Hindawi
Scientific Programming
Volume 2017, Article ID 4504589, 11 pages
https://doi.org/10.1155/2017/4504589

Research Article

Information-Balance-Aware Approximated Summarization of Data Provenance

Jisheng Pei and Xiaojun Ye

Department of Computer Science and Technology, Tsinghua University, Beijing, China

Correspondence should be addressed to Jisheng Pei; pjs07@mails.tsinghua.edu.cn

Received 28 February 2017; Revised 21 April 2017; Accepted 2 May 2017; Published 12 September 2017

Academic Editor: Chi-Hung Chi

Extracting useful knowledge from data provenance information has been challenging because provenance information is often overwhelmingly enormous for users to understand. Recently, it has been proposed that we may summarize data provenance items by grouping semantically related provenance annotations so as to achieve concise provenance representation. Users may provide their intended use of the provenance data in terms of provisioning, and the quality of provenance summarization could be optimized for smaller size and closer distance between the provisioning results derived from the summarization and those from the original provenance. However, apart from the intended provisioning use, we notice that more dedicated and diverse user requirements can be expressed and considered in the summarization process by assigning importance weights to provenance elements. Moreover, we introduce information balance index (IBI), an entropy based measurement, to dynamically evaluate the amount of information retained by the summary to check how it suits user requirements. An alternative provenance summarization algorithm that supports manipulation of information balance is presented. Case studies and experiments show that, in summarization process, information balance can be effectively steered towards user-defined goals and requirement-driven variants of the provenance summarizations can be achieved to support a series of interesting scenarios.

1. Introduction

With the development of data-generating devices and services such as intelligent mobile phones, tablets, sensor networks, and large-scale social network sites, it has become a common and important practice to collect, store, and aggregate large amount of data from multiple sources to generate useful information for users. Real-world examples include scientific workflow systems and crowd-sourcing applications such as open-source encyclopedia and crowd rating websites. The results produced by these applications are often used to help users make various kinds of decisions in both life and business. Therefore, as the stakeholders in these scenarios are desiring to get more information about how the application comes up with its results and how different data are contributing to them, questions such as how and why data were derived have often been raised [1, 2]. For example, how are different group of users (e.g., younger/senior users, male/female users, and users from different expertise areas) contributing the results? Furthermore, in order to get a feeling of the derivation process in a hands-on manner, users may also want to try discarding some of the contributions to see their original influence on the results. For example, users may discard some parts they consider to be scams or irrelevant, or they may discard some parts until only what they are interested in are among the inputs.

To answer questions like these, we may refer to the provenance of the data derivation process, as it records the context of data input and how the information was derived. For example, movie rating websites such as IMDB usually present estimated movie ratings by aggregating ratings submitted by a large number of users, whose diverse demographic characteristics, preferences, and previous reviews are all recorded as part of the provenance. We may use such information to analyze why the data derivation process has been executed in certain way or what are the influences applied onto the final result by different groups of users. To achieve this, provenance semiring [2] has been proposed and used to

support both the storage and the analytical manipulations to analyze the influence of various data elements. For example, based on provenance semiring, we can support provisioning of the result according to hypothetical insertions, removals, or modifications to the input.

However, listing all recorded provenance in full all at once is not the proper way for users to understand the messages contained by the provenance, as the size and the complexity of detailed provenance information could be overwhelming. Approximated summarization of data provenance has therefore been proposed in [3] to reduce the provenance size by grouping multiple "similar" data provenance annotations as a single annotation through mapping. Intuitively, as annotations are being merged to form new feasible annotations, each annotation would have to symbolize more and more annotations from the original provenance. Thus, the information in each provenance annotations will become ambiguous. Although this would lead to a more concise and high-level representation, it might also cause possible losses in information content or ambiguity, since the grouped annotation no longer makes distinctions between the original annotations. Therefore, we need to find a way to retain useful information for the users in the summaries as much as possible.

Previously, semantic constraints that keep the grouped annotations make sense semantically are imposed such that only "similar" or "related" annotations may be grouped together. However, these are relatively loose constraints (e.g., annotations sharing at least one attribute in common may be grouped together) that are meant to keep the grouped annotations make sense, rather than to retain information that is useful for the users. On the other hand, to achieve a balance between provisioning accuracy and the size of provenance summarization, it has been proposed in [3] that the provisioning results derived from the provenance summary should be retained as much as possible compared to the one derived from the original provenance. Based on this requirement, the current provenance summarization algorithm in [3] searches for an approximated optimal provenance summary, by grouping semantically feasible annotations that will lead to maximum size reduction and minimal distance increment (combined with some weights), one pair at a time in a stepwise manner. However, in this constraint, only the deviation in provisioning results, which is but one of the consequences of the information loss, is considered. But again, the loss of information caused by annotation grouping has not yet been evaluated or dealt with.

Actually, in general data grouping tasks, where raw data are grouped in classes to cope with complexity, balancing the information amount and homogeneity of the grouped classes has for long been recognized as one of the key requirements by users [4]. We believe that this should also be the case for provenance annotation grouping. Users may want to have the freedom to express what kind of information they want to include (or exclude) in the summary. In other words, when choosing from different options of annotation groupings, the influence of size reduction and distance increment should be considered in the context of the information balance. Consequently, among some possible provenance summaries of similar size and distance, the one that preserves more "useful" provenance items for the users should be more favorable than the others. It would be of interest to investigate how we could take control of the loss of information content during provenance summarization and to see how it might affect the quality outcome of the summarization.

Contribution in this article is as follows: we present a novel algorithm for provenance summarization that adopts information balance as an additional factor for provenance summarization quality control. The new summarization process not only takes semantic constraint and provisioning distance into consideration but also uses information balance to dynamically assess the "usefulness" of the summary contents for users. We define a dynamic entropy based heuristic function that keeps the balance between information amount loss and homogeneity according to user requirement inputs as weights assigned to each provenance tuple. Case studies and comparative experiments on real-world datasets are conducted to show that, by controlling information balance during provenance summarization, our approach can allow provenance summarization results meet customized user requirements while achieving comparable or even better size-distance performance with the previous works.

2. Preliminaries

We first recall some background of provenance information management, semiring provenance model, and the summarization of provenance from [1–3] before discussing our motivation and proposal.

2.1. Collection and Storage of Provenance Information. Provenance information can be collected in various forms including scientific workflow logs, data access logs, file system records, and relational query logs. In other words, the raw form of provenance information can be highly heterogeneous (e.g., text files, tables, relational graphs, and time series) and both structured and unstructured information can be involved. In order to cope with the complexities and heterogeneity of these captured provenance information, we need to organize them in structural format. In the experiment part of this article, we consider the case of movie rating websites and adopt the widely used MovieLens dataset, in which users ratings from multiple sources are collected with automatic crawlers along with demographic information of the users. As Figure 1 shows, the raw provenance information collected is stored and managed in a relational database management system as a rating table, a movie information table, and a user information table. These three tables together provide information on which users rated which movies and the ratings they assign. We match the relations of the rows in these tables to the notations in semiring provenance model (e.g., Example 1) so as to reflect how each individual rating influences the eventual aggregation analysis result of movie ratings. To achieve this, we implement semiring algebraic structures (e.g., Figure 1) in our programs by organizing data items from these tables as different attributes of a semiring element class and indicating their roles in the provenance semiring expression (e.g., tuple annotation, value, and tensor).

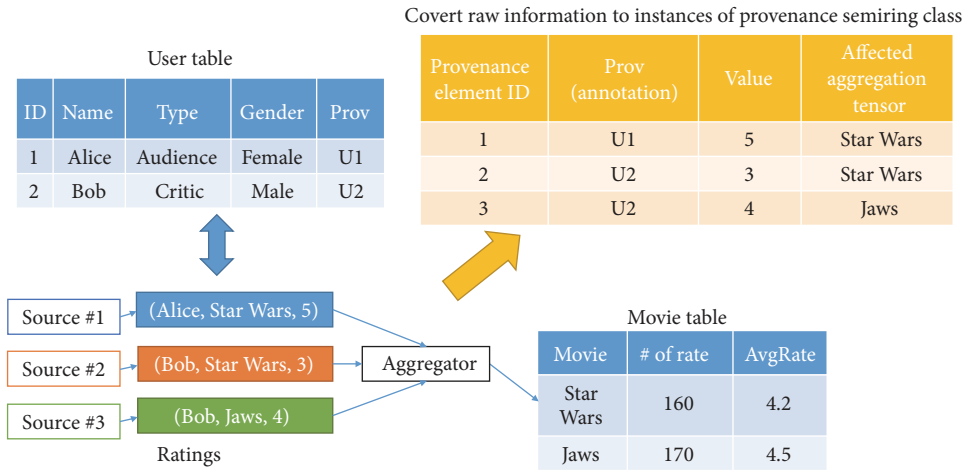

FIGURE 1: Raw provenance information collected from applications is converted to provenance semiring structures.

Thus, a set of instances of such semiring element class may be used to represent and store a provenance expression. In next subsection, we will introduce these roles and symbols as well as their meaning.

2.2. Semiring Provenance Model. Our study focuses on the semiring provenance model [1, 2, 4], but the same result can also be easily extended to other types of models [5, 6]. Semiring provenance model records provenance information with a finite set X of provenance annotations, which can be understood as the basic data items or elements. For example, an annotation may be used to symbolize a row or a field in databases, a user of an information system, or a transaction recorded by the system, and so forth. The provenance information is recorded using these annotations as basic identifiers and organized as an algebraic structure called provenance semiring. Provenance semiring has been used to capture provenance for positive relational queries. In provenance semiring structure, the + symbol is used to describe the fact that some of the data item connected by the symbol are chosen for use, whereas data items connected by • will always be used together; the presence and absence of data during derivation are marked by 1 and 0 in the expression, respectively. As provenance semiring develops, descriptions of aggregation functions in the form like $\sum_i t_i \otimes v_i$ are proposed to capture aggregate queries. In such forms, v_i records the value to be aggregated (e.g., SUM, AVG, and MAX), and t_i symbolizes the provenance (annotation) attached on it. They are paired together with ⊗ to indicate the fact that t_i describes the provenance of v_i. The pair $t_i \otimes v_i$ as a whole is called a tensor. Tensors are collected together by the symbol Σ and symbolize the derivation process of aggregation. We use the following example similar to the one used in [1] to illustrate the use of provenance semiring.

2.3. Valuations and Provisioning. Supporting the operation of provisioning is the main reason why provenance semiring is proposed and also its main design goal. Provisioning is the operation of computing the changes to the results by applying some user-specified modifications to the data. We may do this by alternating the truth valuations applied to semiring expressions. For example, in the expression P_2 of Example 1, if we suspect UID_2 to be an abnormal user, we can then map UID_2 to false and calculate a new result based on the modified expression, thus discarding the contribution of UID_2. In existing literatures, such operation is formalized as the notation of $V : N[X] \to \{\mathrm{true}, \mathrm{false}\}$.

2.4. Summarization through Grouping and Mapping. As the derivation process gets more complex, the length and complexity of the corresponding provenance expression become more and more difficult for users to understand. Instead of offering the whole expressions to the users as raw information, provenance expressions should be summarized to reduce its size and highlight the major messages that need to be conveyed. It has been proposed [1] that the summarization of data provenance be achieved through a series of mapping of annotations. Put in simple words, multiple annotations are mapped to one common annotation so as to reduce the size of provenance expression (such mapping is denoted as $h(x)$). The mapped expression serves as a homomorphic but smaller form of the original provenance. During this process, the distinction between some original annotations is sacrificed for the reduced size of the whole expression.

2.5. Evaluate and Control Summarization Quality. Previous approaches evaluate the quality of provenance summarization mainly by size, distance, and semantic relatedness of the grouped annotations. It is worthwhile to first recall these three existing considerations. The first and most obvious consideration size is simply the number of annotations of a provenance expression, which largely determines its complexity. The second consideration is the semantic similarities between annotations to be grouped together. To ensure that the grouping process and summarization outcome make sense, only similar annotations pairs should be considered for mapping. In Example 1, for example, we allow ID_2 and ID_3 to be grouped together only because they share the *female* gender attribute. In other words, we allow two annotations

x_1, x_2 to be grouped together as one annotation when they share some common attribute or characteristics.

The third consideration is the distance between the original and summarized expression depending on the output value of the expressions under the hypothetical manipulations specified by users. Given a set of user-specified valuation V_X, a mapped valuation of V_X will be built for the summarized provenance expressions (denoted as V_X'). In [3] a function φ (combiner function) is provided to perform this mapping. In simple words φ provides descriptions about how summarized annotations will be discarded or retained in the mapped valuation according to valuation choices of the original provenance. For example, φ may decide that a summarized annotation would be discarded only if all original annotations it corresponds are discarded by the original valuation. The original valuation and the mapped valuation are applied on the provenance expression p and its summary $h(p)$, respectively, and the differences of the resulting between p and $h(p)$ are then collected and aggregated as the distance between them regarding the valuation. In detail, a function named VAL-FUNC would be needed to describe how such differences are collected and aggregated. Various sorts of function instances have been proposed to implement VAL-FUNC. For example, we may use the absolute difference between the two expressions values under the valuation as VAL-FUNC. Alternatively, we may introduce a function that returns zero if the two expressions produce the same output and one if any difference exits. For more choices of distance measures we refer readers to [3].

Apart from the three existing considerations listed as above, we propose to introduce *information amount* as an additional consideration to reflect and support more user requirement. The distance between original and summarized expression measures the error introduced by annotation grouping in terms of end-to-end provisioning result value. On the other hand, in terms of overall information loss, we lose track of the information about the original provenance annotations and elements every time we group some annotations together, as the grouped annotation make no distinction between them and consequently the underlying provenance elements (e.g., tuples and tensors) would have to be mixed together. Due to users' changing requirements in various real-world applications, there are a lot of scenarios in which user requirements can be better satisfied by retaining or reducing the information amount of certain kind of provenance elements or annotations during the summarization. This quality factor has not been considered in the work of [3, 5], and we will discuss more about how to measure the quality of the summarization in terms of its information amount in the next section.

Example 1. Consider three provenance expressions (inspired by and adapted from [3])

$$\begin{aligned} P &= \mathrm{ID}_1 \otimes (1,1) \oplus \mathrm{ID}_2 \otimes (3,1) \oplus \mathrm{ID}_3 \otimes (5,1) \\ P' &= \mathrm{ID}_1 \otimes (1,1) \oplus \mathrm{Female} \otimes (5,2) \\ P'' &= \mathrm{Audience} \otimes (3,2) \oplus \mathrm{ID}_3 \otimes (5,1)\,. \end{aligned} \tag{1}$$

In this case, P' and P'' are both summarized version of P in the sense that ID_2 and ID_3 are mapped to an abstract annotation "Female" and ID_1, ID_2 are mapped to an abstract annotation "Audience."

Both P' and P'' incur decrease of information amount since information about the original annotations and tensors are mixed in the new provenance expression. However, whether such decrease is good or bad to the users depends on use cases and requirements.

Since computing an optimal summarization is #P-hard, [3] has proposed an absolute approximation algorithm for computing the distance between two provenance expressions, by sampling the possible valuations and an approximation algorithm to compute optimal summarization with respect to the first three considerations, but not the fourth, that is information amount.

3. Capturing User Requirements with Weighted Information Balance

Observing that user requirements for the provenance summarizations can be expressed as importance weights assigned to each provenance element, we could include information amount as part of the quality consideration of provenance summarization. To do this, we need a quantitative measurement to evaluate the amount of remaining information during annotation grouping and provenance summarization. In this work, we introduce a generalization of entropy proposed by Guiaşu in [7, 8]. Intuitively, as the process of data grouping goes on, the distinctive power provided by the initial symbols or elements is gradually lost and converted to the homogeneity of the newly grouped symbols or elements. Consequently, information amount represented using the grouped annotations as a whole will decrease. We could quantify the information loss by computing the difference of information amount contained in the initial provenance annotation set and the one after summarization.

Let us suppose that we need to perform an analysis of a set of raw data items, for example, the set of provenance annotations. In this paper, we denote them as the set $X = \{x_1, \ldots, x_N\}$. In order to allow users to specify their preferences over the annotations for being preserved, we allow a corresponding set of weights $w_1, \ldots, w_N$ to be associated with the elements in X, respectively. In order to reduce the complexity of X, we consider the possibility that X is partitioned into a partition (scheme of annotation grouping) consisting of n sets with the form of

$$\mathscr{P}_n = \{X_1, \ldots, X_n\}, \tag{2}$$

where X_i are nonoverlapping subsets of X whose union is the set of X itself.

The initial amount of information supplied by X is defined in the form of the following:

$$\mathscr{I}(X) = -\sum_{k=1}^{N} w_k p_k \log_2 p_k, \tag{3}$$

where $P_1, \ldots, P_N$ are the probability distribution that elements in X are subject to.

However, as the original dataset is partitioned, the information amount contained in the partitioned symbols decreases. More specifically, the relation between the original raw data set X and the partition can be described as follows (information balance [7]):

$$\mathcal{I}(X) = \mathcal{I}(\mathcal{P}_n) + \mathcal{H}(\mathcal{P}_n). \quad (4)$$

Here $\mathcal{I}(\mathcal{P}_n) = -\sum_{i=1}^{n} w_k(X_i) p(X_i) \log_2 p(X_i)$ is the amount of information supplied by the classes of partition $\mathcal{P}_n$ (e.g., the provenance annotations after summarization), and $\mathcal{H}(\mathcal{P}_n)$ is the degree of homogeneity of the partition $\mathcal{P}_n$. After grouping, an amount $\mathcal{I}(X) - \mathcal{I}(\mathcal{P}_n)$ of information is removed as the distinction between the annotations mapped to the same summarized annotation is lost. On the other hand, this lost part is added to the data homogeneity $\mathcal{H}(\mathcal{P}_n)$. For detailed definition and discussions on how to compute $\mathcal{I}(\mathcal{P}_n)$, we refer readers to the related work of [7, 8].

Now let us consider how the variation of $\mathcal{I}(\mathcal{P}_n)$ and $\mathcal{H}(\mathcal{P}_n)$ may affect the outcome of provenance summarization in as the annotations are grouped together. It has been proved in [7] that $\mathcal{I}(\mathcal{P}_n)$ decreases as the grouping goes on whereas $\mathcal{H}(\mathcal{P}_n)$ increases. Therefore [7] also argues that when $\mathcal{I}(\mathcal{P}_n) = \mathcal{H}(\mathcal{P}_n)$, that is, when the amount of remaining information is equal to the amount of information converted into homogeneity, some kind of balance is reached. However, in the problem of provenance annotation summarization, users may want to specify their preferred degree of balance between annotation preserving and grouping. For example, when users want to have a more high-level view of the provenance information on certain aspects of the provenance data, higher homogeneity values may be more preferable. On the other hand, users may also want to reduce the loss of information on certain part of the information which he might deem to be interesting or helpful.

When $\mathcal{I}(\mathcal{P}_n) < \mathcal{I}(X)/2$, we have $\mathcal{I}(\mathcal{P}_n) < \mathcal{H}(\mathcal{P}_n)$, which can be understood as more information is retained rather than being converted to homogeneity. On the other hand, when $\mathcal{I}(\mathcal{P}_n) > \mathcal{I}(X)/2$, we have $\mathcal{I}(\mathcal{P}_n) > \mathcal{H}(\mathcal{P}_n)$ suggesting that more information is being converted to homogeneity than is remaining. Therefore, we could measure the balance between information preservation and information homogeneity using $\mathcal{I}(X)/2$ as a pivot point. To do this, we introduce the notation of *information balance index* as follows:

$$\mathrm{IB}(\mathcal{P}_n) = \frac{\mathcal{I}(\mathcal{P}_n) - \mathcal{I}(X)/2}{\mathcal{I}(X)/2} = \frac{2\mathcal{I}(\mathcal{P}_n)}{\mathcal{I}(X)} - 1. \quad (5)$$

It is easy to prove that $-1 \leq \mathrm{IB}(\mathcal{P}_n) \leq 1$. Intuitively, smaller $\mathrm{IB}(\mathcal{P}_n)$ value suggests that more information amount has been retained whereas greater $\mathrm{IB}(\mathcal{P}_n)$ value suggests that less information has been retained and the homogeneity is higher. In other words, $\mathrm{IB}(\mathcal{P}_n)$ measures the information balance status of a given partition $\mathcal{P}_n$.

Let us now consider how IBI is relevant in our problem of provenance summarization. As provenance annotations are being mapped together, the underlying provenance elements (e.g., tensors and tuples) will also be grouped into partitions. If we treat each element of our provenance expression as one data element and the grouping of the elements due to annotation grouping as the partitions in [7], we may then quantitatively measure the change of information amount during summarization.

Example 2. For example, if we treat each element in the expression $P = \mathrm{ID}_1 \otimes (1,1) \oplus \mathrm{ID}_2 \otimes (3,1) \oplus \mathrm{ID}_3 \otimes (5,1)$ as a data element, then we have a dataset $X_P = \{x_1, x_2, x_3\} = \{\mathrm{ID}_1 \otimes (1,1), \mathrm{ID}_2 \otimes (3,1), \mathrm{ID}_3 \otimes (5,1)\}$. By grouping ID_1 and ID_2 as Female, the grouped dataset becomes $X_{P'} = \{\mathbf{X}'_1, \mathbf{X}'_2\}$ where $\mathbf{X}'_1 = \{x_1, x_2\}$ and $\mathbf{X}'_2 = \{x_3\}$. Based on the probability (uniformed distribution or specified by users) and weights assigned to x_1, x_2, x_3 by the users, we may compute the information balance index of $X_{P'} = \{\mathbf{X}'_1, \mathbf{X}'_2\}$, $\mathrm{IB}(X_{P'})$, according to (5).

By continuously computing the information balance index of these groups, we may dynamically assess the amount of remaining information so as to support the decision-making of next annotation grouping operation. Therefore, users may express their requirements or preferences by assigning their preferred weights to each element. For example, users might assign higher weights to the "useful" provenance elements and lower weights to the "less useful" ones. Under this setting, provenance summarizations with higher amount of remaining information should be more desirable for the users. There exist many alternative ways for users to express their requirement; for example, users may assign higher weights to the items they are less interested in and encourage the amount of remaining information to be as low as possible.

Figures 2 and 3 show the curve of IBI with respect to the iteration steps of the summarization algorithm in [3] under different configurations of aggregation function and combination functions. IBI grows as the provenance summarization algorithm iterates, since more and more annotations are being grouped and the amount of remaining information decreases. We notice that different aggregation function or combination functions may lead to different speeds or patterns of information balance index growth. To deliberately alter this trend towards users' requirements, we could choose mapping candidate based on the additional factor of information balance index.

4. Balanced Provenance Summarization Computation Algorithm

In [3], candidates' mapping is chosen based on their candidate mapping scores defined as

$$\mathrm{CandidateScore} = w\mathrm{Dist} \cdot r\mathrm{Dist} + w\mathrm{Size} \cdot r\mathrm{Size}, \quad (6)$$

where wDist and wSize are the weights for size and distance and rDist and rSize are the rank of size and distance of summary after performing the candidate mapping. To include information balance into consideration, we could extend the original definition of candidate mapping score to

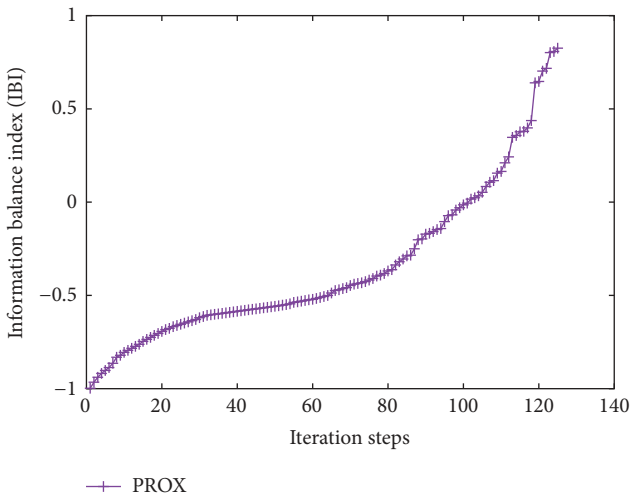

Figure 2: Using expression from MovieLens dataset, AVG aggregation, and "cancel one annotation" valuation, $w\text{Size} = w\text{Dist} = 0.5$, using conjunction as combination function.

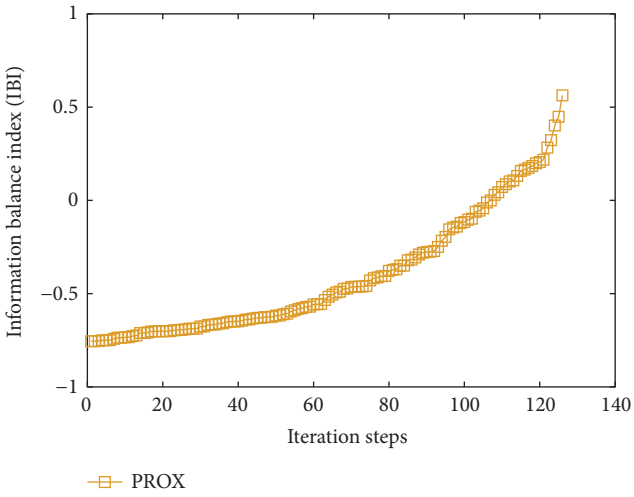

Figure 3: Same setting as in Figure 2, but using MAX as congruence aggregation function.

have an extra item involving the influence of the information balance index, as the following definition shows:

$$\text{CandidateScoreInf} = w\text{Dist} \cdot r\text{Dist} + w\text{Size} \cdot r\text{Size} + \text{scoreIBI}. \quad (7)$$

The influence of information balance index (scoreIBI) here can be defined in various forms to satisfy user requirements. One of the simplest forms could be

$$\text{scoreIBI} = w\text{IBI} \cdot r\text{IBI}, \quad (8)$$

where wIBI and rIBI stand for the weight and rank (in ascending order or descending order) of the information balance index. When using the ascending order, candidate mappings that lead to lower information balance (i.e., more useful information is retained) will be encouraged, whereas the using of a descending order will encourage higher homogeneity. Generally, users can alternate the definition of scoreIBI to a form that suits their requirements best; for example, to dynamically adjust the influence of IBI, we could even make scoreIBI a function of the current information balance index value and number of steps as

$$\text{scoreIBI} = f(r\text{IBI}, i), \quad (9)$$

where i denotes the current step number of summarization.

On the other hand, in our experiment we notice that the introduction of information balance index score (scoreIBI) sometimes has negative impacts on the size-distance quality of the provenance summarization, since it partially undermines the impact of the original size score and distance score in choosing a locally optimal candidate mapping. To counterbalance this negative influence, we propose to guarantee the size-distance quality by considering the candidate mappings with top k (or k percent) best size-distance quality only. That is, we choose candidate that has the highest CandidateScoreInf among those with the k (or k percent) best original CandidateScore.

Based on the above considerations, we present a new provenance summarization algorithm (see Algorithm 1). To satisfy their requirements, users can provide their preferred ranking function, weight function, and top-k-percent selection function as input to the algorithm.

Algorithm 1 extends the existing approximated provenance summarization algorithm in [3] by supporting the additional functionality of consideration information balance, by computing additional IBI information on a selected set of candidate mappings. Since the remaining information amounts matters in our algorithm, we do not perform equivalence grouping at the beginning as is done in [3]. The algorithm constructs the homomorphism h gradually in a greedy manner. The greedy decision is made according to the evaluation score consisting of considerations including not only size and distance but also our proposed IBI values. At each iteration, we examine a set of possible single-step mappings of two annotations to a new abstract annotation. For each mapping a homomorphism $h(p')$ of the current expression p' is computed so as to evaluate its candidate score and support the greedy decision. After that candidates of top-k percent size-distance performance are selected and evaluated for their IBI scores. Notice that since it is #P-hard to compute the exact distance between p_0 and $h(p')$, we approximate the distance value by sampling as is done in [3]. Right after that, the IBI scores of these k (k percent) candidates are computed and the candidate with the best total score is chosen and used for annotation mapping in the current iteration. Of course, the consideration of size and distance is still involved here, and they need to be combined according to some user-specified weights. Our algorithms differ from the original summarization algorithm in [3] in the sense that we perform a two-stage search to find first some promising candidates in terms of size-distance performance before computing and comparing their IBI performances. This will help the algorithm remain temporally efficient in spite of the additional computation requirement of IBI. We

Require: p_0 (original provenance), Ann (annotations in p_0), φ (combiner function) and V_{Ann} (VAL-FUNC function), the weight for distance, size, definition and weight of IBI score, selection size k, size bound TSIZE, distance bound TDIST
Returns: Summarized expression p_1
(1) Initialize p' as p_0
(2) **While** Size(p') > TSIZE or dist(p_0, p', V_{Ann}) < TDIST **Do**
(3) candidateSet = $\varnothing$
(4) **For** every $h \in$ FeasibleMapping(p') **Do**
(5) $p_{\text{cand}} = h(p')$
(6) Add p_{cand} to candidate set
(7) **End For**
(8) selectedSet = p_{cand} from candidateSet with top k percent size-distance performance
(9) **For** every p_{cand} **in** selectedSet **Do**
(10) **If** candScoreWithScoreIBI(p_0, p_{cand}) is optimal **Then**
(11) $p'_{\text{prev}} = p'$
(12) $p' = p_{\text{cand}}$
(13) **End if**
(14) **End For**
(15) **End While**
(17) **If** dist(p_0, p', V_{Ann}) > TDIST **Then**
(17) return p'_{prev}
(18) **End If**
(19) return p'

Algorithm 1: Information-Balance-Aware Approximated Provenance Summarization Algorithm (IB-PROX).

keep performing the mapping of annotations and reducing the size of provenance annotations set and stop when TSIZE is reached or the distance exceeds TDIST.

5. Evaluation

We conduct evaluations on two typical use cases to validate the effectiveness of our IBI-driven algorithm (Algorithm 1) in terms of its ability to steer IBI curve towards user requirements and also the application potentials of the proposed approach. In Use Case 1, we observe how IBI-driven algorithm could effectively retain the information amount of "useful" items at reasonable costs of size-distance performance. In Use Case 2, we pay a first visit to the possibility of using IBI to improve size-distance performance of provenance summarization.

5.1. Use Case 1 (Retain Useful Items Using IBI). In provenance expressions, there are often interesting or useful provenance tuples that users may prefer to be kept in the summary. For example, when provenance tuples annotated with some previously unobserved annotations start to occur in the retrieved provenance expressions, users would like to keep them in the summary to see how they differ with the previously seen ones. Another example is that users may want to retain some highly influential provenance elements (e.g., elements with outlier values and frequent attribute patterns), which could lead to significant deviations to provisioning results. To do this, users may assign higher importance weights to those tuples of higher interestingness and lower weights to the rest. In this way, grouping interesting tuples with the less interesting ones would lead to a more significant loss of information amount than grouping the less interesting ones only. Under this setting, it is obvious that provenance summarization with higher remaining information amount is more favorable.

This case study is to validate the effectiveness of information loss reduction by our IBI-driven algorithm and also observe the negative impact on size-distance performance by such reduction. The experiments for this case study were conducted using the MovieLens dataset for various configurations of weight, VAL-FUNC, and aggregation functions. We would like to point out that although only a subset of results are shown, the rest of the results which are not featured in this paper actually have similar characteristics.

In the experiment shown by Figures 4 and 5, we assume that users choose "Cancel Single Annotation" valuation and the AVERAGE aggregation function. We assign equal weights to wSize, wDist, and wIBI and randomly pick 25% of the tuples in the provenance expression as "interesting" tuples and make their weights one magnitude larger than the rest. In Figure 4, we choose a provenance expression with 200 tuples and compare the size-distance performance and information balance index curve of the results produced by the original provenance summarization algorithm in [3] (labelled as "PROX") and ours ("IB-PROX"). The blue plots and green plots stand for the results produced by our algorithm under the setting of $k = 5$ (selecting only candidates with top 5 size-distance performance for information balance index

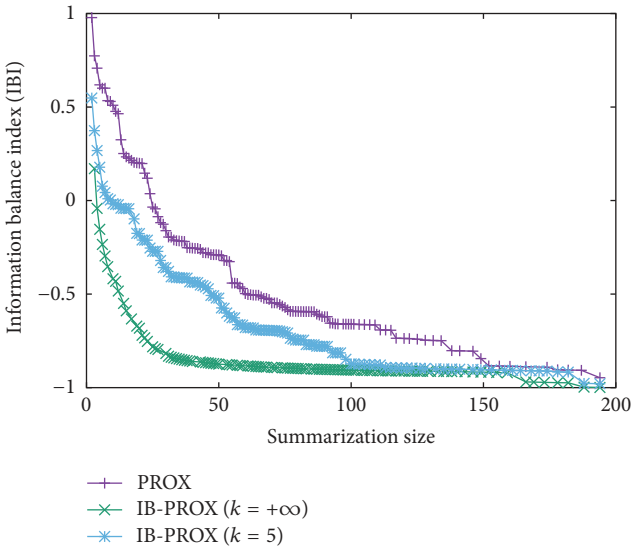

Figure 4: IBI curve can be effectively altered.

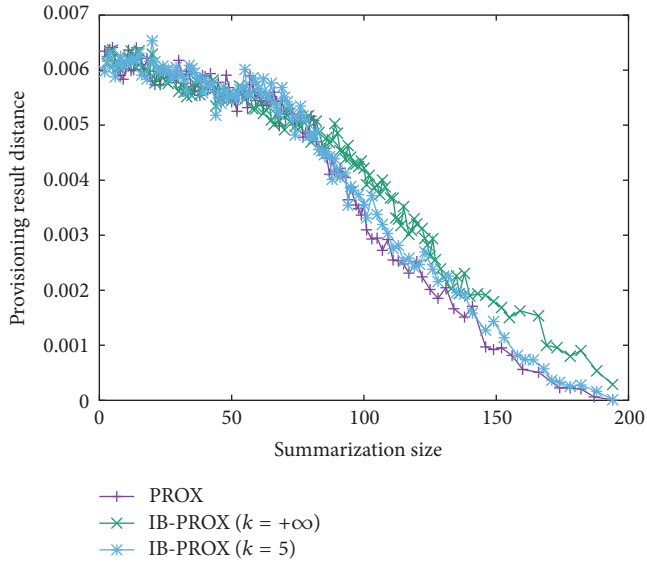

Figure 5: Negative impacts on size-distance performance can be counter-balanced by the select-top-k strategy.

comparison at each iteration) and $k = +\infty$ (selecting all candidate for information comparison). The red plots stand for the results produced by the original algorithm in [3].

By observing the balance index curves, we may observe how effectively information balance has been controlled by our algorithm. In Figure 4, information balance index curves of IB-PROX (green and blue) are significantly lower for summarization at all sizes than the original algorithm PROX (purple) (meaning more information content amount is retained).

From this trend, we may conclude that, by considering the information balance score, our algorithm can effectively alter the information balance index curve towards our desired bias (lower information balance index) to retain more information amount. Figure 5 shows the negative impact on size-distance performance. It can be observed that the negative impact is not quite significant until the size of summarization is less than 100. On the other hand, by using the select-top-k strategy, we can partially counterbalance this negative impact by sacrificing a certain amount of information loss reduction.

5.2. Use Case 2 (Better Summarization Quality Using IBI). In Use Case 1, we show that information balance index curve can be effectively "pushed down" to encourage interesting items to be retained (or not mixed with the lower-weight items as much as possible), by sacrificing a certain amount of size-distance performance. Seeing this, one might naturally come up with the question "Instead of worsening the size-distance performance, could we alternate the information balance index curve to improve size-distance performance? If possible, how?" It is reasonable to assume that if we could identify the "right" information that, when put together by grouped annotations, would incur less size-distance costs, then we are able to improve the size-distance performance by carefully choosing the weights assigned to each annotation. Admittedly, the definition or properties of the "right" information inevitably might vary due to the selection of valuation classes, aggregation function, datasets, and so forth. But it is still important to notice that we could build small successes one at a time in establishing specific correlations between information balance index and size-distance quality for some useful scenarios.

In this case study, we illustrate this point using the case of AVG aggregation. In this experiment we choose the "Cancel One Tuple" valuation class; that is, users may cancel one tuple (or more accurately, a tensor) from the provenance expression at a time for provisioning, and the combination rule is that if any tuple related to the grouped annotation is cancelled, the congruence tuple of the grouped annotation should be cancelled. In this case, we could notice an intuitive correlation between information balance and the size-distance performance. That is, higher homogeneity (higher information balance index value) in each grouped annotation may have positive impact on the size-distance performance. This is because, by creating higher homogeneities, more "raw" tuples will be grouped and considered in the tensor congruencies of the grouped annotations. Consequently, for each provisioning operation involving the cancellation of tuples related to the grouped annotations, the congruence value of the grouped tuples to be cancelled is closer to the global average. Therefore, the inaccuracies introduced by cancelling the congruence value rather than the original tuple will be smaller.

To validate this, we perform experiments to check the effect of information balance manipulation on the size-distance performance with multiple combinations of weights and information bias settings. All combinations demonstrate that, by encouraging higher homogeneity (i.e., higher information balance index values), the size-distance performance of the provenance summarization process can be improved. Due to space constraint, we show the representative results of the information balance index curve and size-distance

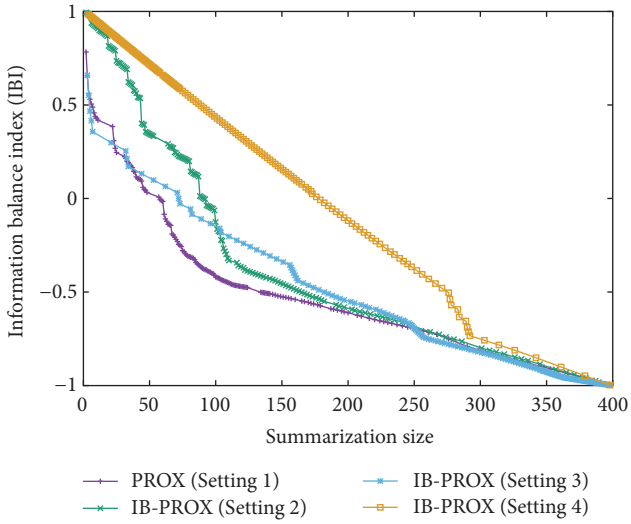

Figure 6: IBI curve under different settings.

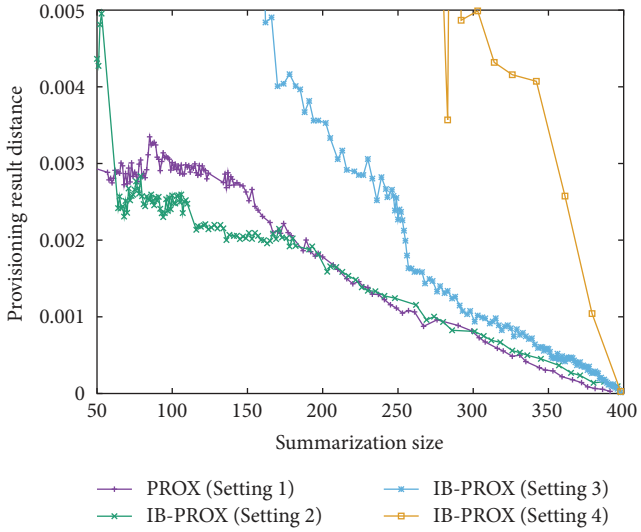

Figure 7: Size-distance performance under different settings.

performance produced by 4 settings of the summarization processes: (1) the original PROX process with wSize = Dist = 0.5, (2) IB-PROX with wSize = 0.25, wDist = 0.25, wIBI = 0.5, $k = +\infty$, and a bias towards "raised" information balance index curve, (3) the same configuration as in (2) but with a bias towards "lowered" information balance index curve, and (4) IB-PROX with wSize = 0, wDist = 0, wIBI = 1, and a bias towards "raised" information balance index curve.

We could observe from Figure 6 that although setting (3) tried to lower the information curve and successfully did that until the summary size drops to around 200, it failed to keep the trend and became even higher than the original summarization process (Setting 1). This can be explained by the fact that we still have wSize = 0.25 and wDist = 0.25 in Setting 3, so the bias towards a lower information balance index curve is counterbalance by the requirements of better size-distance performance, which favors higher homogeneity. This in a way reflects that there indeed exists a correlation between higher homogeneity and better size-distance performance.

We can see from Figure 7 that, by "raising" the information balance index curve (Setting 2), the size-distance performance is significantly improved after when the summary size dropped under 150. Although a small amount of negative impact on size-distance performance can be observed from the diagram, the overall improvement by Setting 2 is still highly significant. Among the settings which deliberately alter the information balance index curve, Setting 2 is the only setting that performs better than the original provenance summarization process. The other two settings (Settings 3 and 4) both create negative impact on the size-distance performance. Moreover, from the performance of Setting 4 we may conclude that although higher homogeneity may help to improve the size-distance performance, it is not the unique determining factor that can settle the size-distance performance once and for all. The size-distance performance can be improved only when the goal of higher homogeneity is considered and balanced in the context of size and distance ranking information.

The above results are representative for the summarization of provenance expressions involving AVG aggregation function only. For other scenarios with different aggregation functions and combination functions, the IBI-driven strategies and weight values would have to be redesigned to fit the scenario. Users may further apply machine learning techniques to automatically generate such strategies.

6. Discussion

Being analogous to the data grouping process in statistical analysis, the summarization of provenance through annotation mapping causes loss of information as the data provenance is summarized. With the analysis and the alternative provenance summarization algorithm presented in this work, we show that such losses could be well defined, measured, and controlled with the concept of information balance during provenance summarization (e.g., Figures 2 and 3). Consequently, information balance and the weights assigned to each provenance elements can serve as methods for users to specify and control the amount of information loss caused by abstract annotation mappings. The perspective provided by information balance differs from the traditional quality considerations used in [3] in the sense that it does not only focus on the end-to-end valuation error caused by the annotation mapping but quantitatively and statistically track the loss of information amount from the perspective of importance weight and information entropy.

In Use Case 1 of our evaluation, we observe that sometimes there exists a trade-off between users' preferred information balance bias and better size-distance performance of the provenance summarization. But still, information balance index (IBI) can be effectively alternated towards users' preferences at acceptable costs of size-distance performance deterioration if we could balance different considerations

properly by tuning the heuristic weights used by the search algorithm. On the other hand, in Use Case 2, we may also observe that under certain cases the importance weights and information balance goals can be carefully chosen such that we could achieve an even better performance in terms of both size and valuation error distance by heuristically optimizing the information balance goal. It turns out that the performance of our algorithm in Use Case 2 can be even better in terms of size-distance ratio than the previous algorithms which considers only size and distance as heuristics information.

With these observations, we may come to the realization that *information balance index* is a promising additional measurement of provenance summary quality that can be effectively optimized according to users' requirements and that it is a powerful way for users to specify their additional goals during provenance summarization and even a promising way to improve traditional quality goals such as summary size and valuation error distance when skillfully used. These findings imply that it would be an interesting and fruitful research direction to explore other sophisticated ways of specifying and optimizing the information balance goals in provenance summarization. As the incentives of the changes in *information balance index* and importance weights assigned to each provenance elements may largely determine the outcome of the provenance summarization process, they may serve useful tools for domain-specific provenance summarization solutions to adaptively observe and manipulate so as to reach their additional quality goals.

7. Related Work

Data provenance [9, 10] has been proposed to record how data is generated, propagated, and modified by different users or system modules. Many studies have demonstrated the wide scope of application that data provenance is capable of [11–13] and also the challenges [14, 15] we are faced with while applying provenance technologies. Among these challenges, the evergrowing size and complexity of provenance data have become a significant obstacle for users to understand the messages inside. Therefore, several provenance summarization or compression approaches have been proposed. In [16], an interactive way for exploring large provenance graph has been proposed to control the complexities presented to the users. In [17] the authors proposed to compress provenance graphs in a lossless manner so as to reduce spatial cost. In [18], abstract provenance graphs have been proposed to provide a homomorphic view of the provenance data to help users spot useful information. The most recent work of [3] proposed summarizing provenance data through a series of annotation mapping. However, to the best of our knowledge no existing works on provenance summarization consider the variation of information amount during provenance summarization, and none of the existing approaches allows users to control the loss of information or balance between homogeneity and information completeness by controlling entropy-like indices.

On the other hand, although the concept of entropy and its related derivatives have been successfully applied to a wide range of problems related to summary and compression [19–23], our work is the first attempt to try to involve the computation and control of entropy measurements with the user requirement specification during approximated provenance summarization. We believe that by allowing more flexible control of approximated data provenance summarization using entropy measurements, a wide scope of provenance-related tasks, for example, provenance based access control rules retrieval [24], provenance visualization [25], and provenance storage [26, 27], can be performed both more effectively and more efficiently.

8. Conclusion

More dedicated and diverse user requirements can be expressed and considered by the provenance summarization process by assigning importance weights to provenance elements. Information balance is introduced to measure the change of information content amount during provenance summarization process and included as part of the quality evaluation to achieve user goals defined in terms of biases on information balance. Experiment results show that IBI can be effectively manipulated at reasonable size-distance costs. As future work, promising directions include exploring more possible use cases of information balance driven provenance summarization and consequently new definitions of IBI score, weight evaluation function, and scenario-specific information balance index manipulation strategies, using domain knowledge and even machine learning techniques. It is also an important potential direction to explore how information balance information can be used to improve provenance summarization quality in more general cases.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the National Key Research and Development Program of China (no. 2016YFB0800901) and the program of China Scholarship Council (CSC) (no. 201606210384).

References

[1] Y. Amsterdamer, D. Deutch, and V. Tannen, "Provenance for aggregate queries," in *Proceedings of the 30th Symposium on Principles of Database Systems, PODS'11*, pp. 153–164, May 2011.

[2] T. J. Green, G. Karvounarakis, and V. Tannen, "Provenance semirings," in *Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2007*, pp. 31–40, June 2007.

[3] E. Ainy, P. Bourhis, S. B. Davidson, D. Deutch, and T. Milo, "Approximated summarization of data provenance," in *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015*, pp. 483–492, October 2015.

[4] Y. Amsterdamer, S. B. Davidson, D. Deutch, T. Milo, J. Stoyanovich, and V. Tannen, "Putting lipstick on pig: Enabling database-style workflow provenance," in *Proceedings of the VLDB Endowment 5.4*, pp. 346–357, 2011.

[5] P. Missier, K. Belhajjame, and J. Cheney, "The W3C PROV family of specifications for modelling provenance metadata," in *Proceedings of the 16th International Conference on Extending Database Technology, EDBT 2013*, pp. 773–776, March 2013.

[6] L. Moreau and M. Paolo, "PROV-DM: The PROV Data Model," 2013.

[7] S. Guiaşu, "Weighted entropy," *Reports on Mathematical Physics*, vol. 2, no. 3, pp. 165–179, 1971.

[8] S. Guiaşu, "Grouping data by using the weighted entropy," *Journal of Statistical Planning and Inference*, vol. 15, no. 1, pp. 63–69, 1986.

[9] P. Buneman, S. Khanna, and W. Tan, "Data Provenance: Some Basic Issues," in *FST TCS 2000: Foundations of Software Technology and Theoretical Computer Science*, vol. 1974 of *Lecture Notes in Computer Science*, pp. 87–93, Springer Berlin Heidelberg, Berlin, Germany, 2000.

[10] 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.

[11] J. Park, D. Nguyen, and R. Sandhu, "A provenance-based access control model," in *Proceedings of the 10th Annual International Conference on Privacy, Security and Trust (PST '12)*, pp. 137–144, Paris, France, July 2012.

[12] R. Lu, X. Lin, X. Liang, and X. Shen, "Secure provenance: the essential of bread and butter of data forensics in cloud computing," in *Proceedings of the 5th ACM Symposium on Information, Computer and Communication Security (ASIACCS '10)*, pp. 282–292, Beijing, China, April 2010.

[13] I. M. Abbadi, "A framework for establishing trust in Cloud provenance," *International Journal of Information Security*, vol. 12, no. 2, pp. 111–128, 2013.

[14] I. M. Abbadi and J. Lyle, "Challenges for provenance in cloud computing," in *Proceedings of the 3rd USENIX Workshop on the Theory and Practice of Provenance (TaPP)*, 2011.

[15] K.-K. Muniswamy-Reddy, P. Macko, and M. Seltzer, "Provenance for the cloud," in *Proceedings of the 8th USENIX Conference on File and Storage Technologies (FAST)*, vol. 10, 2010.

[16] P. Macko, D. Margo, and M. Seltzer, "Provenance map orbiter: Interactive exploration of large provenance graphs," in *Proceedings of the 3rd USENIX Workshop on the Theory and Practice of Provenance (TaPP)*, 2011.

[17] Y. Xie, K.-K. Muniswamy-Reddy, D. D. E. Long, A. Amer, D. Feng, and Z. Tan, "Compressing Provenance Graphs," in *Proceedings of the 3rd USENIX Workshop on the Theory and Practice of Provenance (TaPP)*, 2011.

[18] D. Zinn and B. Ludäscher, "Abstract provenance graphs: Anticipating and exploiting schema-level data provenance," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6378, pp. 206–215, 2010.

[19] L. Ferrier, *A maximum entropy approach to text summarization*, School of Artificial Intelligence, Division of Informatics, University of Edinburgh, 2001.

[20] H. Karloff and K. E. Shirley, "Maximum entropy summary trees," *Computer Graphics Forum*, vol. 32, no. 3, pp. 71–80, 2013.

[21] G. Ravindra, N. Balakrishnan, and K. R. Ramakrishnan, "Multi-document Automatic Text Summarization Using Entropy Estimates," in *Proceedings of the International Conference on Current Trends in Theory and Practice of Computer Science*, Springer Berlin Heidelberg, Berlin, Germany, 2004.

[22] J. Lin, "Divergence measures based on the Shannon entropy," *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[23] D. S. Ornstein and B. Weiss, "Entropy and data compression schemes," *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, vol. 39, no. 1, pp. 78–83, 1993.

[24] J. Pei and X. Ye, "Towards policy retrieval for provenance based access control model," in *Proceedings of the 13th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2014*, pp. 769–776, September 2014.

[25] P. Chen and B. A. Plale, "Big data provenance analysis and visualization," in *Proceedings of the 15th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing, CCGrid 2015*, pp. 797–800, May 2015.

[26] Z. Bao, H. Köhler, L. Wang, X. Zhou, and S. Sadiq, "Efficient provenance storage for relational queries," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM 2012*, pp. 1352–1361, November 2012.

[27] Y. Xie, D. Feng, Z. Tan et al., "A hybrid approach for efficient provenance storage," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM 2012*, pp. 1752–1756, November 2012.