# Approximate Joins: Concepts and Techniques

Nick Koudas
University of Toronto
koudas@cs.toronto.edu

Divesh Srivastava
AT&T Labs–Research
divesh@research.att.com

## 1 Motivation

The quality of the data residing in information repositories and databases gets degraded due to a multitude of reasons. Such reasons include typing mistakes during insertion (e.g., character transpositions), lack of standards for recording database fields (e.g., addresses), and various errors introduced by poor database design (e.g., missing integrity constraints). Data of poor quality can result in significant impediments to popular business practices: sending products or bills to incorrect addresses, inability to locate customer records during service calls, inability to correlate customers across multiple services, etc.

In the presence of data quality errors, a problem central in this context is the ability to identify whether two entities (e.g., relational tuples) are *approximately* the same. Depending on the type of data under consideration, various "similarity metrics" (approximate match predicates) have been defined to quantify the closeness of a pair of data entities in a way that common mistakes are captured. A key operation in this context is, given two large multi-attribute data sets, identify all pairs of entities (tuples) in the two sets that are approximately the same. This operation has been well studied through the years and it is known under various names, including record linkage, entity identification, entity reconciliation and approximate join, to name a few. Given the significance and the inherent difficulty of the approximate join problem, a plethora of techniques have been developed in various communities, including the statistics, pattern matching, and the database communities, deploying diverse approximate match predicates.

The objective of this tutorial is to provide a comprehensive and cohesive overview of the key research results, techniques, and tools used for approximate joins. It complements recent data quality tutorials that present broad overviews of various aspects of data quality, and don't delve into the details of approximate join technology [2, 1].



**Proceedings of the 31st VLDB Conference,
Trondheim, Norway, 2005**

## 2 Tutorial Outline

The tutorial is example driven, and organized as follows.

- Formally define the various flavors of the approximate join problem as optimization problems.
- Contrast different approximate match predicates, based on their algorithmic properties, computational overhead, and adaptability.
- Review multiple methodologies for determining whether tuples of attributes are approximately the same, and techniques for adapting this decision problem into a join framework between two data sets.
- Identify key research areas requiring further work.

## 3 Professional Biographies

**Nick Koudas** is a faculty member at the University of Toronto, department of computer science. He holds a Ph.D. from the University of Toronto, an M.Sc. from the University of Maryland at College Park, and a B.Tech. from the University of Patras in Greece. He serves as an associate editor for the Information Systems journal and the IEEE TKDE journal. He is the recipient of the 1998 ICDE Best Paper award. His research interests include core database management, data quality, metadata management and its applications to networking.

**Divesh Srivastava** is the head of the Database Research Department at AT&T Labs-Research. He received his Ph.D. from the University of Wisconsin, Madison, and his B.Tech. from the Indian Institute of Technology, Bombay, India. He is on the editorial board of the ACM SIGMOD Digital Review, and is a co-chair of the ACM SIGMOD Workshop on Information Quality in Information Systems, 2005. His current research interests include data quality, IP network data management, and XML databases.

## References

[1] C. Batini, T. Catarci, and M. Scannapieco. A survey of data quality issues in cooperative information systems. *Pre-conference ER tutorial*, 2004.

[2] T. Johnson and T. Dasu. Data quality and data cleaning: An overview. *SIGMOD tutorial*, 2003.