Truth Discovery and Crowdsourcing Aggregation: A Unified Perspective

Jing Gao[1], Qi Li[1], Bo Zhao[2], Wei Fan[3], and Jiawei Han[4]
[1]SUNY Buffalo, Buffalo, NY USA [2]LinkedIn, Mountain View, CA USA
[3]Baidu Research Big Data Lab, China [4]Univesrity of Illinois, Urbana, IL USA
[1]{jing,qli22}@buffalo.edu, [2]bozhao@linkedin.com, [3]fanwei03@baidu.com, [4]hanj@illinois.edu

ABSTRACT

In the era of Big Data, data entries, even describing the same objects or events, can come from a variety of sources, where a data source can be a web page, a database or a person. Consequently, conflicts among sources become inevitable. To resolve the conflicts and achieve high quality data, truth discovery and crowdsourcing aggregation have been studied intensively. However, although these two topics have a lot in common, they are studied separately and are applied to different domains. To answer the need of a systematic introduction and comparison of the two topics, we present an organized picture on truth discovery and crowdsourcing aggregation in this tutorial. They are compared on both theory and application levels, and their related areas as well as open questions are discussed.

1. INTRODUCTION

In the past decade, **truth discovery** methods have emerged as a powerful tool to resolve the conflicts among data sources. They can detect truths among conflicting information by integrating source reliability estimation in data fusion. The topic of truth discovery has attracted lots of attentions with a variety of emphases to tackle different challenges. The success has been witnessed in numerous applications, including data integration, web mining, knowledge base construction, etc.

A highly relevant field is **crowdsourcing aggregation**, a hot topic in the field of crowdsourcing. Crowdsourcing is the process of completing some tasks (e.g., answer a set of questions) by soliciting contributions from a large group of people. One important task in crowdsourcing is to aggregate noisy answers contributed by crowd workers to obtain the correct answers. As the workers may have diverse levels of expertise, it is important to estimate worker abilities in the aggregation. Along this direction, many crowdsourcing aggregation approaches have been proposed.

Truth discovery and crowdsourcing aggregation have been studied separately, and they are applied to different domains. However, these two topics have a lot in common: 1) Their goals are to improve the quality of aggregation results; 2) They hold similar assumptions that reliable sources(workers) tend to provide high quality information and information from reliable sources(workers) is more likely to be accurate. On the other hand, differences in the two fields motivate approaches to tackle different challenges. Some major differences are caused by data generation. Truth discovery is typically applied to web and database information integration, where the data is already available, so it is passive in terms of data generation. In contrast, crowdsourcing is active and requesters have better control on what and how data are generated.

With such commonalities and differences between truth discovery and crowdsourcing aggregation, it is important to bring people's attention to the study and exploration of both fields. Despite the high similarity between two fields, few efforts have been contributed to connect the studies on these topics. In this 1.5-hour tutorial, we will present a systematic introduction and comparison of the two topics, including their applications, models, related areas, and open questions.

Proceedings of the VLDB Endowment, Vol. 8, No. 12

2. TUTORIAL OUTLINE

The tutorial is presented based on the following outline.

Overview

We start with an overview on truth discovery and crowdsourcing aggregation, and their broader impact in crowdsourcing, social sensing, web mining, question answering, knowledge base constructing, and data management.

Truth Discovery. As multiple data sources may provide conflict information for the same object, the task of truth discovery is to detect trustworthy information by identifying reliable sources. Truth discovery methods are usually unsupervised or semi-supervised.

Crowdsourcing Aggregation. In crowdsourcing, requesters post their tasks on crowdsourcing platforms and pay crowd workers for their answers to those tasks. Since single worker may provide incorrect answers, requesters usually hire several workers for the same task. As the answering results may not achieve consensus, crowdsourcing aggregation is widely used to find the correct answers by estimating workers' reliability degrees and expertise.

Truth discovery and crowdsourcing aggregation methods are successfully applied in many domains to solve a variety of real-world problems, such as integrating databases, detecting trustworthy information on the Web and in social media, building training data for machine learning tasks, creating "super players" for trivia games.

Comparison of the Two Fields

Similarities. As described above, both fields are trying to find trustworthy and accurate information among the conflicting information from multiple sources. They share a common goal to improve the quality of aggregated results by estimating source reliabilities. They also have similar basic principles, i.e., data from reliable sources are more likely to be accurate and a source is reliable if it provides accurate information. In terms of techniques,

as the ground truth is usually unavailable, methods in both fields need to work without supervision, and an iterative procedure is frequently adopted. Due to these similarities in problem settings and techniques, there are many overlaps of applications.

Differences. The main differences of truth discovery and crowdsourcing aggregation come from the differences of application scenarios. Generally speaking, truth discovery is passive (data is already generated, we just find what is available), while crowdsourcing is active (we can choose what and how much data to generate). In truth discovery, data crawled from the Web or collected from databases may have various types and may change dynamically. As a result, methods for heterogeneous data [8], for streaming data [15], and for online truth discovery [9] are designed. In crowdsourcing, however, there are some unique features that truth discovery does not have. Crowdsourcing aggregation methods can access more information on the source features, such as workers' location, accuracy on historical tasks, and education background [6]. Moreover, the requesters can provide helpful feature information on the instances that can be used for aggregation [12].

Model Comparison

In order to clearly compare the two fields, we summarize popular methods in truth discovery and crowdsourcing aggregation by the techniques they use.

Statistical Models. In general, statistical models are used to estimate the probability of an event. Here, in order to distinguish from the probabilistic graphical model discussed below, we refer to the statistical model as the model that takes no prior distribution. Many classic methods in truth discovery and crowdsourcing aggregation are formulated as statistical models [3, 4, 13]. These methods provide results with concrete statistical meanings. Based on different assumptions and formulations, the methods suit different application scenarios.

Probabilistic Graphical Models. The probabilistic graphical model (PGM) is very popular and important in truth discovery and crowdsourcing aggregation. Thanks to the advantage that prior knowledge can be easily incorporated, many methods use PGM as their core model [1, 14, 12]. Although researchers use different distributions and consider different relations based on their specific tasks, there are three major variable spaces in common: source/worker, object/question, and claim/answer spaces.

Optimization Models. Truth discovery and crowdsourcing aggregation can also be modeled as optimization tasks [16, 8, 7], where the final results are given by minimizing the objective function under constraints.

Related Areas

We provide an overview of the following areas that are most relevant to the proposed tutorial topics, which helps the understanding of the global picture.

Information Integration and Data Cleaning. Information integration and data cleaning are two important research topics in database management. Information integration is a broad research topic including entity resolution, schema mapping, data fusion, etc[5]. Data cleaning is another related topic. Unlike truth discovery that handles multi-source information, data cleaning focuses on the improvement of data quality for a single source [10].

Crowdsourcing. In addition to crowdsourcing aggregation, crowdsourcing consists of many other tasks, such as the designing of questions, the designing of platforms, budget allocation [2], and pricing mechanisms [11]. How these tasks are conducted may significantly affect the result of crowdsourcing aggregation.

There are a few other related topics, including knowledge graph, social sensing, web mining, information extraction, ensemble learning, etc. We provide a brief introduction for these topics too.

Open Questions and Resources

Unstructured Data. Unstructured data are common in many domains, such as the data on the Web and social media platforms. Current truth discovery and crowdsourcing aggregation mainly consider structured data as input. Techniques to extract useful information and aggregate unstructured data need to be developed.

Data with Complex Relations. When considering data relations, current truth discovery and crowdsourcing aggregation methods only consider relations among sources. However, data relation can be more complex. For example, if "Pres Obama is born in Hawaii" from one source is true, then "Pres Obama is born in the USA" from another source should be also true. The data relations need to be considered as they can help the methods improve the performance.

Evaluation. How to evaluate the aggregated results is still an open question, as ground truth information can be difficult to obtain in real life. Theoretical analyses need more attention from both communities.

Finally, we conclude the tutorial with pointers to available resources (e.g., datasets, software, and surveys) in both truth discovery and crowdsourcing aggregation.

3. REFERENCES

[1] Y. Bachrach, T. Graepel, T. Minka, and J. Guiver. How to grade a test without knowing the answers—A Bayesian graphical model for adaptive crowdsourcing and aptitude testing. In *Proc. of ICML*, pages 1183–1190, 2012.

[2] X. Chen, Q. Lin, and D. Zhou. Optimistic knowledge gradient policy for optimal budget allocation in crowdsourcing. In *Proc. of ICML*, pages 64–72, 2013.

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

[4] X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: The role of source dependence. *PVLDB*, 2(1):550–561, 2009.

[5] L. Haas. Beauty and the beast: The theory and practice of information integration. In *Proc. of ICDT*, pages 28–43, 2006.

[6] H. Li, B. Zhao, and A. Fuxman. The wisdom of minority: discovering and targeting the right group of workers for crowdsourcing. In *Proc. of WWW*, pages 165–176, 2014.

[7] Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, and J. Han. A confidence-aware approach for truth discovery on long-tail data. *PVLDB*, 8(4), 2015.

[8] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proc. of SIGMOD*, pages 1187–1198, 2014.

[9] X. Liu, X. L. Dong, B. C. Ooi, and D. Srivastava. Online data fusion. *PVLDB*, 4(11):932–943, 2011.

[10] E. Rahm and H. H. Do. Data cleaning: Problems and current approaches. *IEEE Bulletin on Data Engineering*, 23(4):3–13, 2000.

[11] Y. Singer and M. Mittal. Pricing mechanisms for crowdsourcing markets. In *Proc. of WWW*, pages 1157–1166, 2013.

[12] P. Welinder, S. Branson, P. Perona, and S. J. Belongie. The multidimensional wisdom of crowds. In *NIPS*, pages 2424–2432, 2010.

[13] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. In *Proc. of KDD*, pages 1048–1052, 2007.

[14] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. *PVLDB*, 5(6):550–561, 2012.

[15] Z. Zhao, J. Cheng, and W. Ng. Truth discovery in data streams: A single-pass probabilistic approach. In *Proc. of CIKM*, pages 1589–1598, 2014.

[16] D. Zhou, S. Basu, Y. Mao, and J. C. Platt. Learning from the wisdom of crowds by minimax entropy. In *NIPS'12*, pages 2195–2203, 2012.