	Tru	**0.400**	0.150	0.233	0.225	0.360	**0.350**	0.271	**0.300**	0.267	0.350	0.315	0.323	0.293	0.310	0.293	0.301	0.315	0.311	0.306	-19522
	Est2	**0.400**	0.150	0.267	0.250	0.320	0.333	0.286	0.288	0.267	0.360	0.330	0.317	0.270	0.304	0.292	0.299	0.309	0.306	0.302	-19548
	Est3	**0.400**	0.200	0.333	0.175	0.300	0.250	0.214	0.275	0.300	**0.380**	0.330	**0.370**	**0.348**	0.320	**0.327**	0.313	**0.333**	**0.332**	**0.330**	**-19356**
	Accu	**0.400**	0.150	0.267	0.250	0.320	0.350	0.286	0.263	0.244	0.350	0.310	0.313	0.280	0.302	0.290	0.304	0.310	0.308	0.302	-19551
	CRH	**0.400**	**0.250**	0.233	0.200	0.320	**0.350**	0.243	0.275	0.267	0.350	**0.350**	0.297	0.273	0.294	0.293	0.303	0.308	0.306	0.299	-19496
	SLCA	**0.400**	0.200	**0.367**	0.250	**0.380**	0.300	0.271	0.275	0.244	0.370	**0.350**	0.330	0.315	**0.334**	0.300	0.304	0.313	0.317	0.313	-19376
	GLCA	**0.400**	0.200	0.167	0.250	0.360	**0.350**	0.286	0.288	0.267	0.350	0.325	0.323	0.295	0.314	0.297	0.304	0.309	0.309	0.305	-19531
	Dist.	21.726	14.457	14.318	17.635	19.875	23.601	16.279	19.950	12.884	14.177	16.217	4.243	4.359	13.153	7.937	8.718	6.000	3.464	0.000	16.733
	Cos.	0.887	0.814	0.818	0.713	0.625	0.488	0.774	0.648	0.854	0.821	0.781	0.985	0.985	0.858	0.947	0.936	0.971	0.990	1.000	0.778
	Voting	0.000	0.000	0.267	0.175	0.120	0.167	0.186	0.175	0.167	0.190	0.145	0.153	0.155	0.150	0.145	0.136	0.146	0.151	0.145	-19530

Exp	Sums	0.000	0.000	0.267	0.175	0.120	0.167	0.186	0.175	0.167	0.190	0.145	0.153	0.155	0.150	0.145	0.136	0.146	0.151	0.145	-19488
	Avg	0.000	0.000	0.267	0.175	0.120	0.167	0.186	0.175	0.167	0.190	0.145	0.153	0.155	0.150	0.145	0.136	0.146	0.151	0.145	-19508
	Inv	0.000	0.000	0.367	0.325	0.160	0.233	0.257	0.250	0.211	0.240	0.205	0.203	0.228	0.210	0.203	0.190	0.208	0.206	0.202	-20000
	PInv	0.000	0.000	0.400	0.350	0.160	0.250	0.286	0.275	0.256	0.250	0.235	0.217	0.260	0.238	0.223	0.210	0.236	0.231	0.229	-20185
	Tru	0.000	0.000	0.267	0.175	0.120	0.167	0.186	0.175	0.167	0.190	0.145	0.153	0.155	0.150	0.145	0.136	0.146	0.151	0.145	-19474
	Est2	0.000	0.000	0.267	0.175	0.120	0.167	0.186	0.175	0.167	0.190	0.145	0.153	0.155	0.15	0.145	0.136	0.146	0.151	0.145	-19488
	Est3	**0.420**	**0.420**	**0.420**	**0.420**	**0.420**	**0.420**	**0.420**	**0.420**	**0.420**	**0.420**	**0.420**	**0.420**	**0.420**	**0.420**	**0.420**	**0.420**	**0.420**	**0.420**	**0.420**	**-16596**
	Accu	0.000	0.000	0.267	0.175	0.120	0.167	0.186	0.175	0.167	0.190	0.145	0.153	0.155	0.150	0.145	0.136	0.146	0.151	0.145	-19493
	CRH	0.000	0.000	0.267	0.175	0.120	0.167	0.186	0.175	0.167	0.190	0.145	0.153	0.155	0.150	0.145	0.136	0.146	0.151	0.145	-19482
	SLCA	0.200	0.200	0.333	0.225	0.220	0.233	0.243	0.238	0.278	0.270	0.270	0.300	0.240	0.246	0.268	0.250	0.261	0.268	0.257	-19362
	GLCA	0.000	0.000	0.300	0.200	0.120	0.183	0.186	0.188	0.178	0.200	0.160	0.163	0.168	0.162	0.155	0.146	0.155	0.160	0.153	-19489
	Dist.	8.246	8.246	6.164	2.449	2.828	1.732	3.606	2.449	0.000	0.000	0.000	0.000	0.000	1.414	0.000	0.000	0.000	0.000	0.000	13.114
	Cos.	0.978	0.978	0.974	0.990	0.998	0.995	0.986	0.990	1.000	1.000	1.000	1.000	1.000	0.997	1.000	1.000	1.000	1.000	1.000	0.873

Table 7.4 Experimental Results for Two Real-World Datasets (the Multi-Valued Scenario)

Dataset	Method	Precision	Recall	Accuracy	Specificity	F_1	Average	$\mathscr{C}_m$
Book-Author	Voting	0.749	0.712	0.576	0.022	0.730	0.515	-26258
	Sums	0.851	0.685	0.651	**0.511**	0.759	0.674	-23011
	AvgLog	0.841	0.663	0.629	0.489	0.742	0.656	-23477
	Inv	0.815	0.745	0.659	0.311	0.778	0.633	-23860
	PInv	0.812	0.750	0.659	0.289	0.780	0.628	-23435
	Tru	0.847	0.663	0.633	**0.511**	0.744	0.664	-23303
	Est2	**0.863**	**0.755**	**0.707**	**0.511**	**0.806**	**0.709**	**-21915**
	Est3	0.828	0.734	0.664	0.378	0.778	0.651	-24907
	Accu	**0.858**	**0.788**	**0.725**	0.467	**0.822**	**0.709**	**-21390**
	CRH	0.850	0.679	0.646	**0.511**	0.755	0.672	-22751
	SLCA	**0.861**	**0.810**	**0.742**	0.467	**0.835**	**0.720**	**-21670**
	GLCA	0.846	0.658	0.629	**0.511**	0.740	0.661	-23243
	Dist.	5.099	13.153	11.225	13.153	10.863	4.472	0.000
	Cos	0.980	0.865	0.901	0.861	0.909	0.985	1.000
Parent-Children	Voting	0.919	0.901	0.845	0.462	0.910	0.782	-330234
	Sums	0.938	0.927	0.883	0.585	0.933	0.833	-314582
	AvgLog	0.938	0.926	0.882	0.581	0.932	0.832	-314124
	Inv	0.915	0.919	0.841	0.457	0.917	0.783	-331351
	PInv	0.912	0.912	0.839	0.454	0.912	0.779	-331523
	Tru	0.938	0.926	0.881	0.581	0.932	0.832	-315231
	Est2	**0.940**	**0.927**	**0.885**	**0.595**	**0.933**	**0.836**	**-309873**
	Est3	0.905	0.889	0.822	0.366	0.897	0.746	-340031
	Accu	**0.941**	**0.928**	**0.885**	**0.588**	**0.934**	**0.836**	**-310314**
	CRH	0.938	0.927	0.883	0.586	0.932	0.833	-313421
	SLCA	**0.942**	**0.927**	**0.886**	**0.601**	**0.935**	**0.839**	**-302873**
	GLCA	0.938	0.924	0.876	0.578	0.931	0.829	-321098
	Dist.	2.828	3.742	2.000	1.000	3.162	1.414	0.000
	Cos.	0.994	0.989	0.997	0.999	0.992	0.998	1.000

We observe that none of the twelve methods constantly outperforms the others in terms of precision, and a "one-fits-all" approach does not seem to be achievable. Based on the best performance values (shown in bold), we can see that the best method changed from dataset to dataset. In some cases, an improved method may not even beat its original version as a result of different features of the applied datasets. For example, while in most datasets 2-Estimates performed better than 3-Estimates, it performed worse than 3-Estimates in *FP* and *R*, where most of the claims provided by most sources could be false. This shows that in such cases, the factor that "hardness of facts" should be considered to achieve better truth discovery. This instability of truth discovery methods reveals the importance of evaluating the methods. With a better evaluation approach, users can choose the best method for truth discovery more easily and accurately for a given scenario.

From the table, we can see that CompTruthHyp can always identify the best method for the given dataset. For *80P*, *80O*, and *FO*, the majority of methods performed better than

random guessing with the real precision bigger than 0.5. For *FO*, the ranking of precision stayed stable with the coverage of the ground truth tuned from 1% to 100% and was consistent with the ranking of the real precision. The ranking of the confidence of methods output by CompTruthHyp was also equal to the ranking of their real precision, with $Dist. = 0$, and $Cos. = 1$. While CompTruthHyp and ground truth-based evaluation approach showed similar performance on this type of datasets in terms of accuracy, our approach did not cost any efforts for ground truth collection. For *80P* and *80O*, when the coverage of the ground truth increased, the Euclidean distance decreased until it reached 0 (70% for *80P*, 80% for *80O*), the Cosine similarity increased until it reached 1 (70% for *80P*, 80% for *80O*). The Euclidean distance and Cosine similarity of the confidence ranking were 1.414 and 0.998 for *80P*, which were as good as those of P(40%), while for *80O*, the ground truth-based evaluation approach beat our approach only when they got a ground truth with coverage bigger than 70%. Moreover, in real-world datasets, the collection of a ground truth with coverage bigger than 10% is a rather challenging task.

For *R*, *FP*, and *Exp*, none of the methods was reliable, except for SLCA on *FP*. Almost all the methods performed worse than random guessing with a real precision smaller than 0.5, and the real precision of those methods was similar with each other. For *R*, with the coverage of the ground truth increased, the Euclidean distance and Cosine similarity of the precision ranking fluctuated. Even when the coverage reached 90%, the Euclidean distance was 3.464, which is still not close enough to the real ranking. Though the Euclidean distance of the confidence ranking was 16.733 and the Cosine similarity was 0.778, which are not close to the real ranking, it performed better than the rankings of P(1%), P(4%), P(5%), P(6%), P(8%) in terms of Euclidean distance, and those of P(4%), P(5%), P(6%), P(7%), P(8%) in terms of Cosine similarity. In the case of *FP*, our approach can only identify the best method and performed better than the ground truth-based evaluation approach when the coverage of the ground truth was 1%. However, in this case, only the best method performed better than

random guessing and all the other methods showed very similar bad performance. For *Exp*, where one source always lies and one source always tells the truth for all the objects and the remaining sources range from 1% to 99% of values they claim is true. None of the methods was reliable and all of them performed similarly bad. Even in this case, our approach can still find out the best method, i.e., 3-Estimates.

7.5.3 Experiments on Real-World Datasets

In this section, we report similar comparative studies with two real-world datasets for multi-valued scenarios. As precision cannot reflect the overall performance of a method with the complete ground truth (as analyzed in Section 7.3.1), we compared the confidence ranking of the methods with the ranking of all six metrics calculated on the provided ground truth, including precision, recall, accuracy, specificity, F_1 score, and average. Table 7.4 shows the experimental results, with the top-three best performance in bold. These results also validate the observation that no method constantly outperforms the others. We also observed that the rankings of different metrics differed from one another, which validates our assertion that any one of those metrics can not individually reflect the overall performance of the methods. All methods performed worse on the Book-Author dataset than on the Parent-Children dataset with lower precision, recall, accuracy, and specificity. The possible reasons contain the poorer quality of sources (poorer ground truth distribution), more missing values (i.e., true values that are missed by all the sources), and the smaller scale of the dataset.

For both datasets, our approach can consistently identify the top-three best methods. The confidence ranking is more similar with the ranking of average than the ranking using other metrics. This validates that confidence metric reflects the overall performance of the methods. However, for the Book-Author dataset, the Euclidean distance of the confidence ranking to average was still bigger than 4.0 and the Cosine similarity with average was still lower than 0.99. This is because the ground truth is relatively sparse, so the ranking of average

cannot reflect the real performance ranking of the methods. Another reason is that there may be copying relations among sources, which are neglected by all the methods including our approach. Compared with the Book-Author dataset, the confidence ranking was closer to the rankings of all metrics on the Parent-Children dataset. This is because the ground truth covers all the objects and is obtained by collecting all the latest editions regarding the objects. Although the precision of the ground truth not reaches 1, the quality of sources in this dataset is relatively high. Therefore, the leveraged ground truth is similar to the complete ground truth.

7.6 Summary

In this chapter, we focus on the problem of comparing truth discovery methods without using the ground truth, which has not been studied by previous research efforts. We first motivate this study by revealing the bias introduced by sparse ground truth in evaluating the truth discovery methods, by conducting experiments on synthetic datasets with different coverages of the ground truth. Then, we propose a general approach, called *CompTruthHyp*, to solve this bias. In particular, we propose two approaches for single-valued and multi-valued scenarios, respectively. Given a dataset, we first calculate the precision of each source by the output of each truth discovery method. Based on the source precision and the identified truth, we estimate the probability of observations of the given data set, for each method. The performance of methods is determined by the ranking of the calculated probabilities. Experimental studies on both real-world and synthetic datasets demonstrate the effectiveness of our approach.

Chapter 8

Conclusion

In this chapter, we summarize the contributions of this dissertation and discuss some future research directions regarding knowledge base construction.

8.1 Summary

The last few years have seen a rapid increase of sheer amount of data produced and communicated over the Internet and the Web. As the scale of data increases unprecedentedly, it becomes more urgent than ever to exploit the full values of these data. To this end, many knowledge bases have been constructed for both human use and feeding knowledge-driven applications. Despite the large scale of existing KBs, they are still far from complete and accurate.

In this dissertation, we address several research issues regarding generating actionable knowledge from big data covering topics from knowledge extraction and truth discovery. We propose a novel system to construct a comprehensive knowledge base. We propose approaches to extract new predicates from the Web for ontology augmentation. We tackle the multi-truth discovery problem for estimating value veracity for multi-valued objects. We combine the existing truth discovery methods by utilizing two models for better truth discov-

ery. We also design a novel algorithm to evaluate the performance of truth discovery methods without using ground truth. In particular, we summarize our main research contributions in the following:

- **Generating actionable knowledge from big data.** We proposed a novel system, called *GrandBase*, to **g**ene**r**ate **a**ctionable k**n**owledge from big **d**ata [159]. The system contains two main phases, namely *knowledge extraction* and *truth discovery*. For knowledge extraction, we change the traditional tasks of entity linkage and predicate linkage to new entity extraction and new predicate extraction. For truth discovery, we eliminate the sinlge-truth aasumption, targeting the real-world scenarios where multi-valued objects widely exist.

- **Extracting new predicate from multiple types of sources.** We designed a novel framework that extracts and merges the predicates from four types of sources, existing KBs (i.e., Freebase and DBpedia), query stream, Web texts, and DOM trees, for comprehensive ontology augmentation [34]. In particular, we first combined predicate extractions from DBpedia with Freebase, adopted new patterns and filtering rules for better query stream extraction. Then, we utilized those seeds to learn tag path patterns (from DOM trees), and lexical and parse patterns (from Web texts). Those patterns were in turn leveraged to extract new predicates from DOM trees and Web texts.

- **Tackling multi-truth discovery via graph-based approaches.** With theoretical analysis, we found that the agreements among sources indicate endorsement, which motivates us to model the quality of each source by quantifying the agreements and endorsement relations among sources. we first proposed a novel approach, *SourceVote* [171], to estimate value veracity for multi-valued data items. SourceVote modeled the endorsement relations among sources by quantifying the inter-source agreements on positive and negative claims. In particular, two graphs were constructed to model inter-source

relations. Then two aspects of source reliability (positive precision and negative precision) were derived from these graphs and were used for estimating value veracity and initializing existing truth discovery methods. To improve the accuracy of SourceVote, we further proposed a full-fledged graph-based model, *SmartVote* [175, 172, 177], by incorporating four implications including *two types of source relations, object popularity, loose mutual exclusion*, and *long-tail phenomenon on source coverage*. In particular, ±supportive agreement graphs were constructed to model the endorsement among sources on their positive and negative claims, from which the improved evaluations of two-sided source reliability were derived. Copying relations among sources were captured by constructing the ±malicious agreement graphs based on the consideration that sources sharing the same false values are more likely to be dependent. SmartVote additionally took supportive relations and copying relations among sources into consideration. Popularities of objects were measured based on object occurrences and source coverage, to minimize the number of people misguided by false values. SmartVote applied source confidence scores to differentiate the extent to what a source believes its positive claims and negative claims. For the ubiquitous long-tail phenomenon on source coverage, SmartVote also added smoothing weights to the ±supportive agreement graphs to avoid the reliability of small sources from being over- or under-estimated.

- **Ensembling for better truth discovery.** We analyzed the feasibility of applying an ensemble approach to combining existing truth discovery methods for better truth discovery. Two novel models, namely *serial model* and *parallel model*, and several implementations based on both models for both single-truth and multi-truth discovery problems, were proposed [182].

- **Evaluating existing truth discovery methods without using ground truth.** We examined the bias introduced by sparse ground truth in evaluating the truth discovery

methods, by conducting experiments on synthetic datasets with different coverages of the ground truth. A generic approach, called *CompTruthHyp*, was proposed to **comp**are the performance of **truth** discovery methods without using ground truth, by considering the output of each method as a **hyp**othesis about the ground truth [185]. In particular, we proposed two algorithms for single-valued and multi-valued scenarios, respectively. Given a dataset, we first calculated the precision of each source by the output of each truth discovery method. Based on the source precision and the identified truth, we estimated the probability of observations of the given data set, for each method. The performance of methods was determined by the ranking of the calculated probabilities.

8.2 Future Directions

There are many opportunities to extend this work for full-fledged knowledge base construction. In this section, we lay out a research agenda by proposing several future research directions.

New entity extraction. Based on the discovered new attributes, we will consider to create new entities automatically by improving the existing techniques [160]. Specifically, we will seek to solve entity-linking and entity-discovery jointly. To improve the scalability of the solution, based on the more comprehensive attribute set, we will develop a novel model that reasons over the compact hierarchical entity representations, as well as a new distributed inference architecture, which is inherent in the MapReduce architectures, that avoids the synchronicity bottleneck. Finally, we will apply this enhanced ontology to explore more facts from the open Web environment.

Theoretical guarantee analysis. Inspired by a recent research effort by Xiao et al. [186], one of our future research efforts is to conduct the convergence analysis of SmartVote and study the theoretical guarantee of the results of SmartVote.

Performance evaluation without using groundtruth. Chapter 7 is our first step towards truth discovery methods comparison without using the ground truth. Our future work will focus on enhancing the approach by considering more complex application scenarios. For example, we are interested in the scenarios with complex source relationships such as copying and mutual supportive relations (i.e., two sources with similar facts) [49].

Quantifying extraction uncertainty. While many extractors have been proposed, rare research efforts have been devoted to investigating the uncertainty of extractions. Few knowledge extraction techniques simultaneously assign confidence scores to their extractions [15, 93], and consequently, these scores are rarely leveraged to improve the quality of extractions. Moreover, the criterion of confidence assignment in different extractor is varied from one another, making the confidence scores incomparable and tricky to be utilized. In our future work, we plan to assign a confidence score to each triple extracted by our extractors by following a unified criterion and incorporate those scores into our graph-based truth discovery model for better value veracity estimation.

Considering noises introduced by extractors. Existing truth discovery methods refer to the real-world sources, e.g., Websites, as the provenance of data. However, the datasets, on which the existing approaches conduct truth discovery, are extracted from the real-world sources by various extractors with different capabilities. The issue is, not only the real-world sources are error-prone, but also the extractors may introduce additional errors into the datasets, including predicates linkage errors, triple identification errors, and entity linkage errors. Ignoring the noises introduced by extractors would impair the accuracy of truth discovery. By additionally considering extractors as one of the provenances of data, a more challenging problem, *knowledge fusion*, should be considered. Dong et al. [20] recently

investigate data fusion techniques and find that some of them are still promising in solving the knowledge fusion problem. However, these methods are all under the single-valued assumption. We will further incorporate the multi-valued knowledge fusion approach into our system.

Detecting inter-source and inter-extractor relations. There are complex relations among real world sources, for example, one source may directly or transitively copies the other sources and several sources may copy data from one authoritative source. The relations among extractors can be even richer. There may be correlations among extractors, if they focus on the same types of Web content or apply the same extraction techniques. On the other hand, there may also be anti-correlations among extractors if they apply significantly different extraction techniques. Taking these relations into consideration may lead to better fusion results.

Considering hierarchical value spaces. Previous research efforts [47] have proposed to improve the accuracy of truth discovery by considering value similarity. However, they all focus on the similarity of string or numeric values. To the best of our knowledge, there is no existing work that considers value hierarchy. For example, for "the hometown of a person", "Wuhan" and "Hubei" can both be the true values (Wuhan is the capital city of Hubei province). In the future, we will propose a strategy that can infer the hierarchy and similarity of the values of predicates, where the information is presented by our extracted triples.

References

[1] Gerhard Weikum and Martin Theobald. From Information to Knowledge: Harvesting Entities and Relationships from Web Sources. In *In Proc. of the 29th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS'10, pages 65–76, Indianapolis, Indiana, USA, 2010.

[2] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A Core of Semantic Knowledge. In *Proc. of the 16th Intl. Conference on World Wide Web*, WWW'07, pages 697–706, Banff, Alberta, Canada, 2007.

[3] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka, Jr., and Tom M. Mitchell. Toward an Architecture for Never-ending Language Learning. In *Proc. of the 24th AAAI Conference on Artificial Intelligence*, AAAI'10, pages 1306–1313, Atlanta, Georgia, 2010.

[4] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A Nucleus for a Web of Open Data. In *Proc. of the 6th Intl. The Semantic Web and 2th Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, pages 722–735, Busan, Korea, 2007.

[5] Feng Niu, Ce Zhang, Christopher Ré, and Jude Shavlik. DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference. In *Proc. of the 2th Intl. Workshop on Searching and Integrating New Web Data Sources*, VLDS'12, pages 25–28, Istanbul, Turkey, 2012.

[6] Feng Niu, Ce Zhang, Christopher Ré, and Jude Shavlik. Elementary: Large-scale Knowledge-base Construction via Machine Learning and Statistical Inference. *Intl. Journal on Semantic Web and Information Systems (IJSWIS)*, 8(3):42–73, 2012.

[7] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Unsupervised Named-Entity Extraction from the Web: An Experimental Study. *Artificial Intelligence*, 165(1):91–134, 2005.

[8] Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying Relations for Open Information Extraction. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'11, pages 1535–1545, Edinburgh, United Kingdom, 2011.

[9] Jia Deng, Wei Dong, Richard Socher, Li-jia Li, Kai Li, and Fei-fei Li. ImageNet: A Large-scale Hierarchical Image Database. In *Proc. of the IEEE Computer Society*

Conference on Computer Vision and Pattern Recognition, CVPR'09, pages 248–255, Florida, USA, 2009.

[10] Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250, 2012.

[11] Robert Speer and Catherine Havasi. Representing General Relational Knowledge in ConceptNet 5. In *Proc. of the 8th Intl. Conference on Language Resources and Evaluation*, LREC'12, pages 3679–3686, Istanbul, Turkey, 2012.

[12] Vrandecic Vrandečić and Markus Krötzsch. Wikidata: A Free Collaborative Knowledge Base. *Communications ACM*, 57(10):78–85, 2014.

[13] Vivi Nastase and Michael Strube. Transforming Wikipedia into a Large scale Multilingual Concept Network. *Artificial Intelligence*, 194:62–85, 2013.

[14] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proc. of the 2008 ACM SIGMOD Intl. Conference on Management of Data*, SIGMOD'08, pages 1247–1250, Vancouver, Canada, 2008.

[15] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. In *Proc. of the 20th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, KDD'14, pages 601–610, New York, New York, USA, 2014.

[16] Rahul Gupta, Alon Halevy, Xuezhi Wang, Steven Whang, and Fei Wu. Biperpedia: An Ontology for Search Applications. *Proc. the VLDB Endowment*, 7(7):505–516, 2014.

[17] Marius Paşca. Acquisition of Open-domain Classes via Intersective Semantics. In *Proc. of the 23th Intl. Conference on World Wide Web*, WWW'14, pages 551–562, Seoul, South Korea, 2014.

[18] Zaiqing Nie, Ji-Rong Wen, and Wei-Ying Ma. Statistical Entity Extraction From the Web. *Proc. of the IEEE*, 100(9):2675–2687, 2012.

[19] Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q. Zhu. Probase: A Probabilistic Taxonomy for Text Understanding. In *Proc. of the 2012 ACM SIGMOD Intl. Conference on Management of Data*, SIGMOD'12, pages 481–492, Scottsdale, Arizona, USA, 2012.

[20] Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Kevin Murphy, Shaohua Sun, and Wei Zhang. From Data Fusion to Knowledge Fusion. *Proc. the VLDB Endowment*, 7(10):881–892, 2014.

[21] Xiu Susie Fang. Generating Actionable Knowledge from Big Data. In *Proc. of the 2015 ACM SIGMOD on PhD Symposium*, SIGMOD'15 PhD Symposium, pages 3–8, 2015.

[22] Bing Liu, Robert Grossman, and Yanhong Zhai. Mining Data Records in Web Pages. In *Proc.of the 9th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, KDD'03, pages 601–606, Washington, D.C., USA, 2003.

[23] Lidong Bing, Wai Lam, and Yuan Gu. Towards a Unified Solution: Data Record Region Detection and Segmentation. In *Proc. of the 20th ACM Intl. Conference on Information and Knowledge Management*, CIKM'11, pages 1265–1274, Glasgow, Scotland, UK, 2011.

[24] Arlind Kopliku, Mohand Boughanem, and Karen Pinel-Sauvagnat. Towards a Framework for Attribute Retrieval. In *Proc. of the 20th ACM Intl. Conference on Information and Knowledge Management*, CIKM'11, pages 515–524, Glasgow, Scotland, UK, 2011.

[25] Ralph Grishman. Information Extraction: Capabilities and Challenges. In *Notes for the 2012 International Winter School in Language and Speech Technologies*, Rovira i Virgili University, Tarragona, 2012.

[26] Mahnoosh Kholghi, Laurianne Sitbon, Guido Zuccon, and Anthony N. Nguyen. External Knowledge and Query Strategies in Active Learning: a Study in Clinical Information Extraction. In *Proc. of the 24th ACM Intl. Conference on Information and Knowledge Management*, CIKM'15, pages 143–152, Melbourne, VIC, Australia, 2015.

[27] Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam Mausam. Open Information Extraction: The Second Generation. In *Proc. of the 22th Intl. Joint Conference on Artificial Intelligence*, IJCAI'11, pages 3–10, Barcelona, Catalonia, Spain, 2011.

[28] Subhabrata Mukherjee, Gerhard Weikum, and Cristian Danescu-Niculescu-Mizil. People on Drugs: Credibility of User Statements in Health Communities. In *Proc. of the 20th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, KDD'14, pages 65–74, New York, New York, USA, 2014.

[29] Polina Rozenshtein, Aris Anagnostopoulos, Aristides Gionis, and Nikolaj Tatti. Event Detection in Activity Networks. In *Proc. of the 20th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, KDD'14, pages 1176–1185, New York, New York, USA, 2014.

[30] Dong Wang, Md Tanvir Amin, Shen Li, Tarek Abdelzaher, Lance Kaplan, Siyu Gu, Chenji Pan, Hengchang Liu, Charu C. Aggarwal, Raghu Ganti, Xinlei Wang, Prasant Mohapatra, Boleslaw Szymanski, and Hieu Le. Using Humans As Sensors: An Estimation-theoretic Perspective. In *Proc. of the 13th Intl. Symposium on Information Processing in Sensor Networks*, IPSN'14, pages 35–46, Berlin, Germany, 2014.

[31] Jing Gao, Qi Li, Bo Zhao, Wei Fan, and Jiawei Han. Truth Discovery and Crowdsourcing Aggregation: a Unified Perspective. *Proc. the VLDB Endowment*, 8(12):2048–2049, 2015.

[32] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and Fast—but is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'08, pages 254–263, Honolulu, Hawaii, 2008.

[33] Sorokin, Alexander and Forsyth, David. Utility data annotation with Amazon Mechanical Turk. In *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR'08, pages 1–8, Anchorage, AK, USA, 2008.

[34] Xiu Susie Fang, Xianzhi Wang, and Quan Z. Sheng. Ontology Augmentation via Attribute Extraction from Multiple Types of Sources. In *Proc.of the 26th Australasian Database Conference*, ADC'15, pages 16–27, Melbourne, VIC, Australia, 2015.

[35] Djamal Benslimane, Quan Z. Sheng, Mahmoud Barhamgi, and Henri Prade. The Uncertain Web: Concepts, Challenges, and Current Solutions. *ACM Transactions on Internet Technology (TOIT)*, 16(1):1, 2015.

[36] Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart. Corroborating Information from Disagreeing Views. In *Proc. of the 3th ACM Intl. Conference on Web Search and Data Mining*, WSDM'10, pages 131–140, New York, New York, USA, 2010.

[37] Jeff Pasternack and Dan Roth. Latent Credibility Analysis. In *Proc. of the 22th Intl. Conference on World Wide Web*, WWW'13, pages 1009–1020, Rio de Janeiro, Brazil, 2013.

[38] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. Integrating Conflicting Data: The Role of Source Dependence. *Proc. the VLDB Endowment*, 2(1):550–561, 2009.

[39] François Goasdoué, Konstantinos Karanasos, Yannis Katsis, Julien Leblay, Ioana Manolescu, and Stamatis Zampetakis. Fact Checking and Analyzing the Web. In *Proc. of the 2013 ACM SIGMOD Intl. Conference on Management of Data*, SIGMOD'13, pages 997–1000, New York, New York, USA, 2013.

[40] Qi Li, Yaliang Li, Jing Gao, Bo Zhao, Wei Fan, and Jiawei Han. Resolving Conflicts in Heterogeneous Data by Truth Discovery and Source Reliability Estimation. In *Proc. of the 2014 ACM SIGMOD Intl. Conference on Management of Data*, SIGMOD'14, pages 1187–1198, Snowbird, Utah, USA, 2014.

[41] Ravali Pochampally, Anish Das Sarma, Xin Luna Dong, Alexandra Meliou, and Divesh Srivastava. Fusing Data with Correlations. In *Proc. of the 2014 ACM SIGMOD Intl. Conference on Management of Data*, SIGMOD'14, pages 433–444, Snowbird, Utah, USA, 2014.

[42] Jeff Pasternack and Dan Roth. Comprehensive Trust Metrics for Information Networks. In *Proc. of the 27th Army Science Conference*, ASC'10, Orlando, Florida, 2010.

[43] Jeff Pasternack and Dan Roth. Knowing What to Believe (when You Already Know Something). In *Proc. of the 23th Intl. Conference on Computational Linguistics*, COLING'10, pages 877–885, Beijing, China, 2010.

[44] Xiaoxin Yin, Jiawei Han, and Philip S Yu. Truth Discovery with Multiple Conflicting Information Providers on the Web. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 20(6):796–808, 2008.

[45] Xin Luna Dong, Laure Berti-Equille, Yifan Hu, and Divesh Srivastava. Global Detection of Complex Copying Relationships between Sources. *Proc. the VLDB Endowment*, 3(1-2):1358–1369, 2010.

[46] Bo Zhao, Benjamin IP Rubinstein, Jim Gemmell, and Jiawei Han. A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration. *Proc. the VLDB Endowment*, 5(6):550–561, 2012.

[47] Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava. Truth Finding on the Deep Web: Is the Problem Solved? *Proc. the VLDB Endowment*, 6(2):97–108, 2013.

[48] Dalia Attia Waguih and Laure Berti-Equille. Truth Discovery Algorithms: An Experimental Evaluation. *arXiv preprint arXiv:1409.6428*, 2014.

[49] Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. A Survey on Truth Discovery. *ACM SIGKDD Explorations Newsletter*, 17(2):1–16, 2016.

[50] Amit Singhal. *Introducing the Knowledge Graph: Things, not Strings*. Official Google Blog, 2012.

[51] David A. Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Christopher Welty. Building Watson: An Overview of the DeepQA Project. 31(3):59–79, 2010.

[52] Yihong Zhang, Claudia Szabo, Quan Z. Sheng, and Xiu Susie Fang. Classifying Perspectives on Twitter: Immediate Observation, Affection, and Speculation. In *Proc. of the 16th Intl. Conference on Web Information Systems Engineering*, WISE'15, pages 493–507, Miami, Florida, USA., 2015.

[53] Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. Community-based Bayesian Aggregation Models for Crowdsourcing. In *Proc. of the 23th Intl. Conference on World Wide Web*, WWW'14, pages 155–164, Seoul, South Korea, 2014.

[54] Vikas C. Raykar and Shipeng Yu. Ranking Annotators for Crowdsourced Labeling Tasks. In *Proc. of the 24th Intl. Conference on Neural Information Processing Systems*, NIPS'11, pages 1809–1817, Granada, Spain, 2011.

[55] Jingbo Shang, Yu Zheng, Wenzhu Tong, Eric Chang, and Yong Yu. Inferring Gas Consumption and Pollution Emission of Vehicles Throughout a City. In *Proc. of the 20th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, KDD'14, pages 1027–1036, New York, New York, USA, 2014.

[56] Andy Yuan Xue, Rui Zhang, Yu Zheng, Xing Xie, Jianhui Yu, and Yong Tang. Desteller: a system for destination prediction based on trajectories with privacy protection. *Proc. the VLDB Endowment*, 6(12):1198–1201, 2013.

[57] Lina Yao, Quan Z Sheng, and Schahram Dustdar. Web-Based Management of the Internet of Things. *IEEE Internet Computing*, 19(4):60–67, 2015.

[58] Xin Luna Dong and Divesh Srivastava. Big data integration. In *Proc. of the 29th IEEE Intl. Conference on Data Engineering*, ICDE'13, pages 1245–1248, Brisbane, Australia, 2013.

[59] Xin Luna Dong and Divesh Srivastava. *Big Data Integration (Synthesis Lectures on Data Management)*. Morgan Claypool Publishers, 2015.

[60] Aditya Pal, Vibhor Rastogi, Ashwin Machanavajjhala, and Philip Bohannon. Information Integration over Time in Unreliable and Uncertain Environments. In *Proc. of the 21th Intl. Conference on World Wide Web*, WWW'12, pages 789–798, Lyon, France, 2012.

[61] Wenfei Fan, Zhe Fan, Chao Tian, and Xin Luna Dong. Keys for Graphs. *Proc. the VLDB Endowment*, 8(12):1590–1601, 2015.

[62] Pei Li, Xin Luna Dong, Songtao Guo, Andrea Maurino, and Divesh Srivastava. Robust Group Linkage. In *Proc. of the 24th Intl. Conference on World Wide Web*, WWW'15, pages 647–657, Florence, Italy, 2015.

[63] Anish Das Sarma, Xin Dong, and Alon Halevy. Bootstrapping Pay-as-you-go Data Integration Systems. In *Proc. of the 2008 ACM SIGMOD Intl. Conference on Management of Data*, SIGMOD'08, pages 861–874, Vancouver, Canada, 2008.

[64] Xin Dong, Alon Y. Halevy, and Cong Yu. Data Integration with Uncertainty. In *Proc. of the 33th Intl. Conference on Very Large Data Bases*, VLDB'07, pages 687–698, Vienna, Austria, 2007.

[65] Xin Luna Dong, Barna Saha, and Divesh Srivastava. Less is More: Selecting Sources Wisely for Integration. *Proc. the VLDB Endowment*, 6(2):37–48, 2012.

[66] Theodoros Rekatsinas, Xin Luna Dong, and Divesh Srivastava. Characterizing and Selecting Fresh Data Sources. In *Proc. of the 2014 ACM SIGMOD Intl. Conference on Management of Data*, SIGMOD'14, pages 919–930, Snowbird, Utah, USA, 2014.

[67] Zohra Bellahsene, Angela Bonifati, and Erhard (Eds.) Rahm. *Schema Matching and Mapping*. Springer, 2011.

[68] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1–16, 2007.

[69] Lise Getoor and Ashwin Machanavajjhala. Entity Resolution: Theory, Practice & Open Challenges. *Proc. the VLDB Endowment*, 5(12):2018–2019, 2012.

[70] Douglas B. Lenat and R. V. Guha. *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[71] Douglas B. Lenat. Cyc: A Large-Scale Investment in Knowledge Infrastructure. *Communications ACM*, 38(11):33–38, 1995.

[72] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[73] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. *Artificial Intelligence*, 194:28–61, 2012.

[74] Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. Open Language Learning for Information Extraction. In *Proc. of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL'12, pages 523–534, Jeju Island, South Korea, 2012.

[75] James Fan, David Ferrucci, David Gondek, and Aditya Kalyanpur. Prismatic: Inducing Knowledge from a Large Scale Lexicalized Relation Resource. In *Proc. of the NAACL HLT 2010 1th Intl. Workshop on Formalisms and Methodology for Learning by Reading*, FAM-LbR'10, pages 122–127, Los Angeles, California, 2010.

[76] Ndapandula Nakashole, Martin Theobald, and Gerhard Weikum. Scalable Knowledge Harvesting with High Precision and High Recall. In *Proc. of the Fourth ACM Intl. Conference on Web Search and Data Mining*, WSDM'11, pages 227–236, Hong Kong, China, 2011.

[77] Sherif Sakr and Ghazi Al-Naymat. Relational Processing of RDF Queries: A Survey. *SIGMOD Record*, 38(4):23–28, 2009.

[78] Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*, pages 1–136. Morgan & Claypool, 2011.

[79] Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, and Robert Baumgartner. Web Data Extraction, Applications and Techniques. *Knowledge-Based Systems*, 70(C):301–323, 2014.

[80] Alberto H. F. Laender, Berthier A. Ribeiro-Neto, Altigran S. da Silva, and Juliana S. Teixeira. A Brief Survey of Web Data Extraction Tools. *ACM SIGMOD Record*, 31(2):84–93, 2002.

[81] Hongkun Zhao. *Automatic Wrapper Generation for the Extraction of Search Result Records from Search Engines*. PhD thesis, Binghamton, NY, USA, 2007.

[82] Utku Irmak and Torsten Suel. Interactive Wrapper Generation with Minimal User Effort. In *Proc. of the 15th Intl. Conference on World Wide Web*, (WWW'06), pages 553–563, Edinburgh, Scotland, 2006.

[83] Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, and Khaled F. Shaalan. A Survey of Web Information Extraction Systems. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1411–1428, 2006.

[84] Hassan A. Sleiman and Rafael Corchuelo. A Survey on Region Extractors From Web Documents. *IEEE Transactions on Knowledge and Data Engineering*, 25(9):1960–1981, 2012.

[85] Xuan-Hieu Phan, Susumu Horiguchi, and Tu-Bao Ho. Automated Data Extraction from the Web with Conditional Models. *International journal of Business Intelligence and Data Mining*, 1(2):194–209, 2005.

[86] Jordi Turmo, Alicia Ageno, and Neus Català. Adaptive Information Extraction. *ACM Computing Survey*, 38(2), 2006.

[87] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Web Mining: Information and Pattern Discovery on the World Wide Web. In *Proc. of the 9th IEEE Intl. Conference on Tools with Artificial Intelligence*, ICTAI-97, pages 558–567, CA, USA, 1997.

[88] Rekha Jain and G. N. Purohit. Page Ranking Algorithms for Web Mining. *International Journal of Computer Applications*, 13(5):22–25, 2011.

[89] Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[90] Wenpu Xing and Ali Ghorbani. Weighted Page Rank Algorithm. In *Proc. of the 2th Annual Conference on Communication Networks and Services Research*, CNSR'04, pages 305–314, Fredericton, NB, Canada, 2004.

[91] Jon M Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632, 1999.

[92] Xiang Li and Ralph Grishman. Confidence Estimation for Knowledge Base Population. In *Proc. of the 9th Intl. Conference on Recent Advances in Natural Language Processin*, RANLP'13, pages 396–401, Hissar, Bulgaria, 2013.

[93] Michael Wick, Sameer Singh, Ari Kobren, and Andrew McCallum. Assessing Confidence of Knowledge Base Content with an Experimental Study in Entity Resolution. In *Proc. of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC'13, pages 13–18, San Francisco, California, USA, 2013.

[94] Simone Paolo Ponzetto and Michael Strube. Deriving a Large Scale Taxonomy from Wikipedia. In *Proc. of the 22th National Conference on Artificial Intelligence - Volume 2*, AAAI'07, pages 1440–1445, Vancouver, British Columbia, Canada, 2007.

[95] Fei Wu and Daniel S. Weld. Automatically Refining the Wikipedia Infobox Ontology. In *Proc. of the 17th Intl.l Conference on World Wide Web*, WWW'08, pages 635–644, Beijing, China, 2008.

[96] Girija Limaye, Sunita Sarawagi, and Soumen Chakrabarti. Annotating and Searching Web Tables Using Entities, Types and Relationships. *Proc. the VLDB Endowment*, 3(1-2):1338–1347, 2010.

[97] Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. Discovering and Exploring Relations on the Web. *Proc. the VLDB Endowment*, 5(12):1982–1985, 2012.

[98] Omar Alonso, Michael Gertz, and Ricardo Baeza-Yates. Clustering and Exploring Search Results Using Timeline Constructions. In *Proc.of the 18th ACM Conference on Information and Knowledge Management*, CIKM'09, pages 97–106, Hong Kong, China, 2009.

[99] Klaus Berberich, Srikanta Bedathur, Omar Alonso, and Gerhard Weikum. A Language Modeling Approach for Temporal Information Needs. In *Proceedings of the 32Nd European Conference on Advances in Information Retrieval*, ECIR'2010, pages 13–25, Milton Keynes, UK, 2010.

[100] Marius Pasca. Towards Temporal Web Search. In *Proc. of the 2008 ACM Symposium on Applied Computing*, SAC'08, pages 1117–1121, Fortaleza, Ceara, Brazil, 2008.

[101] Yafang Wang, Mingjie Zhu, Lizhen Qu, Marc Spaniol, and Gerhard Weikum. Timely YAGO: Harvesting, Querying, and Visualizing Temporal Knowledge from Wikipedia. In *Proc. of the 13th Intl. Conference on Extending Database Technology*, EDBT'10, pages 697–700, Lausanne, Switzerland, 2010.

[102] Qi Zhang, Fabian M. Suchanek, Lihua Yue, and Gerhard Weikum. TOB: Timely Ontologies for Business Relations. In *Proc. of the 11th Intl. Workshop on Web and Databases*, WebDB'08, Vancouver, Canada, 2008.

[103] Dian Yu, Hongzhao Huang, Taylor Cassidy, Heng Ji, Chi Wang, Shi Zhi, Jiawei Han, Clare Voss, and Malik Magdon-Ismail. The Wisdom of Minority: Unsupervised Slot Filling Validation based on Multi-dimensional Truth-Finding. In *Proc. of the 25th Intl. Conference on Computational Linguistics*, COLING'14, pages 1567–1578, Dublin, Ireland, 2014.

[104] Furong Li, Mong Li Lee, and Wynne Hsu. Entity Profiling with Varying Source Reliabilities. In *Proc. of the 20th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, KDD'14, pages 1146–1155, New York, New York, USA, 2014.

[105] Xin Luna Dong and Divesh Srivastava. Compact explanation of data fusion decisions. In *Proc. of the 22th Intl. Conference on World Wide Web*, WWW'13, pages 379–390, Rio de Janeiro, Brazil, 2013.

[106] Manish Gupta, Yizhou Sun, and Jiawei Han. Trust Analysis with Clustering. In *Proc. of the 20th Intl. Conference Companion on World Wide Web*, WWW'11, pages 53–54, Hyderabad, India, 2011.

[107] Amélie Marian and Minji Wu. Corroborating information from web sources. *IEEE Data Engineering Bulletin*, 34(3):11–17, 2011.

[108] Jens Bleiholder and Felix Naumann. Conflict Handling Strategies in an Integrated Information System. In *Proc. of 5th Intl. Workshop on Information Integration on the Web*, IIWeb'06, Edinburgh Scotland, 2006.

[109] Felix Naumann, Alexander Bilke, Jens Bleiholder, and Melanie Herschel. Data Fusion in Three Steps: Resolving Schema, Tuple, and Value Inconsistencies. *IEEE Data Engineering Bulletin*, 29(2):21–31, 2006.

[110] Jens Bleiholder and Felix Naumann. Data Fusion. *ACM Computing Surveys (CSUR)*, 41(1):1–41, 2009.

[111] Xin Luna Dong and Felix Naumann. Data Fusion: Resolving Data Conflicts for Integration. *Proc. the VLDB Endowment*, 2(2):1654–1655, 2009.

[112] Minji Wu and Amélie Marian. Corroborating Facts from Affirmative Statements. In *Proc. of the 17th Intl. Conference on Extending Database Technology*, EDBT'14, pages 157–168, Athens, Greece, 2014.

[113] Guo-Jun Qi, Charu C. Aggarwal, Jiawei Han, and Thomas Huang. Mining Collective Intelligence in Diverse Groups. In *Proc. of the 22th Intl. Conference on World Wide Web*, WWW'13, pages 1041–1052, Rio de Janeiro, Brazil, 2013.

[114] Jeff Pasternack and Dan Roth. Making Better Informed Trust Decisions with Generalized Fact-Finding. In *Proc. of the 22th Intl. Joint Conference on Artificial Intelligence*, IJCAI'11, pages 2324–2329, Barcelona, Catalonia, Spain, 2011.

[115] Xiaoxin Yin and Wenzhao Tan. Semi-supervised Truth Discovery. In *Proc. of the 20th Intl. Conference on World Wide Web*, WWW'11, pages 217–226, Hyderabad, India, 2011.

[116] Xianzhi Wang, Quan Z. Sheng, Xiu Susie Fang, Lina Yao, Xiaofei Xu, and Xue Li. An Integrated Bayesian Approach for Effective Multi-Truth Discovery. In *Proc. of the 24th ACM Intl. on Conference on Information and Knowledge Management*, CIKM'15, pages 493–502, Melbourne, Australia, 2015.

[117] Gjergji Kasneci, Jurgen Van Gael, David Stern, and Thore Graepel. CoBayes: Bayesian Knowledge Corroboration with Assessors of Unknown Areas of Expertise. In *Proc. of the 4th ACM Intl. Conference on Web Search and Data Mining*, WSDM'11, pages 465–474, Hong Kong, China, 2011.

[118] Bo Zhao and Jiawei Han. A Probabilistic Model for Estimating Real-Valued Truth from Conflicting Sources. In *Proc. of 10th Intl. Workshop on Quality in Databases*, QDB'12, Istanbul, Turkey, 2012.

[119] Balaji Lakshminarayanan and Yee Whye Teh. Inferring Ground Truth from Multi-Annotator Ordinal Data: A Probabilistic Approach. *arXiv preprint arXiv:1305.0015*, 2013.

[120] Hengtong Zhang, Qi Li, Fenglong Ma, Houping Xiao, Yaliang Li, Jing Gao, and Lu Su. Influence-Aware Truth Discovery. In *Proc. of the 25th ACM Intl. on Conference on Information and Knowledge Management*, CIKM'16, pages 851–860, Indianapolis, Indiana, USA, 2016.

[121] Dong Wang, Lance Kaplan, Hieu Le, and Tarek Abdelzaher. On Truth Discovery in Social Sensing: A Maximum Likelihood Estimation Approach. In *Proc. of the 11th Intl. Conference on Information Processing in Sensor Networks*, IPSN'12, pages 233–244, Beijing, China, 2012.

[122] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[123] Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, and Jiawei Han. A Confidence-Aware Approach for Truth Discovery on Long-Tail Data. *Proc. the VLDB Endowment*, 8(4):425–436, 2014.

[124] Chuishi Meng, Wenjun Jiang, Yaliang Li, Jing Gao, Lu Su, Hu Ding, and Yun Cheng. Truth Discovery on Crowd Sensing of Correlated Entities. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, SenSys'15, pages 169–182, Seoul, South Korea, 2015.

[125] Xianzhi Wang, Quan Z. Sheng, Lina Yao, Xue Li, Xiu Susie Fang, Xiaofei Xu, and Boualem Benatallah. Truth Discovery via Exploiting Implications from Multi-Source Data. In *Proc. of the 25th ACM Intl. on Conference on Information and Knowledge Management*, CIKM'16, pages 861–870, Indianapolis, Indiana, USA, 2016.

[126] Shi Zhi, Bo Zhao, Wenzhu Tong, Jing Gao, Dian Yu, Heng Ji, and Jiawei Han. Modeling Truth Existence in Truth Discovery. In *Proc. of the 21th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, KDD'15, pages 1543–1552, Sydney, NSW, Australia, 2015.

[127] Xianzhi Wang, Quan Z. Sheng, Lina Yao, Xue Li, Xiu Susie Fang, Xiaofei Xu, and Boualem Benatallah. Empowering Truth Discovery with Multi-Truth Prediction. In *Proc. of the 25th ACM Intl. on Conference on Information and Knowledge Management*, CIKM'16, pages 881–890, Indianapolis, Indiana, USA, 2016.

[128] Mengting Wan, Xiangyu Chen, Lance Kaplan, Jiawei Han, Jing Gao, and Bo Zhao. From Truth Discovery to Trustworthy Opinion Discovery: An Uncertainty-Aware Quantitative Modeling Approach. In *Proc. of the 22th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, KDD'16, pages 1885–1894, San Francisco, California, USA, 2016.

[129] Xian Li, Xin Luna Dong, Kenneth B. Lyons, Weiyi Meng, and Divesh Srivastava. Scaling up Copy Detection. In *Proc. of the 31th Intl. Conference on Data Engineering*, ICDE'15, pages 89–100, Seoul, South Korea, 2015.

[130] Lorenzo Blanco, Valter Crescenzi, Paolo Merialdo, and Paolo Papotti. Probabilistic Models to Reconcile Complex Data from Inaccurate Data Sources. In *Proc. of the 22th Intl. Conference on Advanced Information Systems Engineering*, CAiSE'10, pages 83–97, Hammamet, Tunisia, 2010.

[131] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. Truth Discovery and Copying Detection in a Dynamic World. *Proc. the VLDB Endowment*, 2(1):562–573, 2009.

[132] Xuan Liu, Luna Xin Dong, Beng Chin Ooi, and Divesh Srivastava. Online Data Fusion. *Proc. the VLDB Endowment*, 4(11):932–943, 2011.

[133] M. Lamine Ba, Roxana Horincar, Pierre Senellart, and Huayu Wu. Truth Finding with Attribute Partitioning. In *Proc. of the 18th Intl. Workshop on Web and Databases*, WebDB'15, pages 27–33, Melbourne, VIC, Australia, 2015.

[134] Fenglong Ma, Yaliang Li, Qi Li, Minghui Qiu, Jing Gao, Shi Zhi, Lu Su, Bo Zhao, Heng Ji, and Jiawei Han. FaitCrowd: Fine Grained Truth Discovery for Crowdsourced Data Aggregation. In *Proc. of the 21th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, KDD'15, pages 745–754, Sydney, NSW, Australia, 2015.

[135] Xianzhi Wang, Quan Z. Sheng, Xiu Susie Fang, Xue Li, Xiaofei Xu, and Lina Yao. Approximate Truth Discovery via Problem Scale Reduction. In *Proc. of the 24th ACM Intl. on Conference on Information and Knowledge Management*, CIKM'15, pages 503–512, Melbourne, Australia, 2015.

[136] Laure Berti-Équille. Data Veracity Estimation with Ensembling Truth Discovery Methods. In *Proc. of the 2015 IEEE Intl. Conference on Big Data*, pages 2628–2636, CA, USA, 2015.

[137] Zhou Zhao, James Cheng, and Wilfred Ng. Truth Discovery in Data Streams: A Single-Pass Probabilistic Approach. In *Proc. of the 23th ACM Intl. Conference on Conference on Information and Knowledge Management*, CIKM'14, pages 1589–1598, Shanghai, China, 2014.

[138] Houping Xiao, Yaliang Li, Jing Gao, Fei Wang, Liang Ge, Wei Fan, Long Vu, and Deepak Turaga. Believe It Today or Tomorrow Detecting Untrustworthy Information from Dynamic Multi-Source Data. In *Proc. of the 2015 SIAM Intl. Conference on Data Mining*, SDM'15, pages 397–405, British Columbia, Canada, 2015.

[139] Yaliang Li, Qi Li, Jing Gao, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. On the Discovery of Evolving Truth. In *Proc. of the 21th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, KDD'15, pages 675–684, Sydney, NSW, Australia, 2015.

[140] Chenglin Miao, Wenjun Jiang, Lu Su, Yaliang Li, Suxin Guo, Zhan Qin, Houping Xiao, Jing Gao, and Kui Ren. Cloud-Enabled Privacy-Preserving Truth Discovery in Crowd Sensing Systems. In *Proc. of the 13th ACM Conference on Embedded Networked Sensor Systems*, SenSys'15, pages 183–196, Seoul, South Korea, 2015.

[141] Houping Xiao, Jing Gao, Qi Li, Fenglong Ma, Lu Su, Yunlong Feng, and Aidong Zhang. Towards Confidence in the Truth: A Bootstrapping Based Truth Discovery Approach. In *Proc. of the 22th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, KDD'16, pages 1935–1944, San Francisco, California, USA, 2016.

[142] Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. Knowledge-based Trust: Estimating the Trustworthiness of Web Sources. *Proc. VLDB Endowment*, 8(9):938–949, 2015.

[143] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, pages 20–28, 1979.

[144] Dengyong Zhou, John C. Platt, Sumit Basu, and Yi Mao. Learning from the Wisdom of Crowds by Minimax Entropy. In *Proc. of the 25th Intl. Conference on Neural Information Processing Systems*, NIPS'12, pages 2195–2203, Lake Tahoe, Nevada, 2012.

[145] Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In *Proc. of the 22Nd Intl. Conference on Neural Information Processing Systems*, NIPS'09, pages 2035–2043, Vancouver, British Columbia, Canada, 2009.

[146] Merrielle Spain and Pietro Perona. Some Objects Are More Equal Than Others: Measuring and Predicting Importance. In *Proc. of the 10th European Conference on Computer Vision: Part I*, ECCV'08, pages 523–536, Marseille, France, 2008.

[147] Peter Welinder, Steve Branson, Serge Belongie, and Pietro Perona. The Multidimensional Wisdom of Crowds. In *Proc. of the 23rd Intl. Conference on Neural Information Processing Systems*, NIPS'10, pages 2424–2432, Vancouver, British Columbia, Canada, 2010.

[148] Hongwei Li, Bo Zhao, and Ariel Fuxman. The Wisdom of Minority: Discovering and Targeting the Right Group of Workers for Crowdsourcing. In *Proc. of the 23rd Intl. Conference on World Wide Web*, WWW'14, pages 165–176, Seoul, South Korea, 2014.

[149] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Anna Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy. Supervised Learning from Multiple Experts: Whom to Trust when Everyone Lies a Bit. In *Proc. of the 26th Annual Intl. Conference on Machine Learning*, ICML'09, pages 889–896, Montreal, Québec, Canada, 2009.

[150] Padhraic Smyth, Usama Fayyad, Michael Burl, Pietro Perona, and Pierre Baldi. Inferring Ground Truth from Subjective Labelling of Venus Images. In *Proc. of the 7th Intl. Conference on Neural Information Processing Systems*, NIPS-94, pages 1085–1092, Denver, Colorado, 1994.

[151] Bahadir Ismail Aydin, Yavuz Selim Yilmaz, Yaliang Li, Qi Li, Jing Gao, and Murat Demirbas. Crowdsourcing for Multiple-choice Question Answering. In *Proc. of the 28th AAAI Conference on Artificial Intelligence*, AAAI'14, pages 2946–2953, Québec City, Québec, Canada, 2014.

[152] Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. In *Proc. of the 14th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, KDD'08, pages 614–622, Las Vegas, Nevada, USA, 2008.

[153] Lu Su, Qi Li, Shaohan Hu, Shiguang Wang, Jing Gao, Hengchang Liu, Tarek F. Abdelzaher, Jiawei Han, Xue Liu, Yan Gao, and Lance Kaplan. Generalized Decision Aggregation in Distributed Sensing Systems. In *Proc. of the IEEE Real-Time Systems Symposium*, RTSS'14, pages 1–10, Rome, Italy, 2014.

[154] Hieu Le, Dong Wang, Hossein Ahmadi, Yusuf S. Uddin, Boleslaw Szymanski, Raghu Ganti, Tarek Abdelzaher, Omid Fatemieh, Hongyang Wang, Jeff Pasternack, Jiawei Han, Dan Roth, Sibel Adali, and Hui Lei. *Demo: Distilling likely truth from noisy streaming data with Apollo*, pages 417–418. SenSys'11. Seattle, Washington, 2011.

[155] Shiguang Wang, Dong Wang, Lu Su, Lance Kaplan, and Tarek F. Abdelzaher. Towards cyber-physical systems in social spaces: The data reliability challenge. In *Proc. of the IEEE Real-Time Systems Symposium*, RTSS'14, pages 74–85, Rome, Italy, 2014.

[156] Shiguang Wang, Lu Su, Shen Li, Shaohan Hu, Tanvir Amin, Hongwei Wang, Shuochao Yao, Lance Kaplan, and Tarek Abdelzaher. Scalable Social Sensing of Interdependent Phenomena. In *Proc. of the 14th Intl. Conference on Information Processing in Sensor Networks*, IPSN'15, pages 202–213, Seattle, Washington, 2015.

[157] Dong Wang, Lance Kaplan, and Tarek F. Abdelzaher. Maximum Likelihood Analysis of Conflicting Observations in Social Sensing. *ACM Transactions on Sensor Networks (ToSN)*, 10(2):30:1–30:27, 2014.

[158] Charu C. Aggarwal and Tarek Abdelzahe. *Social Sensing*, pages 237–297. 2013.

[159] Xiu Susie Fang, Quan Z. Sheng, Xianzhi Wang, Anne H.H. Ngu, and Yihong Zhang. GrandBase: Generating Actionable Knowledge from Big Data. *Intl. Journal on PSU Research Review*, 1(2):105–126, 2017.

[160] Michael Wick, Sameer Singh, Harshal Pandya, and Andrew McCallum. A Joint Model for Discovering and Linking Entities. In *Proc. of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC'13, pages 67–72, San Francisco, California, USA, 2013.

[161] Taesung Lee, Zhongyuan Wang, Haixun Wang, and Seung-won Hwang. Attribute Extraction and Scoring: A Probabilistic Approach. In *Proc. of 29th Intl. Conference on Data Engineering*, ICDE'13, pages 194–205, Brisbane, Australia, 2013.

[162] Marius Pasca and Benjamin Van Durme. What You Seek is What You Get: Extraction of Class Attributes from Query Logs. In *Proc. of the 20th Intl. Joint Conference on Artificial Intelligence*, IJCAI'07, pages 2832–2837, Hyderabad, India, 2007.

[163] Marius Paşca, Enrique Alfonseca, Enrique Robledo-Arnuncio, Ricardo Martin-Brualla, and Keith Hall. The Role of Query Sessions in Extracting Instance Attributes from Web Search Queries. In *Proc. of the 32th European Conference on Information Retrieval*, (ECIR'10), pages 62–74, Milton Keynes, United Kingdom, 2010.

[164] Brad Adelberg. NoDoSE - A Tool for Semi-automatically Extracting Structured and Semistructured Data from Text Documents. *ACM SIGMOD Record*, 27(2):283–294, June 1998.

[165] Ling Liu, Calton Pu, and Wei Han. XWRAP: an XML-enabled Wrapper Construction System for Web Information Sources. In *Proc. of the 16th Intl. Conference on Data Engineering*, (ICDE'00), pages 611–621, San Diego, California, USA, 2000.

[166] Jun Zhu, Zaiqing Nie, Ji-Rong Wen, Bo Zhang, and Wei-Ying Ma. Simultaneous Record Detection and Attribute Labeling in Web Data Extraction. In *Proc. of the 12th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, KDD'06, pages 494–503, Philadelphia, PA, USA, 2006.

[167] Trausti Kristjansson, Aron Culotta, Paul Viola, and Andrew McCallum. Interactive Information Extraction with Constrained Conditional Random Fields. In *Proc. of the 19th National Conference on Artifical Intelligence*, AAAI'04, pages 412–418, San Jose, California, 2004.

[168] Valter Crescenzi, Giansalvatore Mecca, and Paolo Merialdo. RoadRunner: Automatic Data Extraction from Data-intensive Web Sites. In *Proc. of the 2002 ACM SIGMOD Intl. Conference on Management of Data*, SIGMOD'02, pages 624–624, Madison, Wisconsin, 2002.

[169] Arvind Arasu and Hector Garcia-Molina. Extracting Structured Data from Web Pages. In *Proc. of the 2003 ACM SIGMOD Intl. Conference on Management of Data*, SIGMOD'03, pages 337–348, San Diego, California, 2003.

[170] Aria Haghighi and Dan Klein. Simple Coreference Resolution with Rich Syntactic and Semantic Features. In *Proc. of the 2009 Conference on Empirical Methods in Natural Language Processing*, EMNLP'09, pages 1152–1161, Singapore, 2009.

[171] Xiu Susie Fang, Quan Z. Sheng, Xianzhi Wang, Mahmoud Barhamgi, Lina Yao, and Anne H.H. Ngu. SourceVote: Fusing Multi-Valued Data via Inter-Source Agreements. In *Proc. of the 36th Intl. Conference on Conceptual Modeling*, ER'17, pages 164–172, Valencia, Spain, 2017.

[172] Xiu Susie Fang, Quan Z. Sheng, Xianzhi Wang, and Anne H.H. Ngu. SmartVote: A Full-Fledged Graph-Based Model for Multi-Valued Truth Discovery. *World Wide Web Journal (WWWJ), under minor revision.*, 2017.

[173] David F. Gleich, Paul G. Constantine, Abraham D. Flaxman, and Asela Gunawardana. Tracking the Random Surfer: Empirically Measured Teleportation Parameters in PageRank. In *Proc. of the 19th Intl. Conference on World Wide Web*, WWW'10, pages 381–390, Raleigh, North Carolina, USA, 2010.

[174] Anish Das Sarma, Xin Luna Dong, and Alon Halevy. Data Integration with Dependent Sources. In *Proc. of the 14th Intl. Conference on Extending Database Technology*, EDBT/ICDT'11, pages 401–412, Uppsala, Sweden, 2011.

[175] Xiu Susie Fang, Quan Z. Sheng, Xianzhi Wang, and Anne H.H. Ngu. Value Veracity Estimation for Multi-Truth Objects via a Graph-Based Approach. In *Proc. of the 26th Intl. World Wide Web Conference*, WWW'17 Companion, pages 777–778, Perth, Australia, 2017.

[176] Xiu Susie Fang. Truth Discovery from Conflicting Multi-Valued Objects. In *Proc. of the 26th Intl. World Wide Web Conference*, WWW'17 Companion, pages 711–715, Perth, Australia, 2017.

[177] Xiu Susie Fang, Quan Z. Sheng, Xianzhi Wang, and Anne H.H. Ngu. SmartMTD: A Graph-Based Approach for Effective Multi-Truth Discovery. In *Proc. of the 41th Intl. ACM SIGIR Conference on Research and Development in Information Retrieval*, (SIGIR'18), submitted in 2018.

[178] Kilem L Gwet. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Advanced Analytics, LLC, 2014.

[179] Wenfei Fan. Data Quality: Theory and Practice. In *Proc. of the13th Conference on Web-Age Information Management*, WAIM'12, pages 1–16, Harbin, China, 2012.

[180] Wenfei Fan, Floris Geerts, Shuia Ma, Nan Tang, and Wenyuan Yu. Data Quality Problems beyond Consistency and Duduplication. In *In Search of Elegance in the Theory and Practice of Computation*, pages 237–249, 2013.

[181] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media. In *Proc. of the 26th Intl. World Wide Web Conference*, WWW'17, pages 1003–1012, Perth, Australia, 2017.

[182] Xiu Susie Fang, Quan Z. Sheng, and Xianzhi Wang. An Ensemble Approach for Better Truth Discovery. In *Proc. of the 12th Anniversary of the Intl. Conference on Advanced Data Mining and Applications*, (ADMA'16), pages 298–311, Gold Coast, Australia, 2016.

[183] Thomas G. Dietterich. Ensemble Methods in Machine Learning. In *Proc. of the First Intl. Workshop on Multiple Classifier Systems*, (MCS'00), pages 1–15, Cagliari, Italy, 2000.

[184] Pedro Domingos and Michael Pazzani. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*, 29(2):103–130, 1997.

[185] Xiu Susie Fang, Quan Z. Sheng, Xianzhi Wang, Wei Emma Zhang, and Anne H.H. Ngu. How to Compare Truth Discovery Methods When Ground Truth is Missing? In *Proc. of the 41th Intl. ACM SIGIR Conference on Research and Development in Information Retrieval*, (SIGIR'18), submitted in 2018.

[186] Houping Xiao, Jing Gao, Zhaoran Wang, Shiyu Wang, Lu Su, and Han Liu. A Truth Discovery Approach with Theoretical Guarantee. In *Proc. of the 22th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, KDD'16, pages 1925–1934, San Francisco, California, USA, 2016.