J Inf Process Syst, Vol.12, No.2, pp.234~248, June 2016
http://dx.doi.org/10.3745/JIPS.04.0024

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Comparative Study of Evaluating the Trustworthiness of Data Based on Data Provenance

Kuldeep Gurjar* and Yang-Sae Moon*

Abstract

Due to the proliferation of data being exchanged and the increase of dependency on this data for critical decision-making, it has become imperative to ensure the trustworthiness of the data at the receiving end in order to obtain reliable results. Data provenance, the derivation history of data, is a useful tool for evaluating the trustworthiness of data. Various frameworks have been proposed to evaluate the trustworthiness of data based on data provenance. In this paper, we briefly review a history of these frameworks for evaluating the trustworthiness of data and present an overview of some prominent state-of-the-art evaluation frameworks. Moreover, we provide a comparative analysis of two key frameworks by evaluating various aspects in an executional environment. Our analysis points to various open research issues and provides an understanding of the functionalities of the frameworks that are used to evaluate the trustworthiness of data.

Keywords

Data Provenance, Trustworthiness of Data, Data Quality, Trustworthiness Evaluation, Trust Score

1. Introduction

In recent applications, a large amount of data that conveys important information has been collected from various distributed sources. Therefore, to produce an accurate analysis, it is imperative to ensure that the data received is trustworthy. *Data provenance* is comprised of techniques to evaluate the trustworthiness of data received from various sources. Data provenance, which is also referred to as lineage, has been assigned different definitions by various research groups with different viewpoints. Lanter [1] defined data provenance in the context of a geographic information system (GIS). Smith et al. [2] and Eagan and Ventura [3] described data provenance as being for the purpose of retrieving environmental data. Woodruff and Stonebraker [4] defined the data provenance for the database environment as a processing history, which includes its origin and all subsequent processing steps applied to it. Buneman et al. [5] described data provenance as the description of the origins of data and the process by which it arrives. According to Wadhwa and Kamalapur [6], data provenance is one kind of metadata that tracks the steps by which the data is derived. Moreau and Missier [7] explained data provenance as being the record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing. For the purpose of this paper, we have

Manuscript received December 3, 2014; accepted January 19, 2016.

Corresponding Author: Yang-Sae Moon (ysmoon@kangwon.ac.kr)
* Dept. of Computer Science, Kangwon National University, Chuncheon, Korea (dkekuldeep@gmail.com, ysmoon@kangwon.ac.kr)

defined data provenance as a process that traces and records the origins and development paths of data, and that it can be applied as a useful tool for evaluating the trustworthiness of data.

We briefly explain the use of data provenance and functionality of these frameworks through a real life example. If we were to conduct an Internet search on a health topic, there would be a lot of sources providing different or similar information on the same topic. However, it would be difficult for a user to identify the only trustworthy source from these numerous sources. Thus, a trustworthiness evaluation framework is required in order calculate the reliability of these sources based on their provenance information and derivation history. The framework also provides trust scores, which allow users to evaluate the trustworthiness of these sources.

In this paper, we provide a detailed view of current data provenance research scenarios with several application areas. In particular, we focus on the frameworks for evaluating the trustworthiness of data and discuss two key frameworks, the Bertino approach and TRUTHFINDER, in detail. Subsequently, we provide a comparative discussion of these two frameworks. Our discussion may help in understanding the data provenance concept in current scenarios. We also present the application domains of the data provenance field in brief. However, the scope of this paper is limited to modern trustworthiness evaluation frameworks only, as Simmhan et al. [8] and Bose and Frew [9] provide detailed information about the domains and history of data provenance.

The rest of the paper is organized as follows: Section 2 explains the motivating domains of frameworks for evaluating the trustworthiness of data. Section 3 presents the structure and functions of two key frameworks in detail. In Section 4, we present a comparison between representative trustworthiness evaluation frameworks. We summarize and conclude our paper in Section 5.

2. Motivating Domains

Data provenance is widely used in various domains, and the trustworthiness obtained by using data provenance techniques improves the quality of results for these domains. In this section, we discuss five motivating domains: 1) social network, 2) location based service (LBS), 3) sensor network, 4) organizational data, and 5) Web service.

Social network: Undoubtedly social media is the most suitable communication medium to broadcast and share certain information to a large number of individuals or groups. The personal use of social media means that it is also being used as a public domain for commercial or political purposes. One report says that 34% of adults use social media to obtain health and wellness information [10]. In general, most of the information shared on social media is contributed by public (untrustworthy) users. Thus, the quality of user-generated content shared on social media may vary from highest to lowest (or, wrong or false) levels, and using it for important and mission critical tasks is not fully acceptable due to its low trustworthiness. Consequently, evaluating the trustworthiness of information shared on social media is getting significant attention these days [11].

Location-based service: LBS allows for the geographical location of an object to be traced. With the emergence of cellular networks and global positioning systems (GPSs), accurate and real-time locations of mobile digital devices can be traced and recorded at a server. Hence, individuals carrying these devices can also be traced. In the field of disease control, tracing the location visited by a disease carrier at a particular time is of significant importance. Similarly, in the field of forensics, investigating the location of a suspect or his/her vehicle at the crime scene is crucial for the interested party. However, the location information generated by mobile devices may contain errors or it can be manipulated for personal benefits. Therefore, obtaining a high level of trustworthiness on location data is crucial to the LBS domain. Consequently, a lot of researchers are currently involved in solving the trustworthiness-related issues of LBS applications and services [12].

Sensor network: Large-scale sensor networks are being deployed in numerous application domains [13], such as environmental monitoring, cyber-physical infrastructure systems, and power grids. The purposes of these sensor networks are often for critical decision-making. In general, sensor data is streamed from multiple sources through intermediate processing nodes for generating the aggregate information. With the limited hardware and software resources, sensor nodes and intermediate nodes may generate inaccurate sensing or aggregating data. Also, sensors are usually operated in an untrusted environment, where a malicious adversary may tamper with the data by introducing additional nodes in the network or by compromising existing ones. Thus, being able to assess the trustworthiness of collected data and making decision makers aware of the trustworthiness of this data has become a crucial issue in sensor network services.

Organizational data: Due to the advancements in cloud computing technologies; the data in local servers have been shifted to a globally distributed system of servers. Now an organization's data and applications that used to be stored in their own machine(s) have gradually become a part of the cloud. Storing organizational data in the cloud has become a recent trend since it is more convenient in terms of cost, performance, and availability. However, it has its own drawbacks: for instance, processing the organizational data is usually distributed among multiple servers controlled by different entities, and thus, it is not possible to determine what part of the computation produced the final results. Hence, considering the confidentiality of the organizational data, which is responsible for critical decision-making, the trustworthiness of this data is worth accessing [14].

Web service: The Internet has become a necessary part of our lives and might be the significant information source for most people. We retrieve all kinds of information from the Web on a daily basis. For example, from shopping online at Amazon or eBay to finding a movie of our choice on Netflix or IMDB, people are very much dependent on the Web. In addition, when we are looking for an answer to a certain question, we usually look at Google or Wikipedia. However, the Web, which provides rich information about a variety of objects, might not be as trustworthy as it should be due to the fact that it is open. If an information source copies from other unreliable sources, outdated data can be provided or the information can be intentionally tampered with for certain benefits. Considering the possible financial benefits and other fraud issues involved, it is of utmost necessity to have a process to evaluate the trustworthiness of the information on the Web.

3. Trustworthiness Evaluation Frameworks

A considerable amount of research work has been devoted to studying trustworthiness evaluation techniques in recent years. In Section 3.1, we first briefly explain the history of the research work in. Second, in Sections 3.2 and 3.3 we summarize the two key frameworks proposed for evaluating the trustworthiness of data. Third, in Section 3.4, we discuss some of the key features of other prominent frameworks.

3.1 History of Trustworthiness Frameworks

Trustworthiness evaluation frameworks have been basically built around integrity, quality, reputation systems, and the provenance of data. We explain the previous efforts categorizing these four aspects. First, many research works have focused on the integrity of data. Biba [15] presented the first approach, which addressed the hierarchical lattice of integrity levels for information systems, where the integrity levels could be determined by blocking the flow of information from low-integrity objects to high-integrity subjects. However, Biba's approach has a critical limitation in that it does not provide any criteria to determine the integrity levels. Clark and Wilson [16] proposed another approach that exploited the following two key concepts: well-formed transactions and the separation of duties. The former manipulates data in trusted ways to preserve the consistency of the data, and the latter mandates separating all operations into several subparts and executing each subpart by a different subject. Bellare and Rogaway [17] and Goldreich [18] have contributed to the advancement of data integrity techniques. In particular, they have worked to improve digital signature systems (DSSs) ability to achieve high integrity levels.

Second, data quality has been used as an important feature in many real applications. High-quality data increases the probability for organizations to make better decisions. Juran [19] addressed the fact that data is considered to be of high quality "if it is fit for their intended uses in operations, planning and decision-making." Organizations have recognized the importance of data quality and started spending huge amounts of money to improve it. For example, the United States government enacted the Data Quality Act in 2002 [20]. There are several theoretical approaches and tools, such as record linkage and business rules [21,22], which have been introduced to evaluate and improve the quality of data.

Third, in the field of recommendation systems, reputation is a key approach towards securing and the entities containing less than a desirable score might be filtered out from the system. Several approaches have been developed to encourage adherence in the field of e-commerce [23]. Popular e-commerce sites like Amazon and eBay use their reputation systems to detect fraud-related activities. eBay runs one of the first reputation systems, which gathers comments from both buyers and sellers of each transaction [24]. The Web-based community of Advogato uses a reputation system for filtering spam [25].

Fourth, recently, the data provenance or data lineage has been used to assess the trustworthiness of data. The basic notion behind data provenance is the information that helps determine the derivation history of a data product, starting from its original sources [8]. Several frameworks have been developed for computing the confidence levels of query results by evaluating the provenance of these results [26-29]. Buneman et al. [5] have proposed the "why and where" characterization of data provenance by defining "why" a piece of data is in existence and "where" the data comes from. In addition, Ragan et al. [30] have proposed a framework for characterizing provenance information in the fields of visual analysis. Widom [31] have proposed the Trio system that supports information management in regards to the accuracy of data and provenance. They incorporated both accuracy and provenance as an integrated part of data management and query processing. Vijayakkumar and Plale [32] introduced the provenance collection framework for an event processing system. Sarma et al. [33] developed an architecture based on a decoupled strategy to compute provenance and confidence in probabilistic databases. Yin et al. [34] and Gupta et al. [35] proposed a trustworthy framework called TRUTHFINDER that focuses on the trustworthiness of Web services. Bertino and his associates

[12,13,36] developed and further improved the frameworks that evaluate the trustworthiness of data based on data provenance. Data provenance has also been investigated in online health care analytics for a biomedical data stream system in IBM's Century [37,38]. Malaverri et al. [39] introduced a provenance-based approach for evaluating data managed by E-Science applications. While data provenance is used as a tool to evaluate the trustworthiness of data, Cheah and Plale [40] proposed a framework for assessing the quality of the provenance data itself.

In the next two subsections we explain in detail: 1) Bertino's comprehensive approach and 2) TRUTHFINDER. We have limited the scope of this paper to the comparison to these two approaches only due to the following reasons: first, these two approaches can be considered as the basic frameworks for the trustworthiness of data and data provenance; so studying these frameworks gives a basic idea of this area. Second, this comparison leads us to the current issues and challenges of the areas of provenance and trustworthiness.

3.2 Bertino's Comprehensive Approach

Bertino and his associates [12,41,42] proposed a novel cyclic framework of accessing the trustworthiness of data based on its provenance. Fig. 1 shows the cyclic framework that computes the trustworthiness of data items, source nodes, and intermediate nodes in a cyclical manner. That is, it computes the trust scores of data items from source or intermediate nodes and the trust scores of source or intermediate nodes from data items. The framework takes into account four factors: data similarity, path similarity, data conflict, and data deduction.

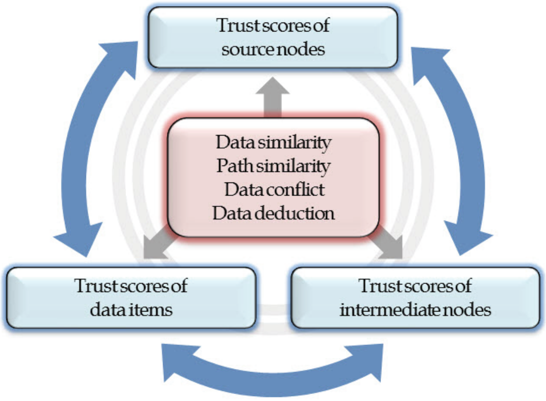
Trust scores of source nodes
Data similarity
Path similarity
Data conflict
Data deduction
Trust scores of data items
Trust scores of intermediate nodes

Fig. 1. Bertino's comprehensive approach with the cyclic framework.

Data similarity refers to the process of determining how similar two data items are. It is evident that if two data items are similar, they are supportive of each other. Bertino and his associates [36,41] proposed a clustering algorithm to examine the similarity of data items. The algorithm works as follows: it defines the first few items as separate clusters and regards those items as the representatives of those clusters. Then, for every next item, the algorithm compares it with representatives of existing clusters—that is, if the distance between the current representative and an item is within the given threshold, the item is added to the current cluster. This process is repeated until all items are clustered. Finally, each cluster keeps data items whose distances from the representative are within the given threshold. In this scheme, the trust score of an item will be high if the cluster of containing the item has many other similar items. The algorithm is further extended by combining three commonly used types of attributes: numerical, categorical, and string. Readers are referred to [36,41,42] for more details.

Path similarity refers to the process of determining whether two data items followed the same or similar data generation path. We will now elaborate on how to evaluate the similarity of paths and how it affects the overall trust score. To evaluate this, all intermediate nodes from source to destination nodes need to be considered. Bertino and Lim [36] described a path as a list of identifiers, which represent data source, destination, and all intermediate nodes that fall in the path. To compute the similarity of two paths, their lists of identifiers are compared. For example, if we are comparing paths P_1 and P_2, then we compare the list of identifiers for P_1 = "*Busterminal* → *policestation1* → *policestation2*" and for P_2 = "*Railwaystation* → *policestation1* → *policestation2*." As we can see, the difference between P_1 and P_2 is only the first identifier, and using the difference they calculated the path similarity between them. Likewise, they computed the path similarity among all given paths. In their framework, the path similarity results in a negative impact on the overall trust score. That is, a data item provided by different sources is most likely to be true if it comes from different independent sources and through different paths. Overall, path similarity helps in evaluating the provenance independence of two or more data items.

Data conflict refers to the situation in which the data provided by two or more source providers have a conflict in regards to information or a description for the same event or entity. It is evident that a data conflict has a negative impact on the overall trust score. To understand data conflict, consider the simple example where two or more sources report different locations for the same person at a particular time. The common reasons behind data conflict are typos, malicious source providers, and misleading knowledge generation by intermediate nodes. Since data conflict deeply depends on application domains, Bertino et al. [36,41] permit users to define their own data conflict functions according to what is suitable for their application domain. To evaluate a conflict among two or more data items, the meaning of the conflict, called *prior knowledge*, is defined first. Then, based on prior knowledge, the final results are obtained. Two data items are considered to be conflicting with each other if they fail to satisfy the conditions established by the prior knowledge.

Data deduction is the process of evaluating the impact of source and intermediate nodes on the delivered data item. The trustworthiness of a data item greatly depends on the source information that generates it and the intermediate nodes that handle it. Therefore, if the source information and intermediate nodes are highly trusted, the delivered data will also be highly trusted. To evaluate the impact of source information and intermediate nodes, Bertino et al. [41] proposed several actions. These actions are application-specific and may vary for different applications. For example, they have described two typical actions of "PASS" for simply passing the input data to successive nodes and "INFER" for generating new knowledge based on the input data and some local knowledge. Since PASS does not change the input data at all, the trustworthiness of delivered data remains the same, whereas, INFER changes the input to the output, and, thus, it affects the trustworthiness of delivered data. Bertino et al. [41] also proposed a weight function to compute the trust score of delivered data.

Finally, the four aspects mentioned above are combined to obtain an overall trust score for data items with the help of iterative computations. Readers are referred to [41,42] more details on the computation model. Starting from [41], Bertino and his associates [12,13,36] have made several advancements by improving the cost of computation, security, and the performance of models by combining several techniques and considering specific application domains.

3.3 TRUTHFINDER: An Object-Based Approach

Yin et al. [34] proposed a framework, TRUTHFINDER, to infer the trustworthiness of websites and the information that those websites provide. According to the survey conducted by Princeton Survey Research [43], only 54% of Internet users trust online news sites, only 26% trust e-commerce sites, and only 12% trust blogs. Based on these observations, TRUTHFINDER deals with the conflict of information provided by several websites for a particular subject. They use the term "facts" for data and "objects" for areas, where the data provider provides facts for objects.

Fig. 2 shows an example of the input structure in TRUTHFINDER. In Fig. 2, inputs to TRUTHFINDER are facts f_1 to f_5 provided by different providers P_1 to P_4 for objects o_1 and o_2. The goal of TRUTHFINDER is to deal with the conflicting facts provided by different websites. To represent the asymmetric relationship between facts, they introduced the concept of *implication* between facts. For example, implication from fact f_1 to fact f_2, ($f_1 \rightarrow f_2$), defines how much f_2's confidence increases or decreases due to f_1's confidence. In other words, ($f_1 \rightarrow f_2$) defines f_1's influence on f_2's confidence. TRUTHFINDER uses an iterative method for inferring the trustworthiness of a website and fact confidence from each other. Yin et al. [34] proposed a heuristic-based computational model that has explained how to infer the reliability of websites and confidence about facts separately and how to calculate both measures through matrix operations.

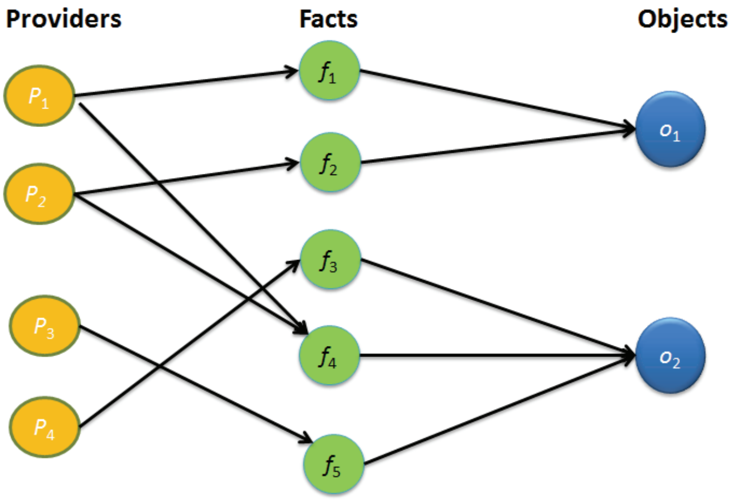

Fig. 2. An example of inputs to TRUTHFINDER.

Gupta et al. [35] extended TRUTHFINDER to obtain even stronger trustworthiness and they used cluster-based methods for evaluating trustworthiness. The idea behind this concept is that every information provider has its own area where it provides good quality data only for its own area. As such, a single data provider may not be a trustworthy source for all kinds of information. Based on intuition, they improved the accuracy of data and data providers by presenting the object-based trustworthiness rather than the global trustworthiness of an information provider provided by the naïve TRUTHFINDER [34].

We will now explain Gupta et al.'s model [35] in more detail. Fig. 3 explains the process of clustering objects, evaluating trustworthiness, and why cluster-based ranking of providers is useful. Two data providers P_1 and P_2 provide facts for five objects o_1 to o_5. P_1 provides good facts for objects o$_1$ and o_2 (shown in solid lines), and bad facts for o_3 and o_4, (shown in dotted lines), whereas, P_2 provides good facts for objects o_3, o_4, o_5 and bad facts for o_1 and o_2. Since the number of good facts provided by P_2 is larger than that of P_1, P_2 is higher than P_1 in the global rankings. But, looking at the trust profile of objects in the provider space, we can observe that objects o_1 and o_2 have a similar profile, while o_3, o_4, and o_5 have a similar profile. So, they can be grouped into two clusters C_1 and C_2. Note that P_1 would be ranked higher for C_1, and P_2 would be ranked higher for C_2. Likewise, interesting results can be obtained by clustering objects in the provider trustworthiness space. Gupta et al.'s algorithm performs two iterative steps. First, it performs the clustering step of bringing similar objects together. Second, it performs the trust analysis step to compute better cluster-conditional trust rankings and better fact confidence values. Both of the iterative steps are repeated until the changes in cluster formation and trustworthiness of providers are negligible. In addition, they use two metrics, accuracy and compactness, for measuring the accuracy of facts and the quality of clusters, respectively. In summary, Gupta et al.'s model provides a cluster-based approach for obtaining the object-based local trustworthiness of websites.

3.4 Some Other Prominent Frameworks

Apart from the models discussed earlier, there are some other frameworks proposed in the field of data provenance and data trustworthiness. In this section, we give a brief introduction of some of these prominent frameworks. We categorize them into two groups: 1) frameworks that are based on streaming data applications, and 2) other prominent frameworks for many different application areas.

First, Vijayakumar and Plale [32], proposed a provenance collection framework for collecting provenance with low overhead in stream filtering applications. They addressed the challenges of the data streaming environment and proposed data collection models with low overhead for handling rapid streams. Additionally, they discussed a prototype that was implemented using the Calder stream processing system. Lim et al. [13] proposed a cyclic method to compute the trustworthiness of sensor data. Their model presents sensor data in the graphical form and computes the trustworthiness of sensor data by exploring the *normal distribution* property of values and provenance similarities. It first computes the current and intermediate trust scores and eventually converts them into final trust scores by using an iterative framework. Advancing Lim's et al.'s work, Batini et al. [44] presented a lightweight provenance encoding and decoding scheme by exploiting *bloom filters*. Considering the specific security challenges of sensor networks, they have proposed network, data, provenance, and threat models. They also presented a security analysis of schema, which defines confidentiality, integrity, and the freshness of streaming data.

Second, Dong et al. [45] introduced an interesting model to determine the copying relationships between Web sources by exploring their update history. Their model measures the quality of the Web data by using coverage, exactness, and freshness, as defined in the paper. To discover copying relationships between the sources and lifespan of each object, they used the hidden Markov model

(HMM) and Bayesian model, respectively. Pasternack and Roth [46] tried to incorporate a user's prior knowledge into the fact-finding system. Their framework uses general reasoning and already known facts as the first logic and translates it into the tractable linear program with three fact-finding algorithms. Moturu and Liu [11] proposed a social related framework of defining the specific problems of social media applications and the sociological notions of trust. Their framework follows iterative steps: 1) it identifies the features of the given information and categorizes them, and 2) it quantifies the contents of social data based on its own scoring models.

4. Comparison of Trustworthiness Frameworks

In this section we present a comparative discussion between the Bertino's comprehensive approach (*Bertino approach*) and TRUTHFINDER. Research works, including [12,13,36,42], are based on the Bertino approach in [41]. Similarly, [34] and [45] are based on TRUTHFINDER in [33]. Thus, our comparison not only includes these two frameworks, but also partially includes other interrelated frameworks. Since both frameworks work with different data types and domains, it is not possible to directly compare them. Thus, we compared them on the basis of several factors. We performed the comparisons in the following three categories: 1) domain comparisons, 2) accuracy comparisons, and 3) efficiency comparisons.

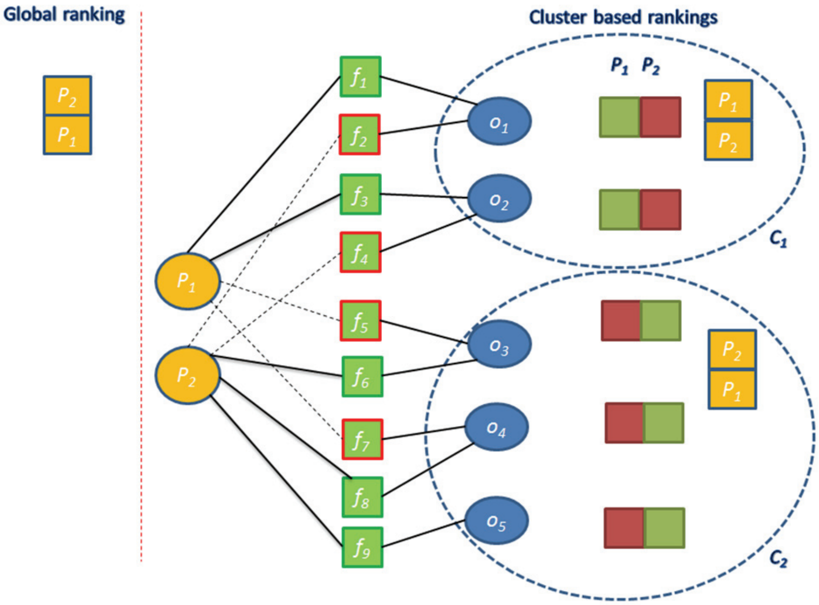

Fig. 3. Gupta et al.'s [35] cluster-based approach for TRUTHFINDER.

4.1 Domain Comparisons

In this section, we elaborate on the different application areas of data trustworthiness evaluation frameworks. Table 1 summarizes what applications were considered and which frameworks were used for those applications. First, the Bertino approach, as shown in Table 1, has been widely used in many domains. It evaluates the trustworthiness of sensor data items and sensor nodes in [13,44]. In [12], it computes the trustworthiness of location information provided by various sources. In [44], it evaluates the trustworthiness of organizational data.

Second, TRUTHFINDER's domain is limited to Web services only. That is, it evaluates the facts claimed by various Internet sites and the credibility of those websites. However, the framework can work on various Web applications, such as verification of book authors or movie runtime. To our knowledge, this framework has not been explored for other data types.

Table 1. Domain areas of trustworthiness evaluation frameworks

Framework	Domain	Ref.	Purpose
Bertino approach	Sensor networks	[13,44]	Assessment of sensor data
	Location based service	[12]	Assessment of location based data
	Organizational data	[44]	Assessment of organizational data
TRUTHFINDER	Web services	[35,45]	Assessment of Web data

4.2 Accuracy Comparisons

As we explained in Section 4.1, the Bertino approach and TRUTHFINDER have different domains and different data types. Thus, definitions of accuracy measures are also different for each of the frameworks. We explain the accuracy of trustworthiness from a different viewpoint for each of the frameworks. Table 2 summarizes the accuracy definitions in the frameworks and their use in their. First, the Bertino approach produces global trustworthiness for the data providers. Global trustworthiness can be referred to as the overall trustworthiness of a data provider, irrespective of what kind of data it provides for a particular object. The trust score of a particular data provider reflects the trustworthiness of the overall data it provides. This global trustworthiness is used for the applications where provider-specific trustworthiness is required. For example, we need to verify a news site's overall trustworthiness without any specific requirements on the trustworthiness of sports news or business news.

Table 2. Definitions and uses of accuracy for trustworthiness evaluation frameworks

Framework	Accuracy definition	Use of accuracy
Bertino approach [41]	Global trustworthiness	Provider-specific trustworthiness
TRUTHFINDER [35]	Cluster-based local trustworthiness	Object-specific trustworthiness

Second, TRUTHFINDER produces the local trustworthiness of data providers. That is, a data provider receives a trust score based on the particular object for which it provides data. This object-specific local trustworthiness brings additional accuracy considering the fact that a data provider can provide good or bad data for different objects. This object-specific trustworthiness can be used for the application where we need to verify a source based on its object. For example, we need object-specific trustworthiness from the same news site. Now, we need to evaluate what the trust score is for sports news, business news, and weather news. The key difference here is that with object-specific trustworthiness we can classify areas of this particular website.

4.3 Efficiency Comparisons

In this section we compare both frameworks based on their efficiency regarding several aspects. We categorize the comparisons in two subcategories: 1) the scalability of the framework and 2) the number of attributes used. First, scalability refers to exploring frameworks by changing different aspects in the experimental phase and calculating its effect on the running time. Table 3 shows the efficiency comparison, which is based on the scalability of frameworks. As shown in Table 3, the Bertino approach produces more efficiency by showing more flexibility towards the various aspects discussed that are discussed below.

The Bertino approach examines its scalability for 1) the size of datasets, 2) the number of seeds, and 3) the path lengths of datasets. As shown in Table 3, with the increase of all of the above-mentioned aspects, the running time increases gradually, but it is still practically applicable. For example, if we increase the size of datasets to 50K, the Bertino approach takes less than two minutes to compute their trustworthiness, which is practical for off-line applications. On the other hand, TRUTHFINDER is only examined with respect to the number of facts used. Moreover, if we increase the number of facts to 50K, the running time increases to 118 times, which is more costly.

Table 3. Efficiency comparison based on scalability of frameworks

Framework	Scalability of frameworks	Impact of scalability on the running time	Type of data relationships
Bertino approach [41]	Size of datasets Number of seeds Path lengths of datasets	Increase gradually First decrease, then increase Increase gradually	Symmetric
TRUTHFINDER [34,35]	Number of facts	Increase largely (118 times)	Symmetric and asymmetric

Second, the Bertino approach is also efficient with the number of attributes used. For example, Dai et al. [42] used a dataset with seven attributes at a time, whereas, TRUTHFINDER works with only one attribute at a time. For example, only the names of book authors or the runtime of movies are used at a time. However, TRUTHFINDER deals with both the symmetric and asymmetric relationships of data, while the Bertino approach only works with the symmetric relationships of data. In addition, TRUTHFINDER uses matrix operations to calculate final trust scores, which is easy and quick to implement and brings additional efficiency in the performance of the framework.

5. Conclusions

Since a lot of huge data that is used in various applications that are for critical decision making is being collected from a large number of distributed sources, accessing the trustworthiness of these sources and their paths becomes an essential research issue. In this paper, we discussed the improvement in the field of data provenance by focusing on the data trustworthiness evaluation frameworks. We first provided a brief introduction about the history of data provenance frameworks.

We then presented a detailed and comparative discussion of two key frameworks: the Bertino approach and TRUTHFINDER. We also presented a comprehensive overview of current data trustworthiness evaluation frameworks and showed that data provenance is still an exploratory area with several open research issues.

We will now present some open research issues that are either not addressed or that have some scope for improvement. The first one is data privacy. For example, when developing an approach that focuses on the privacy of data that travels through various intermediate nodes, we may define certain constraints for intermediate sources based on some prior knowledge of "what amount of data can be seen or modified by a particular source." Moreover, we may determine the specific contributions made by all intermediate sources. The second issue is a common approach, which can simultaneously work on both symmetric and asymmetric relationships of data. In addition, that common approach should work for all type of domains concurrently. The third one is cost models. Based on the time and space complexity discussed earlier in this paper, cost effective models should be investigated.

Acknowledgement

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and future Planning (NRF-2014R1A2A2A01002548).

References

[1] D. P. Lanter, "Design of a lineage-based meta-data base for GIS," *Cartography and Geographic Information System*, vol. 18, no. 4, pp. 255-261, 1990.

[2] T. R. Smith, J. Su, D. Agrawal, and A. El Abbadi, "Database and modeling systems for the earth sciences," *IEEE Special Issue on Databases,* vol. 16, no. 1, pp. 33-37, 1993.

[3] P. D. Eagan and S. J. Ventura, "Enhancing value of environmental data: data lineage reporting," *Journal of Environmental Engineering*, vol. 119, no. 1, pp. 5-16, 1993.

[4] A. Woodruff and M. Stonebraker, "Supporting fine-grained data linage in database visualization environment," in *Proceedings of the 13th Conference on Data Engineering*, Birmingham, UK, 1997, pp. 91-102.

[5] P. Buneman, S. Khanna, and W. C. Tan, "Why and where: a characterization of data provenance," in *Proceedings of the 8th International Conference on Database Theory (ICDT),* London, UK, 2001, pp. 316-330.

[6] P. S. Wadhwa and P. Kamalapur, "Customized metadata solution for a data warehouse: a success story," White paper, Wipro Technologies, Bangalore, India, 2003.

[7] L. Moreau and P. Missier, "PROV–DM: the PROV data model," World Wide Web Consortium, Technical Report, 2013.

[8] Y. L. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance in e-science," *ACM SIGMOD Record*, vol. 34, no. 3, pp. 31-36, 2005.

[9] R. Bose and J. Frew, "Lineage retrieval for scientific data processing: a survey," *ACM Computing Surveys*, vol. 37, no. 1, pp. 1-28, 2005.

[10] N. Elkin, "How america searches: health and wellness," Opinion Research Corporation, Survey Report, 2008.

[11] H. T. Moturu and S. Liu, "Quantifying the trustworthiness of social media content," *Distributed and Parallel Databases*, vol. 29, no. 3, pp. 239-260, 2011.

[12] C. Dai, H. S. Lim, E. Bertino, and Y. S. Moon, "Assessing the trustworthiness of location data based on provenance" in *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Seattle, WA, 2009, pp. 276-285.

[13] H. S. Lim, Y. S. Moon, and E. Bertino, "Provenance-based trustworthiness assessment in sensor networks," in *Proceedings of the 7th International Workshop on Data Management for Sensor Networks*, Singapore, 2010, pp. 2-7.

[14] M. Kuehnhausen, V. S. Frost and, G. J. Minden, "Framework for assessing the trustworthiness of cloud resources," *Proceedings of the IEEE International Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, New Orleans, LA, 2012, pp. 142-145.

[15] K. J. Biba, "Integrity considerations for secure computer systems," MITRE Corp., Bedford, MA, Report No. TR-3153, 1977.

[16] D. Clark and D. Wilson, "A comparison of commercial and military computer security policies," in *Proceedings of the IEEE Symposium on Security and Privacy*, Oakland, CA, 1987.

[17] M. Bellare and P. Rogaway, "The exact security of digital signals: how to sign with RSA and Rabin," in *Proceedings of the International Conference on Theory and Application of Cryptographic Techniques*, Saragossa, Spain, 1996, pp. 399-416.

[18] O. Goldreich, "The foundation of modern cryptography," in *Modern Cryptography, Probabilistic Proofs and Pseudorandomness*. Heidelberg: Springer, 1999, pp. 1-37.

[19] J. M. Juran, Juran on Leadership for Quality: An Executive Handbook. New York, NY: Free Press, 1989.

[20] The Office of Management and Budget, "Federal collection of information," [Online]. Available: http://www.whitehouse.gov/omb/inforeg_infocoll.

[21] C. Batini and M. Scannapieco, *Data Quality, Concepts, Methodologies and Techniques*. Heidelberg: Springer, 2006.

[22] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Computing Surveys*, vol. 41, no. 3, pp. 16-52, 2009.

[23] P. Resnick, R. Zeckhauser, E. Friedmen and K. Kuwabara, "Reputation systems," *Communications of the ACM*, vol. 43, no. 12, pp. 45-48, 2000.

[24] P. Resnick and R. Zeckhauser, "Trust among strangers in internet transactions: empirical analysis of eBay's reputation system," in *The Economics of the Internet and E-commerce.* Amsterdam: Elsevier Science, 2002, pp. 127-157.

[25] R. Levien, "Attack Resistant Trust Metrics," Ph.D. dissertation, University of California, Berkeley, CA, 2004.

[26] S. Abitebourl, P. Kanellakis, and G. Grahne, "On the representation and querying of sets of possible words," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, San Francisco, CA, 1987, pp. 34-48.

[27] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. San Francisco, CA: W. H. Freeman, 1979.

[28] D. Barbara, H. Gracia-Molina, and D. Porter, "The management of probabilistic data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 4, no. 5, pp. 487-502, 1992.

[29] N. Fuhr, "A probabilistic framework for vague queries and imprecise information in databases," in *Proceedings of the 16th International Conference on Very Large Databases*, Brisbane, Australia, 1997, pp. 696-707.

[30] E. D. Ragan, A. Endert, J. Sanval, and J. Chen, "Characterizing provenance in visualization and data analysis: an organizational framework of provenance types and purposes," *IEEE Transactions on Visualization & Computer Graphics*, vol. 22, no. 1, pp. 31-40, 2016.

[31] J. Widom, "Trio: a system for integrated management of data, accuracy, and lineage," in *Proceedings of the 2nd International Conference on Innovative Data Systems Research,* Asilomar, CA, 2005, pp. 262-276.

[32] N. N. Vijayakumar and B. Plale, "Towards low overhead provenance tracking in near real time stream filtering," in *Proceedings of the International Provenance and Annotation Workshop*, Chicago, IL, 2006, pp. 46-54.

[33] A.D. Sarma, M. Theobald, and J. Widom, "Exploiting lineage for confidence computation in uncertain and probabilistic databases," in *Proceedings of the 14th International Conference on Data Engineering*, Cancun, Mexico, 2008, pp. 1023-1032.

[34] X. Yin, J. Han, and P. S. Yu, "Truth discovery with multiple conflicting information providers on the web," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 6, pp. 796-808, 2008.

[35] M. Gupta, Y. Sun, and J. Han, "Trust analysis with clustering," in *Proceedings of the 20th International Conference Companion on World Wide Web*, Hyderabad, India, 2011, pp. 53-54.

[36] E. Bertino and H. S. Lim, "Assuring data trustworthiness-concepts and research challenges," in *Secure Data Management*. Heidelberg: Springer, 2010, pp. 1-12.

[37] M. Blount, "Century: automated aspects of patient care," in *Proceedings of the 13th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications*, Daegu, Korea, 2007, pp. 504-509.

[38] A. Misra, M. Blount, A. Kementsietsidis, D. M. Sow, and M. Wang, "Advances and challenges for scalable provenance in stream processing systems," in *Proceedings of the 2nd International Provenance and Annotation Workshop*, Salt lake City, UT, 2008, pp. 253-265.

[39] J. E. G. Malaverri, A. Santanche, and C. B. Medeiros, "A provenance based approach to evaluate data quality in eScience," *International Journal of Metadata, Semantics and Ontologies*, vol. 9, no. 1, pp. 15-28, 2014.

[40] Y. W. Cheah and B. Plale, "Provenance quality assessment methodology and framework," *Journal of Data and Information Quality*, vol. 5, no. 3, article no. 9, 2015.

[41] E. Bertino, C. Dai, H. S. Lim, and D. Lin, "High-assurance integrity techniques for databases," in *Proceedings of the 25th British National Conference on Databases*, Cardiff, UK, 2008, pp. 244-256.

[42] C. Dai, D. Lin, E. Bertino, and M. Kantarcioglu, "An approach to evaluate data trustworthiness based on data provenance," in *Secure Data Management*. Heidelberg: Springer, 2008, pp. 82-98.

[43] Princeton Survey Research Associates International, *Leap of Faith: Using the Internet Despite the Dangers*. Yonkers, NY: Consumer Reports WebWatch, 2005.

[44] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Computing Surveys*, vol. 41, no. 3, pp. 16-52, 2009.

[45] X. L. Dong, L. Berti-Equille, and D. Srivastava, "Truth discovery and copying detection in a dynamic world," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 562-673, 2009.

[46] J. Pasternack and D. Roth, "Knowing what to believe (when you already know something)," in *Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China, 2010, pp. 877-885.

Kuldeep Gurjar

He received B.S. (2005, Stany Mem. Collage) and M.S. degrees (2010) in Computer Science, from Dept. of Computer Science and Information Technology, University of Rajasthan, Jaipur. From the year 2010 to 2011, he worked for a Website development company (Octal info. Solutions). Since March 2012, he is with the Dept. of Computer Science and Engineering from Kangwon National University as a Ph.D. candidate. His research interests are data mining, data provenance, data trustworthiness.

Yang-Sae Moon

He received B.S. (1991), M.S. (1993), and Ph.D. (2001) degrees in Computer Science from Korea Advanced Institute of Science and Technology (KAIST). From 1993 to 1997, he was a research engineer in Hyundai Syscomm, Inc., where he participated in developing 2G and 3G mobile communication systems. From 2002 to 2005, he was a technical director in Infravalley, Inc., where he participated in planning, designing, and developing CDMA and W-CDMA mobile network services and systems. He is currently a professor of computer science department at Kangwon National University. He was a visiting scholar at Purdue University in 2008 to 2009. His research interests include data mining, knowledge discovery, storage systems, access methods, multimedia information retrieval, big data analysis, mobile communication systems, and network communication systems. He is a member of the IEEE, and a member of the ACM.