Yuan Xue, EECS department, Vanderbilt University, yuan.xue@vanderbilt.edu

Research topics

Programming Models for the Cloud
Cloud Security, Privacy, and Auditing

Summary of current research activities related to Cloud Computing

Record linkage is the process of comparing the records from multiple resources to aggregate information on the same real-world entity, such as a patient, a customer, a business, a consumer report, a bibliographic citation, or a genome sequence. It has many applications across different industry sectors and government agencies. Many business, government agencies and research projects now are collecting massive amounts of data. Linking large-scale data sources is an extremely challenging task, as it faces significant performance barrier and resource consumption cost. Cloud computing platforms provide massive distributed computing resources that would make high-performance large-scale record linkage possible. Record linkage tasks are usually performed on an infrequent manner: The ability to pay for the use of computing resources on a short-term basis as needed and release them after use, can achieve great cost efficiency.

The objective of our research is to develop and evaluate models, algorithms, and system prototypes to support large-scale record linkage using a Cloud computing platform. Record linkage is a multi-step process, which includes data standardization, data encoding, blocking, field comparison, record comparison and classification. Our current research focuses on two issues that are critical to support record linkage in a Cloud: 1) privacy-preserving data encoding and 2) blocking (data partitioning).

Data encoding encodes sensitive fields (e.g., personal identifiers) within a record through a set of well-designed transformation functions. It is essential to protect the data privacy when linkage is performed on a third-party's system. Data encoding has a different requirement from encryption (e.g. applying DES encryption on each field of the record). In addition to protecting the confidentiality of the data, the encoded data need to be compared later for similarity to identify the same identify for linkage. In encryption algorithms, similar inputs with typologically variations will be encrypted into very different ciphertext, preventing accurate data linkage. A number of data encoding schemes are proposed in the existing literature in the context of privacy-preserving field comparators. Our work has performed a comprehensive empirical evaluation of these comparators in terms of their linkage accuracy, computational complexity and security. Our on-going work investigates the information theoretical model of data encoding that captures two aspects: 1) its capability in protecting the information confidentiality and 2) retaining information for accurate linkage.

Blocking is pertaining to parallel execution of record linkage. It determines the group of records that most likely to match, retrieves these records and creates partitions of the record set within which records will be compared and linked independently from other partitions. Existing blocking mechanisms are mostly designed in an ad hoc way. Our current work investigates the theoretical model which formulates blocking as a multi-objective optimization problem. This optimal blocking model can be used in determining optimal data partitions in parallel execution of record linkage, where linkage quality, execution time, and resource requirement are considered as optimization objectives. We are also investigating the impact of data encoding on blocking, as encoding changes the statistical structure of records and affects the capability of record similarity comparison, which poses greater challenge to blocking.

Abstract of future research plan

Our future research aims at building an end-to-end solution that enables record linkage as a core service on the Cloud computing platform. Such a service will support a transition toward cloud-based data management for distributed services, which will in turn facilitate the information exchange in many domains, for example, most notably, national health information exchange network.

Realization of such a solution requires technologies that can ensure (1) flexible usage: novice users can get their record linked without handling the technical details of unfamiliar algorithm development and security mechanisms, expert users can evaluate new record linkage methods using high-level programming primitives without concerning the parallelism and data access in the Cloud; 2) cost and performance awareness: scientists and administrators can estimate the time, the expense, the linkage quality, and thus choose the appropriate linkage method and configuration prior to submitting the task to the Cloud. We plan to focus on following research areas:

(1) Develop high-level parallel programming model and its run-time support in the Cloud that are tailored to the semantic of privacy-preserving record linkage. Multi-level parallelism exists in the record linkage process owe to the existence of data independence: 1) internal parallelism within building components such as within blocking and classification components; 2) external parallelism across building components, such as data partitioning of fields and record sets; and 3) overall parallelism among multiple runs of blocking components. We will develop a set of high-level programming primitives that can fully integrate, coordinate and support the multi-level parallelism within record linkage. The record linkage algorithm researchers only need to implement this set of functions. Their programs are automatically parallelized and executed on the Cloud. Note that record linkage process may have different execution flows depending on the input data, and can not be determined a priori. For example, multiple runs of blocking may be needed for some blocking methods, depending on the intermediate blocking results. Comparison results of one field may eliminate the need for other field comparisons. This requires the parallel model for record linkage to support data-aware dynamic execution flows. Existing parallel models for the Cloud, such as MapReduce and Dryad, which assume the data parallelism being modeled as a fixed data flow graph, cannot be directly applied to record linkage. Further research is needed on new programming models.

(2) Develop estimation and optimization techniques that enable user-aware cost-optimal record linkage in the Cloud. Besides the size of the dataset, the cost of a record linkage task running on the Cloud depends on multiple factors: 1) the record linkage method and implementation; 2) the number of data partitions and their sizes, which in turn affect 3) the number and the size of the virtual machines needed for deployment. This area of research will aid the user to find the optimal configurations for the linkage task in this multi-dimensional space. To do so, we will build on our current work on the empirical evaluation of privacy-preserving record linkage methods and develop profiles of primitive operations, with respect to their resource consumption and running time. Depending on the user's preference on the cost, running-time and linkage accuracy, the problem of Cloud deployment configuration will be formulated into an optimization problem, where the optimal blocking model provides the relationship between the data partitions and the linkage accuracy as constraints. The Pareto optimal solution of this problem will be provided as a guideline to user to define the deployment model.

(3) Develop security analysis framework and new cryptographic models and methods for privacy-preserving record linkage in the Cloud. Different from encryption, data encoding for privacy preservation in record linkage tries to minimize the information of individual plaintext survived in its ciphertext, while retaining the inter-plaintext similarity for linkage. Based on our current work on the information theoretical model of data encoding, we will develop a quantitative framework of privacy-preserving encoding, where the requirement of security and linkage accuracy can be parameterized in the encoding scheme. We will further analyze the security properties of record linkage on the Cloud by considering a wide variety of threats, including information leakage, data modification, malicious change in execution flow, and develop resilient record linkage mechanisms on the Cloud.